MATHEMATICAL MODELING AND COMPUTING, Vol. 11, No. 3, pp. 835–847 (2024)

Twitter-sentiment analysis of Moroccan diabetic using Fuzzy C-means SMOTE and deep neural network

Roudani M.[1], Elkari B.[2], El Moutaouakil K.[1], Ourabah L.[2], Hicham B.[3], Chellak S.[3]

[1]*Engineering Science Laboratory (LSI), Faculty Polydisciplinary of Taza, USMBA, Morocco*
[2]*EIDIA, Euromed Research Center, Euro-Med University (UEMF), Fez, Morocco*
[3]*Faculty of Medicine and Pharmacy University Cadi Ayyad, Sidi Abbad, Marrakech, Morocco*

(Received 21 January 2023; Accepted 13 August 2024)

Effectively managing diabetes as a lifestyle condition involves fostering awareness, and social media is a powerful tool for this purpose. Analyzing the content of tweets on platforms like Twitter can greatly inform health communication strategies aimed at raising awareness about diabetes within the Moroccan community. Unfortunately, the corpus of tweets is imbalanced and the feature extraction leads to data sets with a very high dimension which affects the quality of sentiment analysis. This study focused on analyzing the content, sentiment, and reach of tweets specifically related to diabetes in Morocco. The proposed strategy processes in five steps: (a) data collection from Twitter platforms and manual labilization, (b) feature extraction using TF-IDF technique, (c) dimension reduction using deep neural network, (d) data balancing using Fuzzy C-Means SMOTE, and (e) tweets classification using five well-known classifiers. The proposed approach was compared with the classic system, which works directly on very large, unbalanced tweets. In terms of recall, precision, F1-score, and CPU time, the proposed system can perform highly accurate sentiment analysis in a reasonable CPU time.

Keywords: *diabetes; fuzzy C-means; SMOTE; deep neural network; sentiment analysis; classification.*

2010 MSC: 68T05, 68T50, 68T30 **DOI:** 10.23939/mmc2024.03.835

1. Introduction

The rise of social media has drastically reshaped interpersonal interactions in recent decades, serving as a conduit for widespread communication across various societal segments. Considered a socio-cultural phenomenon of the 21st century, social media has evolved into a robust tool addressing human needs in unprecedented ways. Platforms like Facebook and Twitter have been extensively utilized for discussing health-related concerns and sharing thoughts globally [1, 2]. These platforms have become pivotal sources of information, motivation, and guidance, particularly for those seeking online health-related solutions [3]. They provide a space for patients, health professionals, policymakers, and researchers to engage in vibrant health discussions [4]. Social media-based health communication has emerged as a cost-effective means of disseminating health promotion messages, proving effective in both research and practical implementation [5, 6]. With this context, our study aimed to enhance diabetes-related social media communication by analyzing Twitter messages.

Diabetes, a prevalent chronic non-communicable disease, has seen a surge in low and middle-income countries, notably in Morocco [7]. Effective health communication through social media can play a pivotal role in enhancing diabetes prevention, management, and self-care by encouraging lifestyle modifications and behavioral changes [8–10]. Creating awareness about diabetes among patients, the public, and healthcare professionals is a fundamental step in controlling and preventing diabetes in Morocco [11]. Studies have explored the potential of social media in creating healthcare awareness and knowledge about various health issues [12–15]. These platforms possess the capability to reach a vast population with health information and social support. Given the extensive discussion about diabetes on social media [4], leveraging these platforms can significantly raise awareness and disseminate accu-

rate information about diabetes. Previous research indicates that diabetes patients utilize social media platforms to discuss self-management and care, although this area remains relatively new [16,17]. Qualitative assessments of diabetes-related Facebook groups revealed significant improvements in members' diabetes management knowledge [18, 19]. These groups serve as forums for sharing personal experiences, seeking information, and obtaining social support. In-depth analyses of Norwegian Facebook groups showcased differences and inconsistencies in content among patients, patient organizations, and healthcare personnel groups [20]. Qualitative examinations of online blogs by individuals with diabetes unveiled emerging themes and subthemes [21]. Similarly, various studies have delved into diabetes-related Twitter messages using diverse methodologies. Computational approaches like topic modeling and linguistic analysis have been employed to comprehend public sentiment towards diabetes, diet, and obesity discussions on Twitter [22,23]. Others explored trends in the frequency of diabetes-related messages across different geographical areas and periods [24–26]. Lexicon-based sentiment analysis has also been applied to gauge and evaluate sentiments expressed about diabetes on Twitter [27, 28]. Understanding the nature, content, and structure of messages regarding diabetes shared by the public and healthcare professionals on social media is essential for accurate and widespread dissemination of factual information [20]. A comprehensive assessment of diabetes-related messages is crucial, encompassing topics, contributors, sentiments expressed, and preferences in themes and sentiments shared and appreciated by users on Twitter. Communication patterns, linguistic traits, and information-sharing behaviors on social media vary across regions and countries [29, 30]. Therefore, examining social media usage practices and health information-seeking behaviors specific to regions or countries concerning diabetes is critical to effectively utilize these platforms for promoting diabetes awareness and supporting the population.

To date, there is a lack of published studies conducting in-depth content analysis of diabetes-related Twitter messages originating from Morocco. In light of this, our aim was to evaluate the source (individuals/organizations), topics (themes and sub-themes), sentiments, and audience reach of social media messages about diabetes in Morocco. This study focuses on Twitter, a widely used social networking platform for discussing diverse societal issues [31]. Twitter is a continuous source of current data and community perspectives on health topics and policies [32]. This study centered on the comprehensive analysis of tweets concerning diabetes in Morocco, encompassing content, sentiment, and outreach evaluations. The outlined strategy follows five key stages: (a) retrieving data from Twitter platforms with manual labeling, (b) employing the TF-IDF technique for feature extraction, (c) utilizing deep neural networks for dimension reduction, (d) implementing data balancing via Fuzzy C-Means SMOTE, and (e) employing five established classifiers for tweets classification. A comparative analysis was conducted between the proposed approach and the conventional system, which directly handles extensive, imbalanced tweet datasets. The proposed system exhibits superior performance in terms of recall, precision, F1-score, and CPU time, effectively conducting precise sentiment analysis within reasonable computational timeframes.

The rest of the paper is organized as follows: the second section gives the smart tools used in this paper. The third section presents the proposed approach. The fourth section provides different experimental results. The last section presents some conclusions and propositions for future work.

2. Smart tools

In this section, we give the principles of different smart tools used in this paper to realize our objective, namely analyzing the content, sentiment, and reach of tweets specifically related to diabetes in Morocco.

2.1. TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) is widely used to transform tests to numerical vectors based on two main terms Term Frequency (TF) and Inverse Document Frequency (IDF) [33].

Term Frequency (TF): TF quantifies how often a term is present within a particular document. It is computed by dividing the frequency of a term's occurrence in a document by the total number of

terms in that document,

$$\mathrm{TF}(t,d) = \frac{\text{Total number of terms in document } d}{\text{Number of times term } t \text{ appears in document } d}.$$

Inverse Document Frequency (IDF): the IDF evaluates the scarcity of a term across a set of documents. It is determined by taking the logarithm of the ratio between the total number of documents in the entire corpus and the number of documents containing the specific term t,

$$\mathrm{IDF}(t,D) = \log \frac{\text{Total number of documents in the corpus } |D|}{\text{Number of documents containing term } t}.$$

The TF-IDF score for a term t in document d is obtained by multiplying its Term Frequency (TF) by its Inverse Document Frequency (IDF) values:

$$\text{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D).$$

TF-IDF balances term frequency within a document (TF) with the rarity of the term across all documents (IDF). A high TF-IDF score signifies a term that is both frequent in a specific document and unique across the entire document collection.

2.2. Auto-encoder

As a pre-processing step, we use a deep multilayer neural network (called auto-encoder) to project the corpus to a space of very small dimensions $(50-150)$. The auto-encoder is a deep neural network composed of two principal sections: the encoder (box of neurons) and the decoder (box of neurons) [34,35]. The hidden layer gives the coded information and the last layer must produce the input tests; Figure 1 gives the architecture of an auto-encoder that produces an artificial tweet (eligible) but contains the main information of the real tweet.

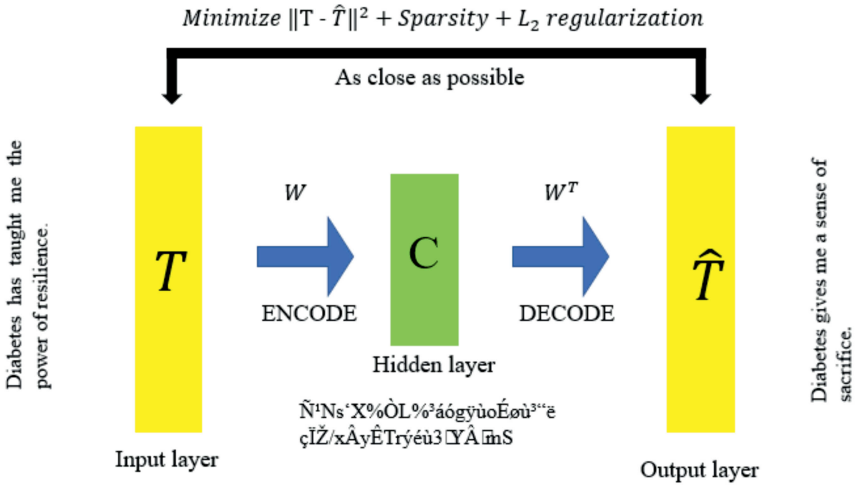

Fig. 1. The architecture of an auto-encoder.

Loss function: the loss function E quantifies the sum of the local loosed information when transforming t to artificial tweet: $a_t\|t-\hat{t}\|$. If the encoding operation is realized by a mapping P and D then $a_t = P(W_e * t + b_e)$ and $\hat{t} = D(W_d * \hat{t} + b_d)$. The global error is given by $E(W_e, W_d) = \sum_{t\in \mathrm{CORPUS}} \|P(W_e * t + b_e) - D(W_d * \hat{t} + b_d)\|$; where CORPUS is the set of the collected tweets.

Sparsity regularizer: the sparseness of an autoencoder can be promoted by incorporating a regularizer into the cost function [36]. The sparsity regularizer aims to constrain the sparsity of the output of the cached layer. Dispersion can be encouraged by the introduction of a regularization factor, which will take on a large magnitude when the average firing rate of a given neuron and its target value is not very close [37]. For example, we can consider Kullback–Leibler divergence as sparsity regularizer noted S_{KL} [38].

L_2 regularized: to guarantee the density of the hidden layer output without increasing the encoder weight values, we can add the L_2-regularizer: $R_2 = \frac{1}{2}\|W_e\|$ [38].

2.3. Fuzzy C-Means

Fuzzy C-Means is a soft clustering method that allows dividing N, non-labeled objects, described in $\mathbb{R}^n$, into K-groups. Unlike the hard methods, this method permits the objects to be in different groups, at the same time, using membership functions [39]. To this end, the Fuzzy C-Means try to solve the following optimization problem:

$$(FP): \quad J(\mu, w) = \sum_{i=1}^{N} \sum_{c=1}^{\boldsymbol{K}} \mu_{c,i}^{m} \left\| z_i - w_c \right\|^2,$$

where z_i is the ith sample from $\mathbb{R}^n$, $m \in]1, +\infty[$, $\mu_{c,i}$ informs us how much the sample z_i is in the group c, and w_c is the center of the cth cluster.

Fuzzy C-Means process in iterative optimization of the problem FP:

(a) $\forall i$ and $\forall c$, $\mu_{c,i}^{m,0}$ and w_c^0 are randomly chosen,

(b) at the iteration k, $\forall i$ and $\forall j$, $\mu_{c,i}^{m,k}$ and w_c^k are known and $\mu_{c,i}^{m,k+1}$ and w_c^{k+1} are calculated using the following learning equations:

$$\mu_{c,i}^{m,k+1} = \left(\sum_{a=1}^{K} \left(\frac{|z_i - w_c^k|}{|z_i - w_a^k|} \right)^{\frac{2}{m-1}} \right)^{-1}, \quad w_c^{k+1} = \left(\sum_{q=1}^{N} \mu_{c,q}^{m,k} z_q \right) \left(\sum_{q=1}^{N} \mu_{c,q}^{m,k} \right)^{-1},$$

(c) return to (b) until $\max\limits_{i,c} \left| \mu_{c,i}^{m,k+1} - \mu_{c,i}^{m,k} \right| \leqslant \varepsilon$, where ε is a very small non-negative real number.

2.4. SMOTE

The Synthetic Minority Over-Sampling Technique (SMOTE) is one of the most well-known methods to solve the unequal class distribution problem in imbalanced datasets. The SMOTE method process follows three steps:

1. For each minority sample m and for each x from D_m, this study calculates the distance.
2. For each minority sample m, this study determines the k neighbor from D_m noted N_{gm}.
3. For a given minority sample m, this study generates y randomly from N_{gm} majority samples and generates a new sample m_{new} using the formula:

$$m_{new} = m + \text{rand}(0, 1) \times (m - y).$$

3. Proposed method

In this part, we formulate a system that permits us to determine whether the sentiment of a tweet is positive or negative considering the diabetes disease. The approach to sentiment analysis adopted in this study is depicted in Figure 2.

First, we use the pre-processing phase to eliminate text noises and cleaning. Second, we transformed the text into a vector using TF-IDF. Third, we use the Auto-encoder to reduce the dimension of the vectors. Fourth, we use Fuzzy C-Means SMOTE (FCM SMOTE) [38] to balance data. Fifth, we classify data into two categories (positive or negative) using a classifier such as KNN, LDA, ANN, SVM, and NB [35, 36, 40–44].

1. **Tweets collection:** manual tweet collection encompasses directly browsing, searching, and documenting tweets from Twitter's platform, omitting the use of automated tools or scripts for the process. Subsequently, we categorized all comments into two groups: positive and negative sentiments [33].
2. **Cleansing Tweets:** this procedure encompasses multiple stages aimed at refining, preparing, and refining data. It includes tasks such as eliminating special characters, managing mentions, hashtags, and URLs, standardizing text by converting to lowercase, managing stop words, and conducting stemming or lemmatization processes [33]. We denote by Cleaning() the tweets cleaning function.
3. **Feature Extraction:** in this step, we utilize TF-IDF to transform text data into numerical vectors. Each dimension of these vectors signifies the significance of a term within a document concerning the entire corpus. TF-IDF finds extensive application in natural language processing,

facilitating tasks like text classification, clustering, and information retrieval. It excels in capturing term importance while diminishing the influence of commonly occurring terms. We denote by TF _ IDF() the function to create numerical tweets.

4. **Dimension Reduction:** employing an auto-encoder entails leveraging neural networks to grasp a condensed depiction of data, achieving dimensionality reduction while endeavoring to preserve crucial information. We denote by DeepNN(,) the function to create artificial tweets.
5. **Data balancing:** this study uses Fuzzy C-Means SMOTE to balance dataset by implementing the following steps:
   — **Grouping:** use Fuzzy C-Means with $C = K$ to separate input data into K groups.
   — **Filtering:** identify clusters for oversampling based on the ratio between minority and majority instances. These selected clusters contain at least 50% minority samples. Determine the number of artificial samples to generate using the provided SW formulas mentioned in equations (1) and (2):

$$\mathrm{SF}(f) = \frac{\text{avrage minority distance } (f)^N}{\text{minoritycount}(f)}, \quad N = \text{number of features in Data set}, \tag{1}$$

$$\mathrm{SW}(f) = \frac{\mathrm{SF}(f)}{\sum_{f \in \text{filtred cluster}} \mathrm{SF}(f)}. \tag{2}$$

   — **Interpolation:** implementing SMOTE in conjunction with determining the number of synthetic samples for each filtered cluster, facilitates the allocation of new samples to their respective clusters,

$$\text{number of artificial}_f = \|n \times \mathrm{SW}(f)\|, \quad n = \text{majoritycount} - \text{minoritycount}. \tag{3}$$

   We denote by FCM _ SMOTE(, , ,) the function to balance numerical tweets.
   It should be noted that most algorithms are complex and tend to generate unnecessary noise. The clustering SMOTE-based method avoids the creation of noise and effectively overcomes imbalances between and within classes.
6. **Sentiment Analysis:** involves classifying text into predefined sentiment categories (positive or negative). Using five different classifiers, we explore various machine learning algorithms and their performance in sentiment analysis tasks. We denote by TrainedClassifier() the function to predict the sentiment of the unseen tweet.

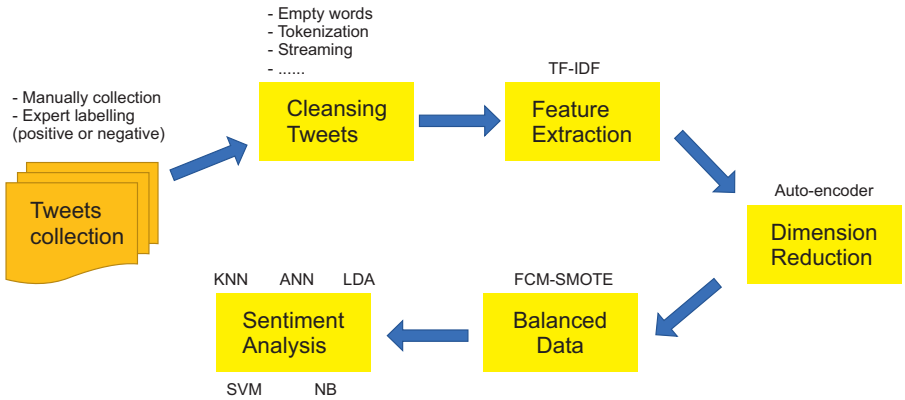

Fig. 2. Methodological system for Sentiment Analysis.

Algorithm 1 describes the steps of our approach. First, let us give some necessary notations that permit us to build our algorithm.

Notation:

UT: Umbalanced tweets data set;
NUT: Numerical Umbalanced Tweets data set;
$ANUT$: Artificial Umbalanced Tweets data set;
$BANT$: Balanced Numerical Tweets data set;

K: Number of cluster;
h: Number of hidden neurons of deep auto-encoder;
t_u: Unseen Tweet;
N_u^t: Numerical Unseen Tweet;
AN_u^t: Artificial Numerical Unseen Tweet;
p_u^t: Sentiment Prediction of unseen tweet t_u.

Algorithm 1 Diabetic sentiment analysis Twitter based.

Input: UT, K, h, t_u
Output: p_u^t
BEGIN
$UT' \leftarrow \text{Cleaning}(UT)$;
$t'_u \leftarrow \text{Cleaning}(t_u)$;
$NUT \leftarrow \text{TF\_IDF}(UT')$;
$N_u^t \leftarrow \text{TF\_IDF}(t'_u)$;
$ANUT \leftarrow \text{DeepNN}(NUT, h)$;
$AN_u^t \leftarrow \text{DeepNN}(V_{t'_u}, h)$;
$BANT \leftarrow \text{FCM\_SMOTE}(ANUT, K, h, 1)$;
$TrainedClassifier \leftarrow \text{trainClassifier}(BANT)$;
$p_u^t = \text{TrainedClassifier}(AN_u^t)$;
END

4. Experimentation

4.1. Data set

Manual tweet collection involves the direct browsing, searching, and recording of tweets from Twitter's platform without using automated tools or scripts, where people express their opinions and thoughts in text form about living with diabetes. We then grouped all comments into two categories: positive and negative.

4.2. Tweets representation

In this section, Pre-Processing is required to eliminate text noises. It is performed before the classification phase, and it is a necessary step due to the unformal text to make correct predictions. The cleaning steps employed in this paper are as follows:

1. Special characters, URLs, and Punctuation are removed.
2. Slang terms are replaced or removed.
3. Emojis are removed.

After pre-processing phase, we use TD-IDF to represent all dataset.

4.3. Artificial tweets

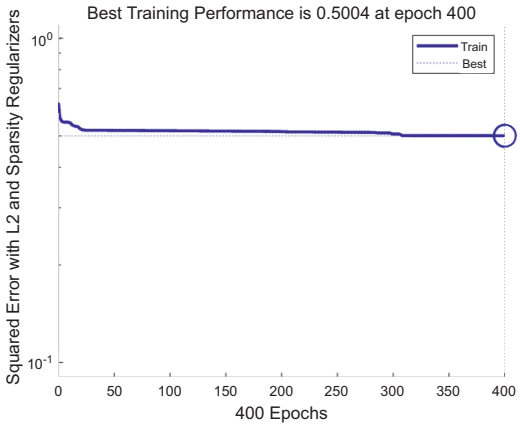

Fig. 3. The size of the projection space 50.

In this section, we use the deep neural network model, called the auto-encoder [34, 35], to project the text into a low-dimensional $(50 - 150)$ description space to minimize the CPU time of sentiment analysis of diabetes data. The auto-encoder configuration is as follows: Max epochs $=$ 400, Encoder transfer function $=$ 'satlin', Transfer function $=$ 'purelin', $L2$ weight regularization.

For example, Figure 3 shows the evolution of the squared error with $L2$ and sparsity regularizations as a function of the number of epochs. We note that the auto-encoder converges very

early (around 80 epochs), which means that the optimizer has been attracted to local minima by the initial choice of auto-encoder parameters. To improve the learning quality of the auto-encoder, we can consider different initialization.

Table 1 gives the mean square error, CPU time, and gradient value for different projection spaces constructed by the auto-encoder for different numbers of hidden layers. We note that the MSE is of the order of 10^{-2} and the smallest error 0.001503 corresponds to a projection space of dimension 140. The CPU time required to project the data is 10^{-2} and the shortest CPU time corresponds to 150 hidden neurons. Finally, the order of the gradient of the loss function is of the order of 10^{-6}, which corresponds to a good local minima of the loss function.

Table 1. MSETimeGrad: mean square error, CPU time, and gradient value for different projection spaces built by auto-encoder for different numbers of hidden neurons.

Projection space size	MSError	Time (s)	Gradient (10^{-6})
50	0.001439924592557	0.002	16.3
60	0.001582820559766	0.0013	0.0141
70	0.002669638811186	0.0015	0.00432
80	0.001706938476901	0.0016	0.0399
90	0.001526629597353	0.0017	0.00408
100	0.019156182944591	0.0023	0.0271
110	0.001662871126882	0.0025	0.0179
120	0.005626419409826	0.0040	0.0204
130	0.001767278343455	0.0058	0.0391
140	0.001502845969274	0.0035	0.00232
150	0.001595545208926	0.0035	0.00225

4.4. Evaluation metrics

In Machine Learning, classification models are evaluated through a series of metrics such as overall accuracy (OA), which is defined as the proportion of correct predictions compared to the total number of predictions.

1) Precision: precision is the quantity of positive samples compared to the ones that the classifier labels positive, used to evaluate the accuracy of classifier's predictions,

$$\text{Precision} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

2) Recall: recall is the amount of samples that were correctly predicted as positive compared to actual positive samples. The importance of such a metric can be observed in sensitive applications whereby accurate predictions of rare instances are essential,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

3) F-Measure (FM) is

$$\text{F-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}.$$

4) Accuracy represents the ratio of correctly predicted instances to the total number of instances in a dataset,

$$\text{Accuracy} = \frac{\text{Total number of predictions}}{\text{Number of correct predictions}}.$$

Table 2 gives the values of different measures of the proposed method applied to the imbalanced initial tweets data set. Table 2 shows that the NB classifier has the highest accuracy and F1-score among the listed models, while SVM also demonstrates strong performance across these metrics. LDA shows

Table 2. Performance of the considered classifiers on imbalanced initial tweets data set.

Model	**Performance measures**			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	68%	70%	59%	64.2%
LDA	82%	77%	77%	77.04%
ANN	75.4%	73.92%	70.22%	72.03%
SVM	84.61%	81.24%	81.38%	81.29%
NB	85.11%	82.35%	82%	82.17%

balanced precision and recall scores. Dealing with imbalanced data requires specific strategies to enhance model performance. High-dimensional data requires more computational resources and time for training models. Some algorithms might struggle to handle large feature spaces efficiently. To overcome this issue we use auto-encoder different numbers of hidden neurons $(50 - 150)$ to reduce dimension.

Table 3. Performance of the proposed system applied to the imbalanced reduced $(50, 60, 70)$ tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	67.2%	69.31%	58.06%	63.65%
LDA	81.69%	76.56%	76.66%	76.67%
ANN	74.99%	73.12%	69.79%	71.63%
SVM	83.77%	80.95%	80.75%	80.6%
NB	84.66%	81.55%	81.76%	81.54%

Table 3 gives the measures of the proposed method applied to artificial the imbalanced reduced $(50, 60, 70)$ tweets data set. From the table, it seems that the NB classifier performs consistently well across different percentages of the reduced dataset, with high scores in accuracy, precision, recall, and F1-score. SVM also demonstrates robust performance across these metrics. LDA shows balanced precision and recall scores. In the experiment, we obtained almost the same values in the three unbalanced reduced datasets $(50, 60, 70)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(50, 60, 70)$.

Table 4. Performance of the proposed system applied to the imbalanced reduced $(80, 90, 100, 110)$ tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	67.71%	69.53%	58.17%	63.96%
LDA	81.77%	76.76%	76.76%	76.87%
ANN	75.14%	73.48%	69.82%	71.85%
SVM	84.29%	81.03%	80.77%	80.72%
NB	84.79%	82.31%	81.83%	81.92%

Table 4 gives the measures of the proposed method applied to artificial the imbalanced reduced $(80, 90, 100, 110)$ tweets data set. From the table, it seems that the NB classifier performs consistently well across different percentages of the reduced dataset, with high scores in accuracy, precision, recall, and F1-score. SVM also demonstrates robust performance across these metrics. LDA shows balanced precision and recall scores. In the experiment, we obtained almost the same values in three unbalanced reduced datasets $(80, 90, 100, 110)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(80, 90, 100, 110)$.

Table 5. Performance of the proposed system applied to the imbalanced reduced $(120, 130, 140, 150)$ tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	67.94%	69.78%	58.84%	64.08%
LDA	82%	76.85%	76.9%	76.96%
ANN	75.31%	73.51%	69.97%	71.95%
SVM	84.52%	81.05%	80.82%	81.23%
NB	85.02%	82.35%	81.98%	82.15%

Table 5 gives the measures of the proposed method applied to artificial the imbalanced reduced $(120, 130, 140, 150)$ tweets data set. From the table, it seems that the NB classifier performs consistently well across different percentages of the reduced dataset, with high scores in accuracy, precision, recall, and F1-score. SVM also demonstrates robust performance across these metrics. LDA shows balanced precision and recall scores. In the experiment, we obtained almost the same values in three unbalanced reduced datasets $(120, 130, 140, 150)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(120, 130, 140, 150)$.

Table 6. Performance of the considered classifiers on balanced initial tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	79.43%	81.67%	72.42%	77.66%
LDA	93.81%	87.45%	89.79%	91.15%
ANN	89.82%	88.82%	82.01%	86.1%
SVM	96.14%	96.11%	96.1%	94.27%
NB	94.04%	95.04%	92.36%	94.7%

Table 6 gives the values of different measures of the proposed method applied to the balanced initial tweets data set. From the table, it seems that SVM shows the highest performance across all

metrics, followed by NB and LDA. The performance of the models is notably higher compared to their performance on imbalanced datasets due to the equal representation of classes, allowing models to learn more effectively without bias towards the majority class.

Table 7 gives the measures of the proposed method applied to artificially balanced reduced $(50, 60, 70)$ tweets data set. From the table, it seems that SVM consistently demonstrates high performance across all metrics. LDA and NB also exhibit strong performance, while KNN and ANN show slightly lower performance in comparison.

Table 7. Performance of the proposed system applied to artificially balanced reduced $(50, 60, 70)$ tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	79.08%	81.31%	71.94%	77.25%
LDA	93.4%	86.96%	89.36%	90.81%
ANN	90.81%	88.34%	81.69%	85.68%
SVM	95.9%	95.75%	95.61%	93.81%
NB	93.8%	94.69%	91.94%	94.45%

Artificially balancing the reduced datasets helps the models perform better by providing a more even representation of classes, allowing the classifiers to learn more effectively. In the experiment, we obtained almost the same values in the three balanced reduced datasets $(50, 60, 70)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(50, 60, 70)$.

Table 8 gives the measures of the proposed method applied to artificially balanced reduced $(80, 90, 100, 110)$ tweets data set. From the table, it seems that SVM consistently demonstrates high performance across all metrics. LDA and NB also exhibit strong performance, while KNN and ANN show slightly lower performance in comparison.

Table 8. Performance of the proposed system applied to artificially balanced reduced (80,90,100,110) tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	79.16%	81.44%	72.14%	77.34%
LDA	93.51%	86.96%	89.49%	90.84%
ANN	90.84%	90.84%	81.84%	85.88%
SVM	95.93%	95.77%	95.96%	93.84%
NB	94.01%	94.91%	92.03%	94.47%

Artificially balancing the reduced datasets helps the models perform better by providing a more even representation of classes, allowing the classifiers to learn more effectively. In the experiment, we obtained almost the same values in three balanced reduced datasets $(80, 90, 100, 110)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(80, 90, 100, 110)$.

Table 9. Performance of the proposed system applied to the balanced reduced $(120, 130, 140, 150)$ tweets data set.

Model	Performance measures			
	Accuracy	***Precision***	***Recall***	***F1-score***
KNN	79.24%	81.53%	72.33%	77.58%
LDA	93.78%	87.2%	89.56%	91.09%
ANN	91.09%	88.56%	81.89%	86.06%
SVM	96.12%	95.87%	95.99%	93.93%
NB	94.03%	95.03%	92.19%	94.65%

Table 9 gives the measures of the proposed method applied to artificially balanced reduced $(120, 130, 140, 150)$ tweets data set. From the table, it seems that SVM consistently demonstrates high performance across all metrics. LDA and NB also exhibit strong performance, while KNN and ANN show slightly lower performance in comparison.

Artificially balancing the reduced datasets helps the models perform better by providing a more even representation of classes, allowing the classifiers to learn more effectively. In the experiment, we obtained almost the same values in three balanced reduced datasets $(120, 130, 140, 150)$ with a difference between 0.001% and 0.0015% which is very negligible. For this reason, we considered a single table for three dimensions $(120, 130, 140, 150)$.

4.5. Comparison results

Model performances show a trend of maintaining relatively similar levels of performance despite more significant dimension reduction. SVM and NB models seem to be the most stable and robust in the face of more significant dimension reductions, maintaining relatively consistent performance levels. KNN, LDA, and ANN models exhibit a slight trend towards decreased performance with more drastic

dimension reductions, although the variations remain relatively moderate. This comparison highlights certain models' ability to maintain stable performances despite more extensive dimension reduction, which can be crucial in scenarios where data size is limited. By contrasting the metrics between imbalanced, reduced datasets (Tables 2 to 5) and artificially balanced datasets (Tables 6 to 9), it becomes evident that Fuzzy C-Means SMOTE methods significantly contribute to improving model accuracy, precision, recall, and F1-score, especially in the face of dimensional reduction.

Tables 6–9 (Balanced): generally exhibit higher accuracy across models compared to Tables 2–5 (Imbalanced). This improvement reflects the positive impact of balancing classes on overall accuracy.

Precision and Recall values in Tables 6–9 (Balanced) often surpass those in Tables 2–5 (Imbalanced). Balanced datasets typically yield higher precision and recall due to equal class representation, showcasing the effectiveness of oversampling.

F1-scores in Tables 6–9 (Balanced) are relatively stable across reductions compared to Tables 2–5 (Imbalanced). Oversampling enhances F1-scores by maintaining a balance between precision and recall.

Assess how reductions impact model performance. Identify models showing stability or degradation in performance with increased reduction levels.

5. Conclusion

Profiling the content of tweets on social media platforms, such as Twitter, could greatly aid health communication tactics aimed at raising diabetes awareness in Moroccan society. When diagnosing in-depth sentiment analysis based on tweets, analyzers are faced with two fundamental problems: data imbalance and the high dimensionality of tweet descriptions. In this paper, we proposed a sentiments analysis system that processes in five steps: data collection from Twitter platforms and manual labialization, feature extraction using TF-IDF technique, dimension reduction using deep neural network (for different numbers of hidden neurons), data balancing using Fuzzy C-Means SMOTE, and tweets classification using five well-known classifiers. We tested these systems on our data set and 11 hidden projection spaces are considered ($50 : 10 : 150$). In addition, five classification methods are used (KNN, LDA, ANN, SVM, and NB). With respect to three popular performance measures (accuracy, F1-Score, precision, and recall), the proposed systems were capable of analyzing tweets with high precision for different projection spaces. And the balancing strategy permits to improve the analysis precision by (accuracy=, F1-Score, precision, and recall).

In future, we will consider the Emojis when analyzing tweets to improve the performance of our system.

Acknowledgement

This work was supported by Ministry of National Education, Professional Training, Higher Education and Scientific Research (MENFPESRS) and the Digital Development Agency (DDA) of Morocco (Nos. Alkhawarizmi/2020/23).

[1] Smailhodzic E., Hooijsma W., Boonstra A., Langley D. J. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Services Research. **16**, 442 (2016).

[2] Rajani R., Berman D. S., Rozanski A. Social networks — are they good for your health? The era of Facebook and Twitter. QJM: An International Journal of Medicine. **104** (9), 819–820 (2011).

[3] Murray C. J. L., Lopez A. D., Wibulpolprasert S. Monitoring global health: time for new solutions. BMJ. **329**, 1096 (2004).

[4] Moorhead S. A., Hazlett D. E., Harrison L., Carroll J. K., Irwin A., Hoving C. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. Journal of Medical Internet Research. **15** (4), e85 (2013).

[5] Korda H., Itani Z. Harnessing Social Media for Health Promotion and Behavior Change. Health Promotion Practice. **14** (1), 15–23 (2013).

[6] Richardson C. R., Buis L. R., Janney A. W., Goodrich D. E., Sen A., Hess M. L., et al. An Online Community Improves Adherence in an Internet-Mediated Walking Program. Part 1: Results of a Randomized Controlled Trial. Journal of Medical Internet Research. **12** (4), e71 (2010).

[7] Diamond J. Diabetes in India. Nature. **469**, 478–479 (2011).

[8] Ho E. Y., Chesla C. A., Chun K. M. Health Communication With Chinese Americans About Type 2 Diabetes. The Science of Diabetes Self-Management and Care. **38** (1), 67–76 (2012).

[9] White R. O., Eden S., Wallston K. A., Kripalani S., Barto S., Shintani A., et al. Health communication, self-care, and treatment satisfaction among low-income diabetes patients in a public health setting. Patient Education and Counseling. **98** (2), 144–149 (2015).

[10] Haghravan S., Mohammadi-Nasrabadi F., Rafraf M. A critical review of national diabetes prevention and control programs in 12 countries in Middle East. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. **15** (1), 439–445 (2021).

[11] Kumar A., Goel M. K., Jain R. B., Khanna P., Chaudhary V. India towards diabetes control: Key issues. Australasian Medical Journal. **6** (10), 524–531 (2013).

[12] Lenoir P., Moulahi B., Azé J., Bringay S., Mercier G., Carbonnel F. Raising Awareness About Cervical Cancer Using Twitter: Content Analysis of the 2015 #SmearForSmear Campaign. Journal of Medical Internet Research. **19** (10), e344 (2017).

[13] Nisar S., Shafiq M. Framework for efficient utilisation of social media in Pakistan's healthcare sector. Technology in Society. **56**, 31–43 (2019).

[14] Diddi P., Lundy L. K. Organizational Twitter Use: Content Analysis of Tweets during Breast Cancer Awareness Month. Journal of Health Communication. **22** (3), 243–253 (2017).

[15] Von Muhlen M., Ohno-Machado L. Reviewing social media use by clinicians. Journal of the American Medical Informatics Association. **19** (5), 777–781 (2012).

[16] Alanzi T. Role of Social Media in Diabetes Management in the Middle East Region: Systematic Review. Journal of Medical Internet Research. **20** (2), e58 (2018).

[17] Elnaggar A., Ta Park V., Lee S. J., Bender M., Siegmund L. A., Park L. G. Patients' Use of Social Media for Diabetes Self-Care: Systematic Review. Journal of Medical Internet Research. **22** (4), e14209 (2020).

[18] Greene J. A., Choudhry N. K., Kilabuk E., Shrank W. H. Online Social Networking by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook. Journal of General Internal Medicine. **26**, 287–292 (2011).

[19] Stellefson M., Paige S., Apperson A., Spratt S. Social Media Content Analysis of Public Diabetes Facebook Groups. Journal of Diabetes Science and Technology. **13** (3), 428–438 (2019).

[20] Årsand E., Bradway M., Gabarron E. What Are Diabetes Patients Versus Health Care Personnel Discussing on Social Media? Journal of Diabetes Science and Technology. **13** (2), 198–205 (2019).

[21] Staite E., Zaremba N., Macdonald P., Allan J., Treasure J., Ismail K., Stadler M. 'Diabulima' through the lens of social media: a qualitative review and analysis of online blogs by people with Type 1 diabetes mellitus and eating disorders. Diabetic Medicine. **35**, 1329–1336 (2018).

[22] Karami A., Dahl A. A., Turner-McGrievy G., Kharrazi H., Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. International Journal of Information Management. **38** (1), 1–6 (2018).

[23] Shaw G., Karami A. Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise. Proceedings of the Association for Information Science and Technology. **54** (1), 357–365 (2017).

[24] Liu Y., Mei Q., Hanauer D. A., Zheng K., Lee J. M. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. JMIR Diabetes. **1** (2), e4 (2016).

[25] Patel K. D., Zainab K., Heppner A., Srivastava G., Mago V. Using Twitter for diabetes community analysis. Network Modeling Analysis in Health Informatics and Bioinformatics. **9**, 36 (2020).

[26] Patel K. D., Heppner A., Srivastava G., Mago V. Analyzing use of Twitter by diabetes online community. ASONAM'19: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 937–944 (2019).

[27] Salas-Zárate M. D. P., Medina-Moreira J., Lagos-Ortiz K., Luna-Aveiga H., Rodríguez-García M. Á., Valencia-García R. Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computational and Mathematical Methods in Medicine. **2017**, 5140631 (2017).

[28] Gabarron E., Dorronzoro E., Rivera-Romero O., Wynn R. Diabetes on Twitter: A Sentiment Analysis. Journal of Diabetes Science and Technology. **13** (3), 439–444 (2018).

[29] Hong L., Ahmed A., Gurumurthy S., Smola A., Tsioutsiouliklis K. Discovering geographical topics in the twitter stream. WWW'12: Proceedings of the 21st international conference on World Wide Web. 769–778 (2012).

[30] Raamkumar A. S., Pang N., Foo S. When countries become the talking point in microblogs: Study on country hashtags in Twitter. First Monday. **21** (1), 1–4 (2016).

[31] Alhabash S., Ma M. A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students? Social Media + Society. **3** (1), 1–13 (2017).

[32] King D., Ramirez-Cano D., Greaves F., Vlaev I., Beales S., Darzi A. Twitter and the health reforms in the English national health service. Health Policy. **110** (2–3), 291–297 (2013).

[33] Bounabi M., El Moutaouakil K., Satori K. The Optimal Inference Rules Selection for Unstructured Data Multi-Classification. Statistics, Optimization & Information Computing. **10** (1), 225–235 (2022).

[34] El Moutaouakil K., Ahourag A., Chellak S., Baïzri H., Cheggour M. Fuzzy Deep Daily Nutrients Requirements Representation. Revue d'Intelligence Artificielle. **36** (2), 263–269 (2022).

[35] El Moutaouakil K., Saliha C., Chellak S. Optimal fuzzy deep daily nutrients requirements representation: Application to optimal Morocco diet problem. Mathematical Modeling and Computing. **9** (3), 607–615 (2022).

[36] El Moutaouakil K., Ahourag A., Chakir S., Kabbaj Z., Chellack S., Cheggour M., Baizri H. Hybrid firefly genetic algorithm and integral fuzzy quadratic programming to an optimal Moroccan diet. Mathematical Modeling and Computing. **10** (2), 338–350 (2023).

[37] El Ouissari A., El Moutaouakil K. Density based fuzzy support vector machine: application to diabetes dataset. Mathematical Modeling and Computing. **8** (4), 747–760 (2020).

[38] El Moutaouakil K., Roudani M., El Ouissari A. Optimal Entropy Genetic Fuzzy-C-Means SMOTE (OEGFCM-SMOTE). Knowledge-Based Systems. **262**, 110235 (2023).

[39] El Moutaouakil K., Palade V., Safouan S., Charroud A. FP-Conv-CM: Fuzzy Probabilistic Convolution C-Means. Mathematics. **11** (8), 1931 (2023).

[40] El Moutaouakil K., El Ouissari A., Hicham B., Saliha C., Cheggour M. Multi-objectives optimization and convolution fuzzy C-means: Control of diabetic population dynamic. RAIRO-Operations Research. **56** (2), 3245–3256 (2022).

[41] Wang Y., Pan Z., Dong J. A new two-layer nearest neighbor selection method for kNN classifier. Knowledge-Based Systems. **235**, 107604 (2022).

[42] Choubey D. K., Kumar M., Shukla V., Tripathi S., Dhandhania V. K. Comparative analysis of classification methods with PCA and LDA for diabetes. Current Diabetes Reviews. **16** (8), 833–850 (2020).

[43] Saritas M. M., Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering. **7** (2), 88–91 (2019).

[44] Chen S., Webb G. I., Liu L., Ma X. A novel selective naïve Bayes algorithm. Knowledge-Based Systems. **192**, 105361 (2020).

Аналіз настрою марокканського діабетика в Twitter з використанням нечітких C-середніх SMOTE та глибокої нейронної мережі

Рудані М.[1], Елькарі Б.[2], Ель Мутауакіл К.[1], Ураба Л.[2], Хічам Б.[3], Челлак С.[3]

[1] *Лабораторія інженерних наук (LSI), Полідисциплінарний факультет Тази, USMBA, Марокко*
[2] *EIDIA, Дослідницький центр Euromed, Євро-Мед університет (UEMF), Фес, Марокко*
[3] *Факультет медицини та фармації Університет Каді Айяд, Сіді Аббад, Марракеш, Марокко*

Ефективне керування діабетом як станом способу життя передбачає підвищення обізнаності, а соціальні мережі є потужним інструментом для цієї мети. Аналіз вмісту твітів на таких платформах, як Twitter, може значно допомогти розробити комунікаційні стратегії щодо здоров'я, спрямовані на підвищення обізнаності марокканської спільноти про діабет. На жаль, корпус твітів незбалансований, і вилучення ознак призводить до наборів даних із дуже високою розмірністю, що впливає на якість аналізу настроїв. Це дослідження було зосереджено на аналізі змісту, настроїв і охоплення твітів, які стосувалися діабету в Марокко. Запропонована стратегія містить п'ять етапів: (a) збір даних із платформ Twitter і ручна лабілізація, (b) виділення ознак за допомогою техніки TF-IDF, (c) зменшення розмірності за допомогою глибокої нейронної мережі, (d) балансування даних за допомогою нечітких C-середніх SMOTE і (e) класифікація твітів за допомогою п'яти добре відомих класифікаторів. Запропонований підхід порівняно з класичною системою, яка працює безпосередньо з дуже великими, незбалансованими твітами. З точки зору повноти, точності, оцінки F1 і процесорного часу, запропонована система може виконувати високоточний аналіз настрою за прийнятний процесорний час.

Ключові слова: *діабет; нечіткі C-середні; SMOTE; глибока нейронна мережа; аналіз настроїв; класифікація.*