Social Media's Toxic Comments Detection Using Artificial Intelligence Techniques

Rabia Rachidi
LaROSERI Laboratory
Faculty of Science,
Chouaib Doukali University,
El Jadida, Morocco
rachidirabia99@gmail.com

Mohamed Amine Ouassil
EEIS Laboratory
ENSET of Mohammedia,
Hassan II University of Casablanca,
Mohammedia, Morocco
ouassil.amine@gmail.com

Mouaad Errami
EEIS Laboratory
ENSET of Mohammedia,
Hassan II University of Casablanca,
Mohammedia, Morocco
mouaad.errami@gmail.com

Bouchaib Cherradi
STIE Team, CRMEF Casablanca-Settat.
EEIS Laboratory, ENSET of Mohammedia,
Hassan II University of Casablanca,
Mohammedia, Morocco
bouchaib.cherradi@gmail.com

Soufiane Hamida
GENIUS Laboratory, SupMTI of Rabat.
EEIS Laboratory, ENSET of Mohammedia
Hassan II University of Casablanca
Mohammedia, Morocco
hamida.93s@gmail.com

Hassan Silkan
LaROSERI Laboratory
Faculty of Science,
Chouaib Doukali University,
El Jadida, Morocco
silkan_h@yahoo.fr

2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) | 979-8-3503-9836-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IRASET57153.2023.10153015

Abstract—**Cyberbullying takes its place in social media and has increased throughout the past few years. The damage that cyberbullying has on the users is undeniable they get attacked either on their appearances, ethnicities, religions, and even their thoughts and personal opinion. The attack causes these users anxiety, depression, low self-esteem, and in the worst scenarios suicide. These harmful actions toward the users drive researchers to identify and detect cyberbullying to fight it. Unfortunately, most of the previous approaches were on English texts, hardly any on other languages. This paper presents a cyberbullying detection system in the Moroccan dialect on an Instagram-collected dataset. The experiment results gave accuracies of around 77% to 91% from both the ML and DL algorithms. The LSTM model gave the best outcome by 91.24% outperforming the ML models.**

Keywords—toxicity, social media, deep learning, machine learning, cyberbullying, Instagram, Moroccan dialect, natural language processing

I. Introduction

Social media went from playing traditional roles to playing hybrid ones in the digital world: it can be a place for companies' promotions despite their size, or for average people to show their personal life and express their beliefs, ideas, and opinions [1]. Social media platforms expand the reach and reduce costs by providing three areas of advantage for customers [2]. First, the marketing company can give customers endless information without human participation. A social media marketing company can foster relationships by personalizing information for each customer, enabling them to create products and offerings that precisely suit their needs [3]. Last but not least, social media platforms can enable business-to-consumer transactions that generally involve face-to-face interaction, as is the case for successful businesses [2]. Numerous daily opportunities for communicating with friends, classmates, and people who have similar interests are provided through social media platforms and apps. Among the popular social media platforms, there is Instagram with millions of users daily, mostly young adults, it is the most frequently used social media platform among the participants, and they favor using it for many purposes [4]. The number of youths and teenagers using such sites has substantially increased during the last five years. A recent survey found that more than half of teenagers log on to social media sites more than once each day and 22% of teenagers log on to their favorite social media sites more than 10 times per day [5]. Currently, 75% of teenagers have a cell phone: 25% of them use it for social media, 54% for texting, and 24% for instant chatting [6]. As a consequence, a significant portion of this generation's social and emotional growth takes place while using the Internet and mobile devices. This makes them vulnerable to people's opinions about them and it reflects on their behaviors and personality. Research has demonstrated that consistent use of social media platforms benefits kids and teens by fostering communication, social connections, and even technical abilities [7]. However, negative online word-of-mouth poses substantial obstacles. Unfortunately, social media platforms contain bullies that attack people for several reasons either for expressing who they are and for expressing their own opinions, or even for their looks and body shame which can leave them with complexes and scars. Among the brightest nowadays Instagram is one of the social media platforms that has a large number of daily users because of its special features [8]. The negative effects of social media have been the subject of numerous studies, however only a few studies have focused on Arabic texts [9], and more on the English ones in the majority of the studies [10]. Cyberbullying has harmful impacts on the victims: Isolation, Anger issues, depression and anxiety [11], low self-esteem, academic issues [12], or in the worst cases self-harm and suicidal thoughts [13]. Let alone the physical effects such as eating disorders especially for girls or sleeping disturbances [14]. Numerous research papers about cyberbullying were made in many languages but only a few of them concentrated on Arabic and especially the Moroccan dialect [15].

According to Internet World Stats. Arabic is the fourth most used Internet language after English, Chinese, and Spanish [16]. There are three main forms of Arabic: classical Arabic (CA), Modern Standard Arabic (MSA), and Arabic Dialect (AD) [17]. The oldest form of Arabic, known as classical Arabic, is used in the Coran, classical literature, and sacred books, while Standard Arabic is the simplified form of it that has undergone certain grammatical modifications. It is employed in business, administration, and the field of education for formal spoken or written communications [18]. However, Arabic dialects refer to dialects that are commonly spoken regionally in each nation. Moroccan dialect (MD) is one of the western group of Arabic dialects spoken in Morocco, it has unique features that distinguish it apart from other Arabic dialects. Officially, the majority of Moroccans use MD in informal situations about 89%, while only less than 27% of the Moroccan population speak Tamazight [19]. The informal language of Moroccan was strongly impacted by The French and Spanish cultures, people picked up certain words and terms from those languages. Currently, MD keeps adding fresh terms from several origins. However, the Moroccan dialect's vocabulary is still impacted by Arabic [20].

Recently, Artificial intelligence (AI), Machine Learning (ML), and Deep Learning (DL) techniques have contributed to the development of many applications such as heart disease and Parkinson's disease prediction [21]–[26], energy forecasting and management [27], brain tumor detection [28]–[30], diabetic retinopathy detection [31]–[36], COVID-19 early diagnosis [37]–[40], Handwritten recognition [41]–[43], Social media content processing [44]–[46], etc.

In this paper, we present a cyberbullying detection system in the context of the Moroccan dialect. For this end, an Instagram- dataset is scraped and prepared. The experiment results gave accuracies of around 77% to 91% from both the ML and DL algorithms. The LSTM model gave the best outcome by 91.24% outperforming the ML models.

The remainder of this paper is organized as follows: Section II presents the materials and methodology. Results and discussion are provided in Section III. Section IV concludes this paper with future work suggestions.

II. MATERIALS AND METHODOLOGY

A. The architecture of the proposed detection system

This experiment went from several steps: the first one started by collecting the dataset from Instagram using selenium. After that in the second step, we did the preprocessing and cleaning of the dataset. The third step consists of building the ML and DL models. Afterward the evaluation of the models, and then the deployment of the models. Figure 1 indicates the steps of the methodology. This experiment aims to develop models that can be used for toxicity detection from social media.

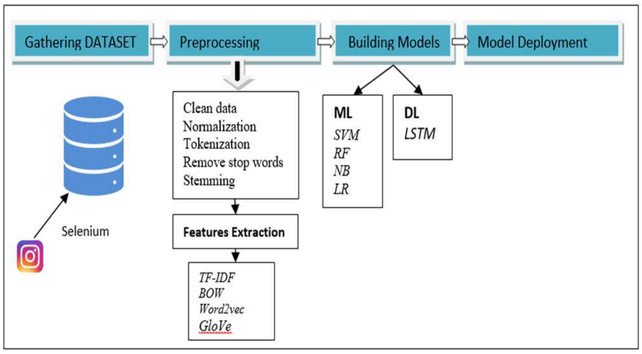


Fig. 1. The methodology steps

B. Gathering and preprocessing of the dataset

Data collection is the process of acquiring information on a particular subject, and social media is the ideal source because it offers a variety of aspects that may be investigated to enhance the volume of available data. When gathering social media data, the social media platform with the highest number of active users should be used to extract huge datasets. Since Instagram witnessed rapid growth recently with millions of users we based on it to extract our dataset. Web scraping refers to obtaining data from web pages. The procedure can be carried out by using a variety of tools, as well as libraries and frameworks that provide numerous capabilities that make web scraping easier. Selenium is an open-source web-based automation tool that performs web scraping with high efficiency. Selenium's web driver provides several features that enable us to browse the necessary websites and fetch various page contents [47]. Web scraping's role is to convert unstructured data on the web into structured data that can be stored in a database or spreadsheet [48]. Selenium is regarded as a powerful tool for web scraping even though there are many other alternatives, and that's due to its rapidity and its flexibility for scraping dynamically populated web pages [49]. We deployed Selenium in the process of compiling the dataset for this paper. The main column is the comments, then we added two more columns: the clean text and the classification of that text. The texts have two categories: positive and toxic. For the preprocessing steps, we had some difficulties since the tools are made for standard Arabic, not for dialects.

The initial preprocessing of the dataset evolves around the 3 steps. First, remove non-Arabic text while classifying our dataset we removed all comments written in Latin characters. Second, remove emojis because NLP wouldn't recognize them while working on text preprocessing. Third, remove punctuation and special characters to make it easier to analyze. After the basic cleaning, it comes to the process of normalization, deleting stop words, stemming... Afterward, we generated the words could for each category as indicated in Figure 2, where (a) is for positive words and (b) is for toxic ones.

(a)

(b)

Fig. 2. Words clouds. (a) Positive words. (b) Toxic words

C. Features extraction

Feature extraction is a significant stage in the process of initial text classification. In this paper, we have used three feature analysis techniques such as BOW, TF-IDF, word2vec, and N-Gram. Different N-Gram techniques, for instance, unigram, bigram, trigram.

1) The Term Frequency-Inverse Document Frequency (TF-IDF): One of the most observable feature extraction methods used in NLP. TF-IDF is normally used to weigh the keywords *[50]*. The Term frequency indicates the frequency of a certain term or phrase appearing in a document. While the inverse document frequency shows how often a term appears across all documents.

2) Word2Vec: it is a technique for natural language processing. Word2vec is not the first, last, or best for word embeddings, but word2vec is simple and accessible *[51]*. It calculates the cosine resemblance between the word vectors to understand the semantic similarity. Similar meaningful words have similar vectors, while dissimilar words have diversified vectors *[52]*.

3) Bag of Words: One of the most widely used feature representation methods is bag-of-words (BoW). It is a popular feature representation method for natural language processing and document representation in information retrieval *[53]*. BOW indicates the words' occurrence in a document. In the classification of documents, a BoW is a vector of the number of word occurrences, which is also called a histogram of that document *[54]*.

III. RESULTS AND DISCUSSIONS

A. Performance measures

To evaluate the system's effectiveness, several standard evaluation measures are usually introduced. Accuracy, Precision, Recall, F1 Score, Confusion Matrix, and Receiver Operating Characteristic Curve (ROC) are some of these measurements. A confusion matrix is a table that is often used to evaluate the effectiveness of a classification model. It gives details regarding four metrics: True Positives (TPs), False Positives (FPs), True Negative (TNs), and False Negative (FNs) as shown in Figure 3. These measurements are used to determine the performance indicators of several models, including accuracy precision, F1 score, and recall,.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig. 3. Confusion Matrix Example

The proposed methodology's effectiveness was assessed using the criteria of precision, specificity, accuracy, and the ROC curve. The mathematical equations of the valuation metrics are detailed in the formulas (1), (2), (3), (4), and (5).

$$Specificity = \frac{TN}{TN+FP} \tag{1}$$

$$Sensivity = \frac{TP}{FN+TP} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$F1\text{-}Score = 2 * \frac{\text{Precision* Sensivity}}{\text{Precision+Sensivity}} \tag{5}$$

B. Training results

The dataset contains 2175 comments from random posts on Instagram. For each category of the classification, there are almost 1000 instances, we used 80% for training and the other 20% for testing. We build the models in Python using the Jupyter Notebook and Google Colab platforms. To build the ML and DL models, we used version 2.9 of the TensorFlow library.

C. Testing results

We used Ml algorithms such as SVM, RF, LR, and NB alongside LSTM. The obtained results of the models are presented in Table 1. LSTM got the highest score among all the models, and among the ML models, it was SVM, especially the ones that used TF-IDF.

TABLE 1. MODELS ACCURACY WITH DIFFERENT WORD EMBEDDING METHODS.

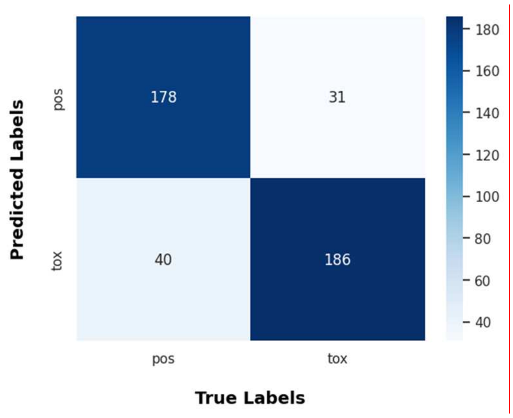
Models	TF-IDF (%)			BAG OF WORDS (%)			WORD2VEC (%)
	Uni-gram	*1g+2g*	*1g+2g +3g*	*Uni-gram*	*1g+2g*	*1g+2g+3g*	
SVM	85.06	83.91	83.22	82.99	81.61	81.61	-
RF	78.85	77.93	77.47	79.31	78.62	77.01	-
LR	84.37	83.91	83.45	83.91	82.76	82.76	-
NB	85.06	84.83	84.6	85.52	84.83	84.83	-
LSTM	-			-			91.24

For the models' evaluation, we used 4 main performance metrics: Accuracy, Precision, F1-score, and Recall. The results of this experiment show that LSTM outperformed the other algorithms with high accuracy up to 91.24 %. With an accuracy of 85.06% and an F1 of 83.29%, the SVM model is seen to perform the second best, demonstrating strength in comparison to the other Ml models. Model RF, on the other hand, had the lowest accuracy (78.85%) and F1 (81.30%). Table 2 displays the commonly used measures for the generated models.

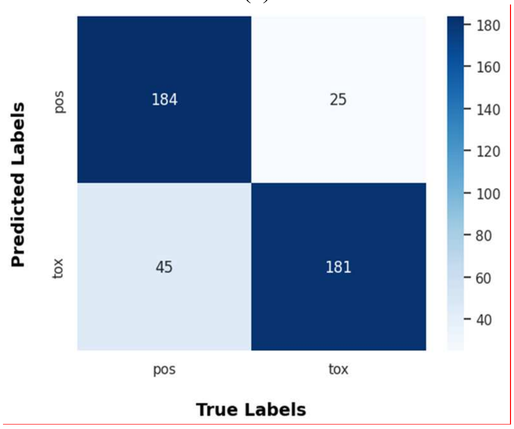
TABLE 2. PERFORMANCE METRICS OF THE MODELS.

models	Accuracy (%)	Precision (%)	F1 score (%)	Recall (%)
SVM	85.06	86.95	83.29	87.08
RF	78.85	84.61	81.30	88.49
NB	85.06	88.51	85.05	88.51
LR	84.37	87.01	83.48	87.08
LSTM	91.24	91.47	91.20	91.13

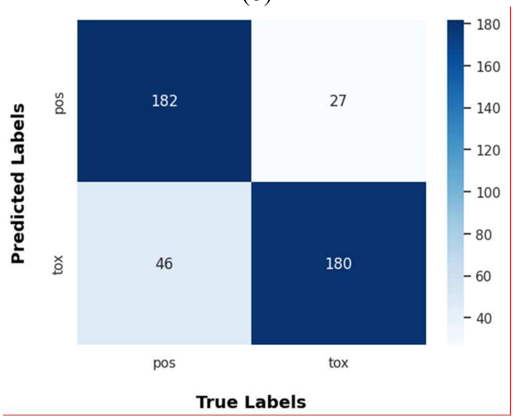
The confusion matrix reflects the performances of the models.



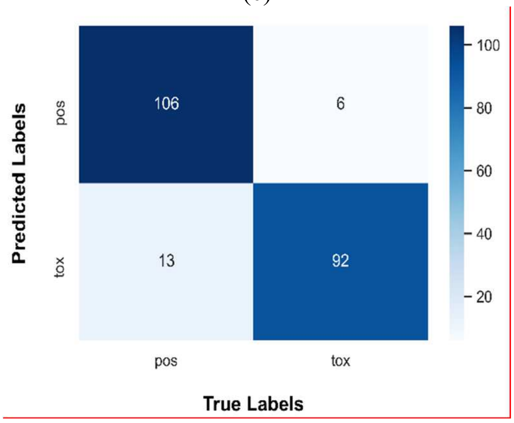
(a)



(b)



(c)

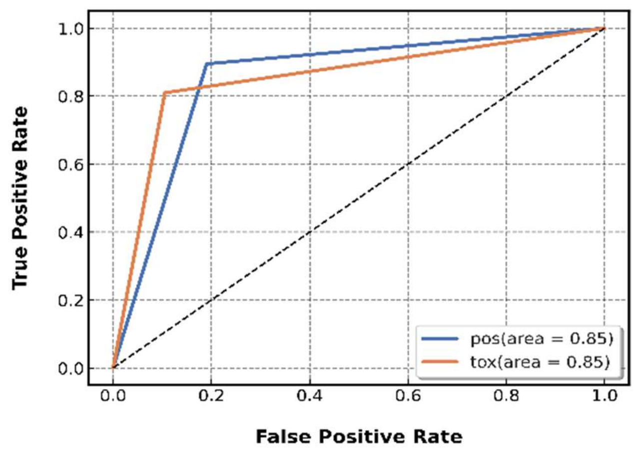


(d)

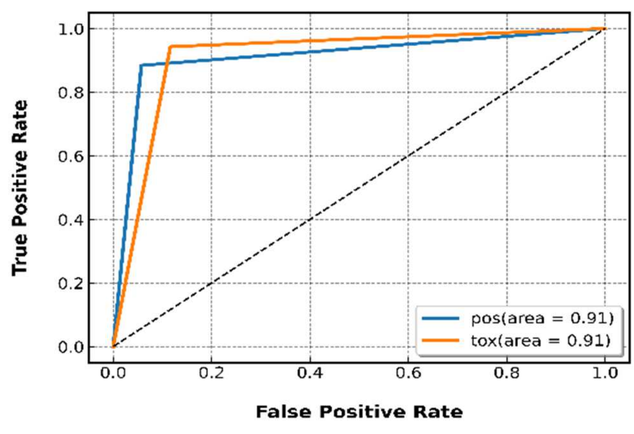
Fig. 4. Confusion matrices of SVM using TF-IDF N-gram method. (a) TF-IDF uni-gram. (b) TF-IDF 1g+2g. (c) TF-IDF 1g+3g. (d) LSTM.

The confusion matrices are displayed in Figure 4: Figure 4.a shows the SVM confusion matrix using the TF-IDF uni-gram method, while Figure 4.b presents the SVM confusion matrix using TF-IDF 1g+2g method, where Figure 4.c presents the SVM confusion matrix using TF-IDF 1g+3g method, and Figure 4.d indicates the confusion matrix of LSTM.

To evaluate the performances of the models we used the ROC curves as well. ROC is a probability curve that demonstrates the model's ability to distinguish between classes. The ROC curves are displayed in Figure 5, Figure 5.a illustrates the ROC curve for SVM, while Figure 5.b displays the ROC curve for LSTM. These graphs give a visual representation of how well the algorithms perform and show that LSTM performs better than SVM in classifying data and making predictions.



(a)



(b)

Fig. 5. ROC curves of the models. (a) SVM. (b) LSTM

D. Discussion

In this paper, we focused on Moroccan-language cyberbullying on social media. We collected the used dataset from scratch from the famous platform Instagram because the datasets that are currently available for this project are primarily in English and we were unable to find the Moroccan dialect ones. The texts in this dataset have been divided into 2 categories: positive and toxic where each category is identified as follows: 0 for toxic, 1 for positive. Overall, the obtained results from this experiment were good even though we encountered some difficulties with the dataset size and the Moroccan dialect.

IV. CONCLUSION AND PERSPECTIVES

Social media has experienced phenomenal growth, especially in the last few years. It's undeniable how much social media helped people in many ways but unfortunately, it still has a dark side filled with hate and toxicity. In this approach, we have used RNN-LSTM and some ML algorithms: SVM, NB, and RF. The experiment proved that the LSTM model performed better than the ML models with an accuracy rate of 91.24% and an F1-score rate of 91.20%.

In order to improve the outcomes, we propose improving the model's suitability for dealing with the Arabic language, particularly the Moroccan Dialect. Another suggestion is by using other Dl models as AraBERT or using hybrid models where we can use combined algorithms to obtain even better results.

REFERENCES

[1] M. Saravanakumar and T. SuganthaLakshmi, 'Social media marketing', *Life science journal*, vol. 9, no. 4, pp. 4444–4451, 2012.

[2] R. Nadaraja and R. Yazdanifard, 'Social media marketing: advantages and disadvantages', *Center of Southern New Hempshire University*, pp. 1–10, 2013.

[3] R. Trichur Narayanan, 'Recommender System: Personalizing User Experience or Scientifically Deceiving Users?', in *2021 the 5th International Conference on Information System and Data Mining*, Silicon Valley CA USA: ACM, May 2021, pp. 138–144. doi: 10.1145/3471287.3471303.

[4] A. Erarslan, 'Instagram as an Education Platform for EFL Learners', *Turkish Online Journal of Educational Technology - TOJET*, vol. 18, no. 3, pp. 54–69, Jul. 2019.

[5] G. S. O'Keeffe and K. Clarke-Pearson, 'Communications Co, Media: the impact of social media on children', *Adolescents Families Pediatrics*, vol. 127, pp. 800–804, 2011.

[6] S. Hinduja and J. W. Patchin, 'Offline consequences of online victimization: School violence and delinquency', *Journal of school violence*, vol. 6, no. 3, pp. 89–112, 2007.

[7] M. Ito *et al.*, *Living and learning with new media: Summary of findings from the digital youth project*. The MIT Press, 2009.

[8] M. Pittman and B. Reich, 'Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words', *Computers in Human Behavior*, vol. 62, pp. 155–167, Sep. 2016, doi: 10.1016/j.chb.2016.03.084.

[9] D. Ostic *et al.*, 'Effects of Social Media Use on Psychological Well-Being: A Mediated Model', *Front. Psychol.*, vol. 12, p. 678766, Jun. 2021, doi: 10.3389/fpsyg.2021.678766.

[10] H. Thygesen *et al.*, 'Use and self-perceived effects of social media before and after the COVID-19 outbreak: a cross-national study', *Health Technol.*, vol. 11, no. 6, pp. 1347–1357, Nov. 2021, doi: 10.1007/s12553-021-00595-x.

[11] W. Cassidy, C. Faucher, and M. Jackson, 'Adversity in university: Cyberbullying and its impacts on students, faculty and administrators', *International journal of environmental research and public health*, vol. 14, no. 8, p. 888, 2017.

[12] N. B. Alotaibi, 'Cyber bullying and the expected consequences on the students' academic achievement', *IEEE Access*, vol. 7, pp. 153417–153431, 2019.

[13] K. Subaramaniam, R. Kolandaisamy, A. B. Jalil, and I. Kolandaisamy, 'Cyberbullying challenges on society: a review', *Journal of positive school psychology*, vol. 6, no. 2, pp. 2174–2184, 2022.

[14] J. M. Nagata *et al.*, 'Cyberbullying and Sleep Disturbance among Early Adolescents in the US', *Academic Pediatrics*, 2022.

[15] F. Shannag, B. H. Hammo, and H. Faris, 'The design, construction and evaluation of annotated Arabic cyberbullying corpus', *Educ Inf Technol*, vol. 27, no. 8, pp. 10977–11023, Sep. 2022, doi: 10.1007/s10639-022-11056-x.

[16] K. Dashtipour *et al.*, 'Multilingual sentiment analysis: state of the art and independent comparison of techniques', *Cognitive computation*, vol. 8, pp. 757–771, 2016.

[17] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, 'Arabic natural language processing: An overview', *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 5, pp. 497–507, 2021.

[18] H. M. Al Chalabi, S. K. Ray, and K. Shaalan, 'Question classification for Arabic Question Answering Systems', in *2015 International Conference on Information and Communication Technology Research (ICTRC)*, May 2015, pp. 310–313. doi: 10.1109/ICTRC.2015.7156484.

[19] R. Tachicart, K. Bouzoubaa, and H. Jaafar, 'Lexical differences and similarities between Moroccan dialect and Arabic', in *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, IEEE, 2016, pp. 331–337.

[20] R. Tachicart, K. Bouzoubaa, and H. Jaafar, 'Building a Moroccan dialect electronic dictionary (MDED)', in *5th International Conference on Arabic Language Processing*, 2014, pp. 216–221.

[21] O. Terrada, A. Raihani, O. Bouattane, and B. Cherradi, 'Fuzzy cardiovascular diagnosis system using clinical data', in *2018 4th International Conference on Optimization and Applications (ICOA)*, Mohammedia: IEEE, Apr. 2018, pp. 1–4. doi: 10.1109/ICOA.2018.8370549.

[22] O. Terrada, B. Cherradi, S. Hamida, A. Raihani, H. Moujahid, and O. Bouattane, 'Prediction of Patients with Heart Disease using Artificial Neural Network and Adaptive Boosting techniques', in *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, Marrakech, Morocco: IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/CommNet49926.2020.9199620.

[23] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, 'Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison', *Computers in Biology and Medicine*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[24] J. Goyal, P. Khandnor, and T. C. Aseri, 'Classification, Prediction, and Monitoring of Parkinson's disease using Computer Assisted Technologies: A Comparative Analysis', *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103955, Nov. 2020, doi: 10.1016/j.engappai.2020.103955.

[25] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, 'Parkinson's Disease Identification using KNN and ANN Algorithms based on Voice Disorder', in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Meknes, Morocco: IEEE, Apr. 2020, pp. 1–6. doi: 10.1109/IRASET48871.2020.9092228.

[26] A. Ouhmida, A. Raihani, B. Cherradi, and O. Terrada, 'A Novel Approach for Parkinson's Disease Detection Based on Voice Classification and Features Selection Techniques', *Int. J. Onl. Eng.*, vol. 17, no. 10, p. 111, Oct. 2021, doi: 10.3991/ijoe.v17i10.24499.

[27] J. M. González-Sopeña, V. Pakrashi, and B. Ghosh, 'An overview of performance evaluation metrics for short-term statistical wind power forecasting', *Renewable and Sustainable Energy Reviews*, vol. 138, p. 110515, Mar. 2021, doi: 10.1016/j.rser.2020.110515.

[28] N. A. Ali, A. E. abbassi, and B. Cherradi, 'The performances of iterative type-2 fuzzy C-mean on GPU for image segmentation', *J Supercomput*, vol. 78, no. 2, pp. 1583–1601, Feb. 2022, doi: 10.1007/s11227-021-03928-9.

[29] H. Moujahid, B. Cherradi, and L. Bahatti, 'Convolutional Neural Networks for Multimodal Brain MRI Images Segmentation: A Comparative Study', in *Smart Applications and Data Analysis*, M. Hamlich, L. Bellatreche, A. Mondal, and C. Ordonez, Eds., in Communications in Computer and Information Science, vol. 1207. Cham: Springer International Publishing, 2020, pp. 329–338. doi: 10.1007/978-3-030-45183-7_25.

[30] Md. S. I. Khan *et al.*, 'Accurate brain tumor detection using deep convolutional neural network', *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4733–4745, 2022, doi: 10.1016/j.csbj.2022.08.039.

[31] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, 'Diabetic retinopathy detection through deep learning techniques: A review', *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020, doi: 10.1016/j.imu.2020.100377.

[32] O. Daanouni, B. Cherradi, and A. Tmiri, 'Predicting diabetes diseases using mixed data and supervised machine learning algorithms', in *Proceedings of the 4th International Conference on*

Smart City Applications, Casablanca Morocco: ACM, Oct. 2019, pp. 1–6. doi: 10.1145/3368756.3369072.
[33] O. Daanouni, B. Cherradi, and A. Tmiri, 'NSL-MHA-CNN: A Novel CNN Architecture for Robust Diabetic Retinopathy Prediction Against Adversarial Attacks', *IEEE Access*, vol. 10, pp. 103987–103999, 2022, doi: 10.1109/ACCESS.2022.3210179.
[34] O. Daanouni, B. Cherradi, and A. Tmiri, 'Self-Attention Mechanism for Diabetic Retinopathy Detection', in *Emerging Trends in ICT for Sustainable Development*, M. Ben Ahmed, S. Mellouli, L. Braganca, B. Anouar Abdelhakim, and K. A. Bernadetta, Eds., in Advances in Science, Technology & Innovation. Cham: Springer International Publishing, 2021, pp. 79–88. doi: 10.1007/978-3-030-53440-0_10.
[35] O. Daanouni, B. Cherradi, and A. Tmiri, 'Automatic Detection of Diabetic Retinopathy Using Custom CNN and Grad-CAM', in *Advances on Smart and Soft Computing*, F. Saeed, T. Al-Hadhrami, F. Mohammed, and E. Mohammed, Eds., in Advances in Intelligent Systems and Computing, vol. 1188. Singapore: Springer Singapore, 2021, pp. 15–26. doi: 10.1007/978-981-15-6048-4_2.
[36] O. Daanouni, B. Cherradi, and A. Tmiri, 'Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis', in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, Marrakech Morocco: ACM, Mar. 2020, pp. 1–5. doi: 10.1145/3386723.3387887.
[37] A. Adadi, M. Lahmer, and S. Nasiri, 'Artificial Intelligence and COVID-19: A Systematic umbrella review and roads ahead', *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5898–5920, Sep. 2022, doi: 10.1016/j.jksuci.2021.07.010.
[38] H. Moujahid *et al.*, 'Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation', *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 723–745, 2022, doi: 10.32604/iasc.2022.022179.
[39] S. Hamida, O. El Gannour, B. Cherradi, A. Raihani, H. Moujahid, and H. Ouajji, 'A Novel COVID-19 Diagnosis Support System Using the Stacking Approach and Transfer Learning Technique on Chest X-Ray Images', *Journal of Healthcare Engineering*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/9437538.
[40] O. El Gannour *et al.*, 'Concatenation of Pre-Trained Convolutional Neural Networks for Enhanced COVID-19 Screening Using Transfer Learning Technique', *Electronics*, vol. 11, no. 1, p. 103, Dec. 2021, doi: 10.3390/electronics11010103.
[41] S. Hamida, B. Cherradi, O. El Gannour, O. Terrada, A. Raihani, and H. Ouajji, 'New Database of French Computer Science Words Handwritten Vocabulary', in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Taiz, Yemen: IEEE, Jul. 2021, pp. 1–5. doi: 10.1109/ICOTEN52080.2021.9493438.
[42] S. Hamida, B. Cherradi, A. Raihani, and H. Ouajji, 'Performance Evaluation of Machine Learning Algorithms in Handwritten Digits Recognition', in *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, Rabat, Morocco: IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICSSD47982.2019.9003052.
[43] S. Hamida, B. Cherradi, O. Terrada, A. Raihani, H. Ouajji, and S. Laghmati, 'A Novel Feature Extraction System for Cursive Word Vocabulary Recognition using Local Features Descriptors and Gabor Filter', in *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, Marrakech, Morocco: IEEE, Sep. 2020, pp. 1–7. doi: 10.1109/CommNet49926.2020.9199642.
[44] P. K. Roy, A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, and R. Raman, 'Disaster related social media content processing for sustainable cities', *Sustainable Cities and Society*, vol. 75, p. 103363, Dec. 2021, doi: 10.1016/j.scs.2021.103363.
[45] M. Errami, M. A. Ouassil, R. Rachidi, B. Cherradi, S. Hamida, and A. Raihani, 'Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection', *IJACSA*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140347.
[46] M.-A. Ouassil, B. Cherradi, S. Hamida, M. Errami, O. E. Gannour, and A. Raihani, 'A Fake News Detection System based on Combination of Word Embedded Techniques and Hybrid Deep Learning Model', *IJACSA*, vol. 13, no. 10, 2022, doi: 10.14569/IJACSA.2022.0131061.
[47] K. U. Manjari, S. Rousha, D. Sumanth, and J. Sirisha Devi, 'Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm', in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India: IEEE, Jun. 2020, pp. 648–652. doi: 10.1109/ICOEI48184.2020.9142938.
[48] D. S. Sirisuriya, 'A comparative study on web scraping', 2015.
[49] P. Thota and E. Ramez, 'Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis', in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, Corfu Greece: ACM, Jun. 2021, pp. 306–314. doi: 10.1145/3453892.3461333.
[50] M. Allahyari *et al.*, 'A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques'. arXiv, Jul. 28, 2017. Accessed: Apr. 09, 2023. [Online]. Available: http://arxiv.org/abs/1707.02919
[51] K. W. Church, 'Word2Vec', *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, Jan. 2017, doi: 10.1017/S1351324916000334.
[52] S. Sivakumar, L. S. Videla, T. R. Kumar, J. Nagaraj, S. Itnal, and D. Haritha, 'Review on word2vec word embedding neural net', in *2020 international conference on smart electronics and communication (ICOSEC)*, IEEE, 2020, pp. 282–290.
[53] C.-F. Tsai, 'Bag-of-words representation in image annotation: A review', *International Scholarly Research Notices*, vol. 2012, 2012.
[54] W. A. Qader, M. M. Ameen, and B. I. Ahmed, 'An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges', in *2019 International Engineering Conference (IEC)*, Jun. 2019, pp. 200–204. doi: 10.1109/IEC47844.2019.8950616.