

Hybrid approach for tweets similarity classification founded on case based reasoning and machine learning techniques

Ismail Bensassi¹, Mohamed Kouissi², Oussama Ndama¹, El Mokhtar En-Naimi¹, Abdelhamid Zouhair¹

¹DSAI2S Research Team, Faculty of Sciences and Technologies of Tangier, Abdelmalek Essaâdi University, Tetouan, Morocco

²DSAI2S Research Team, La Faculté Polydisciplinaire de Larache (FP de Larache), Abdelmalek Essaâdi University, Tetouan, Morocco

Article Info

Article history:

Received Mar 9, 2024

Revised Oct 1, 2024

Accepted Oct 17, 2024

Keywords:

Dynamic case based reasoning

Machine learning

Multi agents system

Term frequency-inverse

document frequency

Tweets similarity classification

ABSTRACT

Twitter sentiment analysis becomes a popular research subject in the last decade. It aims to extract sentiments of users through their public opinion about a given topic. This article proposes a hybrid approach for Twitter sentiment analysis founded on dynamic case based reasoning (DCBR), multinomial logistic regression machine learning algorithm and multi-agent system. Our approach proposes a method to find similar tweets based on content similarity measure using the scientific measurement of keyword weight term frequency-inverse document frequency (TF-IDF). This approach includes gathering and pre-processing tweets, getting score and polarity of tweets, the use of multinomial logistic regression machine learning algorithm to classify our tweets into various classes, using the feature extraction method to extract useful features and then the K-nearest neighbors (KNN) algorithm to make it easier to find similar tweets to our tweet target case. This approach is adaptive and generic and able to track users' tweet to predict their behavior and sentiments in critical situations and delivering personalized content. The current study focuses on Covid-19 tweets, and a public Twitter dataset is used for this purpose.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ismail Bensassi

DSAI2S Research Team, Faculty of Sciences and Technologies of Tangier, Abdelmalek Essaâdi University
Tetouan, Morocco

Email: bensassi.ismail@gmail.com

1. INTRODUCTION

Social network plays an important role for most people, it becomes an indispensable part for human interactions. It represents a relevant way for expressing opinions, thoughts, and sharing more personal emotions and sentiments about various topics as stated in [1]-[4]. Twitter is a widely used social networking platform that generates a significant amount of data from tweets.

In recent years, several studies have been done on Twitter sentiment analyses using big social data by gathering and classifying users' opinions on a topic. Those studies encompass many disciplines, like Covid-19, Covid-19 vaccine, and elections commercial activities. Twitter sentiment analyses is a process which determines the sentiment orientation of a text. The main idea of Twitter sentiment analysis becomes a question of whether a tweet is expressing positive, negative, or neutral towards the discussed subject. During Covid-19 pandemic, an increasing number of people used twitter platform to share their feelings with others [5]. So, feeling analysis has become a frequent research topic [6]. To help decision makers analyze users' opinions and their reactions related to tweets content, and to predict their behavior and sentiments based on past experiences, we propose then a hybrid approach for Twitter sentiment analysis based on dynamic case based reasoning (DCBR), machine learning algorithms, natural language processing, and

multi-agent system. This approach proposes an adaptive system for sentiment analysis classification to ensure a personalized follow-up of users in critical situations.

Several studies have been focused on the analysis of data from social networking platform, in particular those related to significant events. Many researchers have been realized to get relevant information about these events. Sentiment analysis on Twitter has become one of the most popular fields of study in recent years. Imamah and Rachman [7] proposed a solution for sentiment analysis about mental health, caused by Covid-19 disease, through public opinion on Twitter. They used a dataset for Covid-19 tweets and has been classified with logistic regression method and term frequency-inverse document frequency (TF-IDF). The Covid-19 tweets sentiment classification reached an accuracy of 94.71%. Saad and Yang [8] has proposed an approach for tweets sentiment analysis based on ordinal regression. This approach aims to extricate efficient feature. Multiple machine learning algorithms have been used for this purpose. Shofiya and Abidi [9] shown an examination of people sentiment analysis on Covid-19 concerning social distancing in Canada. They used a tool to analyze and extract sentiment polarity of tweets, then a support vector machine algorithm has been used for tweets sentiment classification.

In the last decade, several works have been conducted in the field of Twitter sentiment analysis classification based on machine learning, but mainly, all the earlier studies focused on transforming those problems into an ordinal regression or classification problem. Whereas, to extract and predict users' sentiments through their public opinion about a specific topic or event, it would be more interesting to follow-up the users' tweets traces and sentiment hidden behind them and provide them with an individualized content based on past experiences of another users. To solve this problem, we combine the dynamic case-based reasoning approach, machine learning algorithm and the TF-IDF method. With DCBR, we can solve new problems by reusing the applied solutions of previous problems. It consists in retrieving similar tweets based on similarity measures. Note that, our approach is also applicable to several fields where DCBR is requested.

To fill this gap in the research, we performed our analysis on a Covid-19 dataset with one million (1 M) tweets to understand users' behaviors. Using Twitter sentiment analysis, we can process tweets, check people feelings and assess them during this challenging period. Therefore, to retrieve similar tweets' content using the DCBR approach, that aims to reuse similar past experiences, will help us to track users' tweet to predict their behavior and sentiments in critical situations and delivering personalized content.

This paper is arranged as follows: the section 1, presents related studies about Twitter sentiment analysis classification using machine learning algorithms. Section 2 is devoted to the literature review. In section 3, we explain our approach and our research method. Section 4 presents the results of this approach. A conclusion with some recommendations is drawn in section 5.

2. LITERATURE REVIEW

2.1. Term frequency-inverse document frequency

To construct the classifier model we should first extricate relevant features from the collected tweets. There are many techniques for this purpose like bag of words (BoW), Word2Vec and BERT. This study uses TF-IDF as stated in [10] to extricate relevant features. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). TF-IDF vectorizer retrieves features according to word representation count, frequent words will be assigned less weight and rare words will be assigned more weight.

2.2. Dynamic case based reasoning

Case based reasoning is as a field of artificial intelligence that can be dated back in the late 1980s. CBR is a paradigm that solves new problems by relying on the solutions of previous problems with the same nature. The current problem to solve is called target case and the problem that has been already solved is called source case. Aamodt and Plaza [11] are the first authors who described the CBR cycle. According to these authors, CBR consists of four phases, while others, the cycle of CBR is only on three phases [12]. Currently, the most adapted and used model for CBR is the five-step model which is an extended version of the old models. In this study, we exploit the five-step CBR cycle: elaboration, retrieve, reuse, revision, and retain as illustrated in Figure 1. Those steps are more detailed here [13].

In DCBR, the classical cycle has been adapted by changing the order on the concerned steps. DCBR is continuous and taking into consideration the dynamic change of the descriptors describing the new problem. In DCBR, some steps can be re-executed if there is a change in the specifications of the new problem, certain steps can be stopped, and others can be repeated more than once [13].

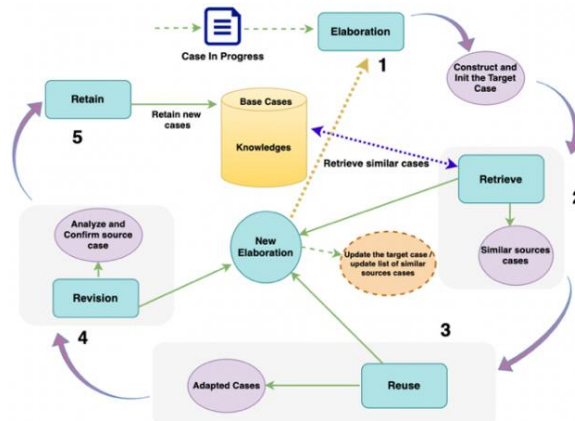


Figure 1. Dynamic case based reasoning cycle inspired from [13]

2.3. Multinomial logistic regression

Multinomial logistic regression is a supervised machine learning algorithm. It is also known as SoftMax regression or multinomial regression analysis [14]. It is used to analyze relationships between multiple independent variables and a categorical dependent variable with more than two categories. In other words, it is used to predict the probabilities of an outcome belonging to each category of a dependent variable based on the values of several independent variables. MLR can classify new data using continuous and discrete datasets. It is a generalization of binary logistic regression, which is used for modeling the relationship between a binary (two-category) dependent variable and one or more independent variables.

3. METHOD

Our proposed system architecture is basically composed of three principal modules. This hybrid approach integrates the benefits of multi-agent systems and DCBR techniques. The first module acquires new data and responsible for data preprocessing, which is a crucial process to clean and prepare the text data for analysis [15]. The second module is responsible of creating a classification model using sentiment analysis, which involves analyzing the emotional tone of a tweet. In this module, we extract useful characteristics and create a balancing and scoring method. The third module consists of applying multinomial logistic regression classifier that classify users' tweets into five clusters of sentiment (high negative, moderate negative, neutral, moderate positive and highly positive) [16] to find the category where our target case belongs to, and then predict in this category, the most similar tweets based on content similarity measure using TF-IDF and then their sentiment to provide users with personalized content. Our proposed system requires the use of multi-agent architecture based on DCBR cycle [17], as you can see in Figure 2. Algorithm 1 displays the overall steps used for the classification to find similar tweets of our tweet target case.

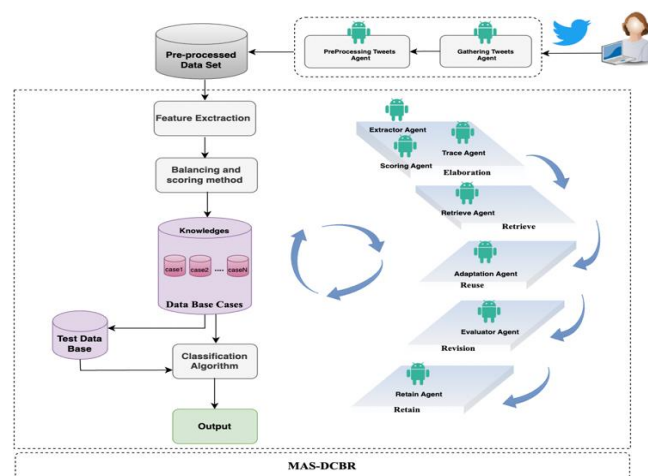


Figure 2. Proposed our MAS and DCBR system architecture

Algorithm 1: Tweet similarity classification

Begin:

Var **InputQuery**: String

Retrieve twitter data

Procedure PreProcessing:

for each tweet in range (len(twitter dataset)) :

Pre-processing(tweet). // syntactic correction of the tweets

Extract features using TF-IDF.

Calculate the polarity and the score of each tweet.

Feed the data base cases with source cases from Twitter.

End Procedure

Procedure Classification

Classify tweets using Multinomial Logistic regression into several ordinal categories.

Find the category where our tweet target case belongs to.

Find similar tweets of our tweet target case in this category using KNN algorithm.

End Procedure

End

3.1. Dynamic case based reasoning in our proposed system model

In this section, we will explain the use of DCBR in our proposed system architecture shown in Figure 2. Case based reasoning approach mainly include five steps, which are case representation, retrieval, reuse, revision and finally retain.

3.1.1. Case representation

Case representation refers to the way in which a case is stored and organized in a case-based reasoning system. It includes information such as the problem description, the process solution, and any relevant contextual details [18]. The case representation should be clear, concise, and structured in a way that makes it easy to retrieve and compare with new problems. Before starting our DCBR cycle, the retrieved data should be first cleaned and structured, so the new cases will be effectively described, and the retrieval case will be properly executed. The case representation is described in vector-based representations. Each case is usually consisting of two parts, the problem description part which is represented by a set of descriptors of the problem to be solved, and the recommended solution part which is represented by a set of steps or descriptors of the solution provided by the reasoning. The solution part provides steps that can be implemented by decision-maker.

$$Case_i = (Problem_i, Solution(Problem_i)) \quad (1)$$

$Problem_i = \{d^s_1 \dots d^s_n\}$ where d^s_j represents the descriptor of the source problem.

$Solution(Problem_i) = \{D^s_1 \dots D^s_m\}$ where D^s_j represents the descriptor of the recommended solution.

The descriptors representations used are vector type representations (attributes, values). The objective of this stage of reasoning is to put in place the different mechanisms to move from an often poorly expressed problem to a well-defined problem to facilitate the following stages. To extract relevant features from the tweets, the IDF is used to weight the features of the case representation. The idea is to assign more weight to the features that are more discriminatory, while giving less weight to the features that are less discriminatory. By using IDF to weight the features, the CBR system can better distinguish between relevant and irrelevant cases and improve the retrieval and reuse process.

3.1.2. Case retrieval

In CBR approach, the case retrieval represents the core step. After the elaboration of the target problem, the case retrieval consists in retrieving similar cases to the target case [19]. A good case is obviously one where its solution represents a good candidate that will allow us to solve the current problem. The search in this step is based on similarity measures, where we compare the descriptors of the problem part of the target case with the source cases stored in the data base case. The most classical approach for similarity measures is to make a weighted sum of criteria on the case descriptor attributes, but generally it depends on the nature of the attributes describing the cases. There are also few metrics such as geometric similarity metrics, Euclidean distance, Mahalanobis distance or KNN. However, this paper uses the KNN algorithm for similarity measure. It is used to classify new objects by calculating the distance between the attributes describing the cases.

3.1.3. Case reuse

In the reuse step of CBR process, the solution from a retrieved similar case is applied to the current problem. During this step, the solution is adapted to fit the specific needs of the new problem [20]. We calculate the similarity between the new case and other cases in the data base case. If the similarity measure is equal to 1, that means, the source case is the same as the new target case and can be directly reused, else, we sort the calculated similarity values from the biggest to the smallest value, then we select the nearest cases to the target case.

In some cases, the solution from the retrieved source case may need only minor modifications to be used for the new target problem. In other cases, the solution from the retrieved case may need to be significantly modified to be used for the new problem. Once the solution has been adapted to fit the new problem, it is evaluated for its effectiveness. If the solution is found to be effective, it is used to solve the new problem. If the solution is found to be ineffective, the CBR system will look for another similar case and repeat the process.

3.1.4. Case revision

In the revision step, we are supposed to give feedback on the proposed solution. This step allows the case-based reasoning system to improve over time, as it incorporates new information and experiences into its data base case. The revision step can involve updating the description or attributes of a case, adding a new case, or discarding a case if it is no longer useful. Finally, after the new target case solution has been tested and verified based on an expert system, real world or simulation and correctness, the retain case is executed.

3.1.5. Case retain

The retain step in case-based reasoning involves storing a new case in the case base with its related problem and solution, typically after a solution has been generated and possibly revised. The retain step allows the CBR system to build up its knowledge over time, and to use that knowledge to solve new problems in the future. The new target case may be stored as a record of the problem, its solution, and any additional information or context that may be relevant. The retain step is important for ensuring that the CBR system becomes richer and richer by adding new and revised target cases and continues to grow and improve, and for maintaining the quality and relevance of the data base case.

4. RESULTS AND DISCUSSION

In this section, we present the obtained results of our approach for Twitter sentiment analysis. The originality of this paper is demonstrated by identifying the general feeling of Twitter users towards Covid-19 tweets and looking for similar tweets according to their content. The KNN algorithm has been used to run different tests to find the nearest similar tweets' content of our tweet target case.

4.1. Data collection

The data being used has been collected using the Twitter API, which contains 1 M tweets. The data was collected in the form of a CSV file with several fields. In this study we require only the username, text of the tweet and the related hashtags, the rest is discarded. Each data tweet will be defined as a vector Casei of a set of characteristics (username, tweet, hash tags, and the related solution of each problem) as mentioned in (1).

4.2. Tweets preprocessing

The collected tweets contain a lot of noise and redundant information. The data preprocessing is a crucial step, and it involves syntactic correction of the tweets as required. It aims to transform data into a better form to feed the tweet data to a machine learning algorithm, to extract valuable insights and minimize ambiguity in the feature extraction process. Many steps are used for tweet preprocessing by replacing URLs, usernames and emojis with their corresponding sentiment, removing repeated letters, converting upper case to lower case and finally applying the stemming method by replacing words with their root [21]-[23]. Table 1 displays an example results of the tweets preprocessing steps.

Table 1. Example of tweets preprocessing

Tweets	HashTags
No one will be safe from Covid-19 until everyone is safe will you commit to ensure	#COVID19
Let all protect ourselves from Covid-19 it real and the number are climbing up fast in the continent	#COVID19
Nagaland police on Covid-19 awareness at city tower junction dimapur Covid-19 keep social distance	#COVID19
Can imagine the same people profiting off the human suffering of Covid-19 will be studying these map to make 207	#COVID19

4.3. Splitting the data

After the pre-processing steps have been performed, we split our dataset into training and testing set for a better evaluation of the model. By evaluating the classifier model on data that it has not seen during training phase, we can get an estimate of its performance on new, unseen data. We allocated 70% of the data for the training and 30% for the testing.

4.4. Feature extraction

A feature refers to an individual measurable property of the data being analyzed. Features are the input variables that are used by machine learning models to make predictions or classifications [24]. Feature extraction refers to the process of selecting and transforming raw data into a set of relevant features that can be used for machine learning and statistical analysis. In machine learning, the features extracted from the data are used as input to a learning algorithm to build a predictive model. Feature extraction is often used to reduce the dimensionality of the data and improving the accuracy and efficiency of the learning algorithms. In this paper, we used TF-IDF to extricate 10000 relevant features matrix from the tweets. Table 2 presents the matrix of TF-IDF features.

Table 2. Matrix of TF-IDF features

Casei	Coronavirus	Covid-19	Covid	Help	Case	Hospital
0	0.0	0.057475	0.0	0.0	0.0	0.0
1	0.0	0.050703	0.0	0.0	0.0	0.0
2	0.420171	0.127932	0.0	0.0	0.0	0.0
3	0.0	0.104732	0.366245	0.0	0.0	0.0
6	0.0	0.052958	0.0	0.0	0.157485	0.0

To build a classifier model for Twitter sentiment analysis to classify the tweets according to their polarity, a scoring method is proposed [25]. It consists of:

- Each tweet is classified as high negative, moderate negative, neutral, moderate positive or high positive according to the count of positive or negative terms in the given tweet.
- Tweet polarity is calculated using natural language processing.
- The score is then assigned according to the tweet polarity. 1 for highly negative tweets, 2 for moderate negative, 3 for neutral, 4 for moderate positive, and 5 for highly positive ones.

4.5. Classification using TF-IDF, multinomial logistic regression and KNN

After processing, splitting the dataset into training and testing dataset, extracting features matrix using TF-IDF [26], [27], and calculating the overall polarity score [28] of each tweet, we retrain our model using the multinomial logistic regression to classify our tweets into several classes (high negative, moderate negative, neutral, moderate positive, high positive) as labels in order to predict a sentiment category for a given target tweet. The evaluation of the model is necessary to understand the performance metrics of the proposed approach for Covid-19 tweets classification. Our classification model reached an accuracy of 88% as shown in Figure 3. The source case tweet of our target case is “Covid is very bad for our health”. Table 3 illustrates the results of the similar tweets using the KNN algorithm.

	precision	recall	f1-score	support
1	0.79	0.36	0.50	715
2	0.86	0.75	0.80	4415
3	0.89	0.99	0.94	12459
4	0.91	0.89	0.90	10351
5	0.83	0.73	0.78	2436
accuracy			0.88	30376
macro avg	0.86	0.74	0.78	30376
weighted avg	0.88	0.88	0.88	30376

Figure 3. Model evaluation of Covid-19 tweets

Table 3. Similar tweets using KNN algorithm

	Similar tweets using the model after applying the KNN algorithm	Distance
Case 80332	I smell something bad brewing for our little country!!! 😞😞😞😞😞😞😞 #Covid19	1.2
Case 65535	This news is very very bad. Mistakes by @BorisJohnson allowed #COVID19 to spread and caused further economic damage	1.21
Case 30565	his is very bad news for the #cruise industry. Hurtigruten crew #COVID19 numbers increase significantly	1.235
Case 60572	This #COVID19 is not good for our older peeps. 😞😞😞	1.249
Case 34303	One of the reasons #COVID19 is very bad in the US is due to the political intonation put into it. US is as a develo...	1.250
Case 92967	How increased screen time during Covid is affecting your mental health #ExerciseIsMedicine #COVID19	1.251

5. CONCLUSION

In this study, we proposed a hybrid approach for Twitter sentiment analysis. This approach proposes an adaptive system for sentiment analysis classification to ensure a personalized follow-up of users in critical situations. It aims to extract relevant features matrix from the tweets using TF-IDF, building a balancing and scoring model, and then feeding the data to a machine learning model to categorize our tweets into several classes and applying the KNN algorithm in the retrieve phase of our DCBR cycle to make it easier to find similar tweets to our tweet target case. The tweets are classified into several ordinal categories with similar tweets sentiment, and then the KNN algorithm is applied to retrieve the nearest similar tweets of the tweet target case, on the category, that the tweet target case belongs to. In this approach we used a public Twitter dataset consisting of 1 M tweets.

Some major constraints and limitations of this study and further research to enhance this approach are discussed. The proposed approach can be improved from several angles. Firstly, the public Twitter dataset is limited to english tweets only. It would be useful to acquire more real-world tweets from different locations, to evaluate the effectiveness of the proposed approach in order to build a robust classifier model. Another limitation is the use of the scientific measurement of keyword weight TF-IDF which cannot understand the context of the words and derive their meaning, the users in different locations may behave with different characteristics, then it will be difficult to follow their tweets traces and predict their behaviors. As such, it is necessary to conduct further evaluation with different algorithms for text classification. So, future work includes the improvement of our classification model for tweet similarity measure. Furthermore, we intend to test different machine learning algorithms to obtain better classification accuracy.




REFERENCES

- [1] D. M. Daoud and S. A. El-Seoud, "Building a Sentiment Analysis System Using Automatically Generated Training Dataset," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 06, p. 48, May 2020, doi: 10.3991/ijoe.v16i06.13623.
- [2] M. Alam, F. Abid, C. Guangpei, and L. V. Yunrong, "Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications," *Computer Communications*, vol. 154, pp. 129–137, Mar. 2020, doi: 10.1016/j.comcom.2020.02.044.
- [3] G. A. de Oliveira, R. T. de Sousa, R. de O. Albuquerque, and L. J. G. Villalba, "Adversarial attacks on a lexical sentiment analysis classifier," *Computer Communications*, vol. 174, pp. 154–171, Jun. 2021, doi: 10.1016/j.comcom.2021.04.026.
- [4] A. M. Alnasrawi, A. M. N. Alzubaidi, and A. A. Al-Moadhen, "Improving sentiment analysis using text network features within different machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 405–412, Feb. 2024, doi: 10.11591/eei.v13i1.5576.
- [5] A. J and J. G., "Sentiment Classification of Tweets with Non-Language Features," *Procedia Computer Science*, vol. 143, pp. 426–433, 2018, doi: 10.1016/j.procs.2018.10.414.
- [6] R. Alegre-Veliz *et al.*, "Machine Learning for Feeling Analysis in Twitter Communications: A Case Study in HEYDRU!, Perú," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 24, pp. 126–142, Dec. 2022, doi: 10.3991/ijim.v16i24.35493.
- [7] Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regression," *2020 6th Information Technology International Seminar (ITIS)*, Oct. 2020, doi: 10.1109/itis50118.2020.9320958.
- [8] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," *IEEE Access*, vol. 7, pp. 163677–163685, Nov. 2019, doi: 10.1109/access.2019.2952127.
- [9] C. Shofiya and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5993, Jun. 2021, doi: 10.3390/ijerph18115993.
- [10] L. Almazaydeh, M. Abuhaleleh, A. Al Tawil, and K. Elleithy, "Clinical Text Classification with Word Representation Features and Machine Learning Algorithms," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 65–76, Apr. 2023, doi: 10.3991/ijoe.v19i04.36099.
- [11] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994, doi: 10.3233/aic-1994-7104.
- [12] A. Cordier, B. Fuchs, and A. Mille, "Engineering and Learning of Adaptation Knowledge in Case-Based Reasoning," *Lecture notes in computer science*, pp. 303–317, Jan. 2006, doi: 10.1007/11891451_27.




- [13] E. M. En-Naimi, and A. Zouhair, "Intelligent dynamic case-based reasoning using multi-agents system in adaptive e-service, e-commerce and e-learning systems," *International Journal of Knowledge and Learning*, vol. 11, no. 1, pp.42–57, Aug. 2016, doi: 10.1504/IJKL.2016.078652.
- [14] K. Kayabol, "Approximate Sparse Multinomial Logistic Regression for Classification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 490–493, Feb. 2020, doi: 10.1109/TPAMI.2019.2904062.
- [15] E. Sutoyo and A. Almaarif, "Twitter sentiment analysis of the relocation of Indonesia's capital city," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1620–1630, Aug. 2020, doi: 10.11591/eei.v9i4.2352.
- [16] C. Aloysius, P. T. Selvan, "Reduction of false negatives in multi-class sentiment analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 1209–1218, Apr. 2023, doi: 10.11591/eei.v12i2.4682.
- [17] N. El Ghouch, E. M. En-Naimi, and M. Kouissi, "Implementation of an Adaptive Learning System based on Agents and Web Services," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 14, p. 162, Jul. 2020, doi: 10.3991/ijet.v15i14.13837.
- [18] Z. Jiang, Y. Jiang, Y. Wang, H. Zhang, H. Cao, and G. Tian, "A hybrid approach of rough set and case-based reasoning to remanufacturing process planning," *Journal of Intelligent Manufacturing*, vol. 30, no. 1, pp. 19–32, Jun. 2016, doi: 10.1007/s10845-016-1231-0.
- [19] W. Gu, A. Moustafa, T. Ito, M. Zhang, and C. Yang, "A Case-based Reasoning Approach for Supporting Facilitation in Online Discussions," *Group Decision and Negotiation*, vol. 30, no. 3, pp. 719–742, Mar. 2021, doi: 10.1007/s10726-021-09731-4.
- [20] S. Demigha, "Computational Methods and Techniques for Case-Based Reasoning (CBR)," *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2020, doi: 10.1109/csci51800.2020.00264.
- [21] A. Addiga and S. Bagui, "Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency," *Journal of Computer and Communications*, vol. 10, no. 08, pp. 117–128, Aug. 2022, doi: 10.4236/jcc.2022.108008.
- [22] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment," *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 52–57, May 2016, doi: 10.1109/BigDataService.2016.36.
- [23] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 05, pp. 42–61, Mar. 2023, doi: 10.3991/ijet.v18i05.35959.
- [24] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python," *International Journal of Computer Applications*, vol. 165, no. 9, pp. 29–34, May 2017, doi: 10.5120/ijca2017914022.
- [25] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019*, pp. 13–15, Mar. 2019.
- [26] H. K. Obayes, K. H. Alhussayni, and S. M. Hussain, "Predicting COVID-19 vaccinators based on machine learning and sentiment analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1648–1656, Jun. 2023, doi: 10.11591/eei.v12i3.4278.
- [27] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.
- [28] S. Bengesi, T. Oladunni, R. Olusegun and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," in *IEEE Access*, vol. 11, pp. 11811–11826, 2023, doi: 10.1109/ACCESS.2023.3242290.

BIOGRAPHIES OF AUTHORS






Ismail Bensassi    is a Ph.D. student in DSAI2S (Data Science, Artificial Intelligence and Smart Systems) Research Team, C3S (Computer Science and Smart Systems) Laboratory, Faculty of Sciences and Technologies (FST), Tangier, Morocco. He is an engineer in Computer Science, Laureate of FST of Tangier, he is a computer systems analyst engineer at the Ministry of Justice, Department of Modernization and Computer Systems, Rabat, Morocco. The research topics of interest are smart systems, sentiment analysis, smart cities, multi-agent systems (MAS), case based reasoning (CBR), ontology, machine learning, and deep learning. He can be contacted at email: bensassi.ismail@gmail.com.






Mohamed Kouissi    is an Associate Professor at the Polydisciplinary Faculty of Larache, University of Abdelmalek Essaâdi (UAE). Membre in DSAI2S (Data Science, Artificial Intelligence and Smart Systems) Research Team, C3S (Computer Science and Smart Systems) Laboratory, Faculty of Sciences and Technologies, Tangier, Morocco. He is an engineer in computer science and specialized in software mobile development. Laureate of Tangier National School of Applied Sciences in 2011. The research topics of interest are multi-agent systems (MAS), case based reasoning (CBR), machine learning, smart cities, and elearning. He can be contacted at email: m.kouissi@uae.ac.ma.






Oussama Ndama    is a Ph.D. student in DSAI2S (Data Science, Artificial Intelligence and Smart Systems) Research Team, C3S (Computer Science and Smart Systems) Laboratory, Faculty of Sciences and Technologies (FST), Tangier, Morocco. He had his Master in Computer Science and Big Data, Laureate of FST of Tangier. He is also a business intelligence engineer with more than 5 years of experience in different multinational companies. The research topics of interest are smart systems, machine learning, deep learning, NLP, ANN, sentiment analysis, and smart cities. He can be contacted at email: oussama.ndama@etu.uae.ac.ma.



Dr. El Mokhtar En-Naimi    is a Full Professor in the University of Abdelmalek Essaâdi (UAE), Faculty of Sciences and Technologies of Tangier (FSTT), Department of Computer Sciences. (He was Temporary Professor: from 1999 to 2003 and Permanent Professor: since 2003/2004 until now. Actually, He is a Full Professor in UAE, FST of Tangier). He was a Head of Computer Sciences Department, since October 2016 until the end of December 2020. He was responsible for a Licence of Science and Technology, LST Computer Engineering ("Licence LST-GI"), from January 2012 to October 2016. He is a chief of DSAI2S (Data Science, Artificial Intelligence and Smart Systems) Research Team, since the academic year 2022/2023. He is also a founding member of the both laboratories: LIST (Laboratoire d'Informatique, Systèmes et Télécommunications) Laboratory (From 2008 To 2022) and C3S (Computer Science and Smart Systems) Laboratory since the academic year 2022/2023 until now, the University of Abdelmalek Essaâdi, FST of Tangier, Morocco. He is also an expert evaluator with the ANEAQ, since the academic year 2016/2017 until now, that an expert of the private establishments belonging to the territory of the UAE and also an expert of the Initial or Fundamental Formations and Formations Continuous at the Ministry of Higher Education, Scientific Research and Executive Training and also at the UAE University and the FST Tangier since 2012/2013 until now. He is an Author/Co-Authors of several articles, published in The International Journals in Computer Sciences, in particular, in multi-agent systems (MAS), cases based reasoning (CBR), artificial intelligent (AI), machine learning (ML), deep learning (DL), eLearning, MOOC/SPOC, big data, data-mining, wireless sensor network, VANet, MANet, and smart city. He was/is also Director of several Doctoral Theses in Computer Sciences. He has too served as a general chair, technical program chair, technical program committee member, organizing committee member, session chair, and reviewer for many international conferences and workshops. In addition, he is an associate member of the ISCN-Institute of Complex Systems in Normandy, the University of the Havre, France, since 2009 until now. He can be contacted at email: en-naimi@uae.ac.ma.



Dr. Abdelhamid Zouhair    is a Professor in the University of Abdelmalek Essaâdi, Faculty of Sciences and Technologies of Tangier (FST), Department of Computer Science, since mars 2020. Since March 2012-10 May 2016: quality project manager, geographic information system and computer head at the Urban Agency of Tetouan, Ministry of Urban Planning and Development. August 2011-March 2012: Senior Executive at the Urban Agency of Tetouan. July 2003-September 2008: IT Manager at Tronico Atlas, Tangier, Morocco. Ph.D. in Computer Science at the University of The Havre, laboratory LITIS, France, and at the FST of Tangier, Morocco (cotutelle doctoral program) in October 2014. He is an author of several articles in computer science. He can be contacted at email: Zouhair07@gmail.com.