International Journal of Difference Equations (IJDE).
ISSN 0973-6069 Volume 18, Number 1 (2023), pp. 163-178

https://www.ripublication.com/ijde.htm

A Novel Approach for Fake News and Rumour Detection

LAMGHARI Nidal[*a], **AHMAM Siham**[a] , **GHAZDALI Abdelghani** [a]

[a]*LIPIM, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco.*

Abstract

During a pandemic situation, many rumors are spreading wildly on social media platforms. This article contributes a dataset related to Covid-19 rumors (COVID-19 RM), which can detect and distinguish false from real news. Our Moroccan dataset covers not only the health field but also all the areas affected by Covid-19 including politics, sports, religion etc. This data set is used to perform a benchmark for rumor detection. It is the first work of its kind in Morocco and could be merged with other international databases, thus serving to detect especially rumors relating to Covid-19 and generally rumors relating to pandemics in the world. In this study, various classifiers were used to see which of them works best with the proposed model to classify the dataset used. The proposed model achieved the best accuracy of 90.04% and the highest f1 score of 0.901 with SVM.

Keywords: Fake news, Rumor detection, Infodemic, Machine learning, Logistic regression, Decision trees, Support vector machine, Gradient-boost.

1. INTRODUCTION

In recent decades, information and communication technologies, namely phones, computers and tablets have become accessible and usable by all members of society. According to the International Telecommunication Union, 95% of the world's population have mobile internet, and 72% of them use social media [1, 2]. So, it is clear that social media has mastered and will continue to dominate our daily lives.

[*]Corresponding Author's email : n.lamghari@usms.ma

Thus, sharing any form of information has become accessible and searchable across the world. In France, the use of social networks is moving more towards access to information [3]. Still, they also facilitate the circulation of false information [4] and this is one of the most issues encountered by users, researchers and companies. Rumors are the subject of research in several disciplines, such as philosophy, social psychology, political science, management science and, more recently, computer science, especially with the expansion of social networks. Different definitions of rumor have been proposed in the literature. Someones define rumor as a controversial statement or circulating information whose veracity status has not yet been verified at the time of publication. For other researchers, a rumor can be true or false. It's a claim whose veracity is in doubt and has no clear source, even if its origins and ideological or partisan intentions are clear. In our view, the definition most used by the scientific community is the one mentioned by DiFonzo and Bordia [5] and Qazvinian [6], who define a rumor as information that emerges and spreads, and whose truth value is unverified or deliberately false. Habitually, rumors are created for different purposes. They are intended to mislead the reader or to influence their points of view, and also to advance a political program. The fields of politics and health are the most affected by the phenomenon. It is in this context that the covid-19 pandemic has been the subject of most rumors published in the past two years. Also, for example, the circulation of information on the vaccine; consisting of injecting a chip to control people; has provoked fear and consequently the rejection of citizens. Similarly, Morocco still suffers from serious misinformation regarding the credibility of disease prevention guidelines established by the state. For example, one of the popular rumours spread was that young people don't present risks of contamination. Indeed, this false information had a serious impact on public safety. To combat these claims we need to detect them. According to our research, there are many datasets dedicated to detecting rumours which are not available in Morocco. Thus, the purpose of this work is to collect a database of fake and real Moroccan news related to COVID-19. Therefore, we will perform data analysis by implementing machine learning methods. The plan of this paper is organized as follows. First, we will show previous studies of rumor detection and some popular datasets. Our new database Covid19-RM of false and real Moroccan rumors about Covid19 is covered in section 3. Then, we present in section 4 several supervised learning models applied to the dataset as well as their results. In section 5, we will discuss the results found using our proposed dataset and the CoVID-19 FNIR dataset [7]. Finally, we will conclude with a summary and some perspectives.

2. RELATED WORKS

The detection of rumors has been the subject of so much research for several years. The one relating to covid-19 news in Arabic countries is still limited. In fact, few Arabic sets exist compared to Latin in this area. ArCOV-19 was the first database in Arabic for the detection of rumors spread on social networks [8]. It contains 2.7 million tweets of which 18.66% were published by verified accounts. The tweets are collected from over 12 Arab countries. This dataset is mainly used for research such as NLP, and social computing... ArCOV19-Rumors, contains tweets in Arabic collected via different sources and annotated manually. It always verifies tweets and complaints by using SOTA models. Furthermore, several international datasets used to detect COVID-19 rumors, and scientists have introduced new ones for an exploratory study of COVID-19 disinformation on Twitter. They collected fake tweets from fact-checking websites and randomly it relating to covid-19. Other ones have developed an ISOT Fake News Dataset containing real information and fake news from different fields. This set held 44,898 items of which 23,481 were false and 21,417 were true [9]. By applying six machine learning algorithms, Thimoty and al. have shown that mixing classifiers gives the best results [10]. In another study, researchers used three datasets taken from the real world, Weibo, Twitter 15 and Twitter 16. Data is represented by a graph in which nodes refer to users and edges to retweet relations. They proposed a model based on Convolutional Graph Networks for the detection of social media rumors. They have also shown that GCN-based approaches give good results in terms of accuracy and efficiency [11]. Other authors have evaluated deep learning approaches, (CNN, LSTM, Bert) on a Constraint @AAAI 2021 dataset which is a covid-19 dataset containing 10,700 media articles in English. They have shown that manual pre-training of the model on a corpus (covid-19) and adaptation of the model to the desired task gives better results. With the pre-training they achieved an accuracy of 98.41% instead of 93.32% by the baseline [12]. There are also researchers who used a dataset of 10,700 fake and real news items from social media and fact-checking websites. They compared the dataset with machine learning models and obtained the best result using SVM with an F1-score of 93.32% [13]. Moreover, other scientists have proposed an approach aimed at detecting rumors about COVID19 in Hindi so they created a dataset of Indian language tweets called indic-covidemic tweet. It is based on the Covid-19 infodemic set which has 504 tweets in English, 218 tweets in Arabic, 100 Bengali tweets translated via the Google Translate API, and Hindi tweets scraped from Twitter. The tweets in this set were collected over the period from March to May 2020 [14]. The classification was based on the Bert model and mBERT to extend the proposed approach to Indian languages. Their model reached 79% in F1-Score for Hindi and 81% F1-Score in Bengali [14]. In

another study [15] the researchers collected 5,182 news which 40% were in English and 60% in 39 different languages. This news was published in 105 countries by 92 fact-checkers. They were manually annotated into 11 different categories and they achieved an F1 score of 76% by using a BERT rumor detection model. Such another credibility analysis of trending news posted on Twitter, researchers evaluated automatic methods to classify whether a post was credible or not. As a result, they demonstrated that there is a difference in the way the messages spread and they classified them as false or real with 70% accuracy and 80% recall [16]. Far from Covid19, Emergent is a new dataset dedicated to debunking rumors which are made up of 300 rumors related to 2595 articles represented as a title and their label. The veracity was classified as true, false or unverified. The aim of this classification was to determine the position of the article's title regarding the complaint. Using logistic regression they achieved an accuracy of 73% [17]. In order to understand how users behave with rumors, authors [18] used the streaming API to collect a set of 330 rumor from 4,842 tweets about 9 events. They have found that using machine learning techniques which provide real-time assistance is necessary to assess the veracity of rumors. In [7], Saenz et al. introduces the CoVID19-FNIR dataset. The latter features factual fake news pulled from Poynter and real news pulled from Twitter. All news is verified by expert fact-checkers. The data samples relate to the period between February 2020 and June 2020. This dataset contains 7,588 English news articles collected from India, the United States of America and European regions. To make better use of the dataset, the authors have taken detrimental steps which include removing special characters and non-vital information.

3. DATASET

According to our research, there are many datasets that are used to detect rumors in different countries and languages. However, there is no dataset for detecting rumors of Covid-19 in Morocco. Therefore, our objective was to create one. So we have collected, from different sources, some false and real news (in English or French) about Covid19 and its different consequences.

We were able to collect 500 real news which 200 were in French and 300 in English. We got 220 pieces of news From the official Hesspress page on Facebook and 180 from the official 360 pages. We also mentioned 60 news published in the official account of the latest government's head and 40 from the official Facebook pages of different ministries. We use unverified Facebook pages to collect 300 fake news and Twitter accounts to get 101 more. So we got 401 rumors divided into 200 in French and 201 in English. The number of rumors collected from social media is limited when compared to the number of rumors circulating in reality. The cause was that the Moroccan state

has established restrictive laws that prevent the publication of disinformation on media platforms. In all, 82% of own dataset Covid19-RM (rumors of morocco) is taken from Facebook as well as 18% from Twitter. Table 1 and fig.1 give an overview of the news collected.

Text	**Label**
Again, today Morocco has received two million injections of the Chinese ”Sinopharma” vaccine against ”Corona”. With this new shipment, Morocco continues to enhance the stock of the national vaccination campaign despite the global worldwide demand for this vital product.	real
The “state conspiracy against mosques” has regained its position on widespread pages on social media, after the decision not to hold Eid alAdha prayers.	real
Je peux me protéger en pulvérisant de l’alcool ou du chlore sur mon corps.(I can protect myself by spraying alcohol or chlorine on my body.)	fake
Se rincer régulièrement le nez avec une solution saline peut aider à prévenir l’infection. (Rinsing your nose regularly with saline solution can help prevent infection.)	fake

Table 1: Overview of examples of the dataset

The government said in a statement: “Abolition of compulsory masks for vaccinated people.”	fake
Morocco is the first country in the world to reach 80% of vaccinated citizens	fake
Morocco has not recorded any cases of infection since the eons of the renewed Delta dynasty	fake
After the total elimination of the pandemic, Morocco will return to normal life, starting from the beginning of next month	fake
les experts recommandent à tout le monde d'éviter les grands rassemblements, en particulier ceux avec des personnes en d	real
les experts disent que la clé est de rester au courant et de pratiquer des habitudes sûres telles que le masquage de l'éloigner	real
des scientifiques du monde entier travaillent sur un certain nombre de vaccins et de traitements contre le covid-19.	real

Figure 1: Example of data

After manually labelling the data, we dispatched them to three different files. A file "TrainMarocRM" contains 70% of the data (630 news) used for training. For validation, we used another file "ValMarocRM" made from 15% of the data. The remaining 15% are recorded in the file "TestMarocRM" and it’s reserved for the test. All files contain two columns, the first represents the information and the other the corresponding

classification label. Table 2 shows the components of each cited file. Another way to visualize data is using Word Cloud. Figure 1 represents the words that make up the false, real data and the combination of the two. The dataset Covid19-RM is available at
.

Split	Real	Fake	Total
Training	350	280	630
Validation	75	61	136
Test	75	60	135
Total	500	401	901

Table 2: Statistic of dataset

Another way to visualize data is by using WordCloud. The Figure 2 Represent the words that make up the false, real data and the combination of the two.

(a) all tokens (b) real tokens (c) fake tokens

Figure 2: WordCloud from the dataset

4. EXPERIMENTATION.

Before using classifiers, a preprocessing and feature extraction step is required to prepare the data. Thus we have removed links, hashtags, non-alphanumeric characters and stop words for both languages. Table 3 shows an example of data before and after cleaning.

Before cleaning	RNA vaccines are based on a molecule, messenger RNA, which has an unlimited lifespan and stays inside a cell for about 365 days before being #brokendown. # covid- 19 #coronavirus
After cleaning	rna vaccines based molecule messenger rna unlimited lifespan stays inside cell 365 days broken covid19 coronavirus

Table 3: Dataset cleaning

To extract the characteristics of the text we used the TF-IDF (Term Frequency-Inverse Document Frequency). It is a very popular topic in natural language processing which usually deals with human languages. In any word processing, cleaning up the text (pre-processing) is vital. In addition, the cleaned data must be converted into a digital format where each word is represented by a matrix (word vectors). This is also known as word incorporation term frequency. The concept is to calculate the relevance of a word in relation to a text. So we calculate the number of times a word appears in a document, as well as the reverse frequency of the word in a set of documents. Performing the TF*IDF report on each word in our data allows it to have a weight. The higher the weight is the rarer the term is. The formulas (1), (2)and (3) below show how to calculate the TF-IDF

$$TF = \frac{\text{(frequency of a term in the document)}}{\text{(total number of terms in the documents)}} \tag{1}$$

$$IDF = log\frac{\text{(total number of documents)}}{\text{(number of documents with term t)}} \tag{2}$$

$$TF - IDF = TF * IDF \tag{3}$$

For the classification, we used decision trees, logistic regression, gradient boost and SVM with a linear kernel. Support-vector machines (SVM) are based on the principle of providing a hyperplane allowing the separation of two or more different classes [19]. SVM also introduces the term support vector, which are vectors passing through the data points closest to the hyperplane for each of the classes to be separated. The principle is therefore to separate the two classes as far as possible by widening the margin which is the distance between the vectors of supports. Logistic regression (LR) is a technique used to know the factors associated with a phenomenon and therefore to develop a prediction model [20]. In other words, it considers a set of qualitative variables Xi and a qualitative variable Y, logistic regression, therefore, makes it possible to study the existing relationships between them using a logistic function as well as to predict the probability of an event occurring. The decision trees (DT) principle is extracted from normal trees, but a node contains an attribute and the branches represent the decisions

as well as the leaves are the results. This simplifies the interpretations, the decision tree is therefore produced from a representation of the decision-making process by taking into account several conditions [21]. Gradient boosting (GB) relies on predicting a set of results for a future model in order to minimize errors. While making a prediction whether a model will reduce prediction errors when merged with models obtained previously [22]. We implemented these algorithms using the sklearn package. And to evaluate the performance of the algorithms, we used different metrics. Most of them are based on the confusion matrix. The confusion matrix is a tabular representation of the performance of a classification model on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative (Table.4).

	Predicted True	Predicted False
Real True	True Positive (TP)	False Negative (FN)
Real False	False Positive (FP)	True Negative (TN)

Table 4: Confusion matrix parameters

Using the F1-score, precision, accuracy, recall, and time, we have measured the quality of our classifiers.

Accuracy

Accuracy is often the most widely used metric representing the percentage of correctly predicted observations, true or false. To calculate the performance accuracy of a model, the following equation can be used (4):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{4}$$

In most cases, a high precision value represents a good model, but considering that we are training a classification model in our case, an item that was predicted as true when it was actually false (false positive) can have negative consequences. Similarly, if an article was predicted to be false despite containing factual data, it can create trust issues. Therefore, we used three other metrics that take into account the misclassified observation, namely precision, recall and F1 score.

Recall

The recall represents the total number of positive classifications out of the true class. In our case, it represents the number of news items predicted as true out of the total

Model	Accuracy	Precison	Recall	F1-score	Time
Logistic Regression	78,68	83,24	78,68	79,44	0,0039
Support Vector Machine	79,41	80,36	79,41	79,61	0,0039
Decision Tree	70,59	74,06	70,59	70,78	0,0039
Gradient Boost	74,26	74,52	74,26	74,20	0,0039

Table 5: Results of different classifiers on the validation data

number of true news items.

$$Recall = TP/(TP + FN) \tag{5}$$

Precision

Unlike recall, precision represents the ratio of true positives to all events predicted as true. In our case, the precision shows the number of items marked as true among all positively predicted (true) items.

$$Precision = TP/(TP + FP) \tag{6}$$

F1-score

The F1 score represents the trade-off between precision and recall. It calculates the harmonic average between each of the two. Thus, it takes into account both false positive and false negative observations. The F1 score can be calculated using the following formula:

$$\text{F1-score} = 2(Precision * Recall)/(Precision + Recall) \tag{7}$$

These metrics were used to determine and compare the efficiency of each method used. Of all the methods, SVM gave the best performance with a precision of 0,803, an accuracy of 79,41%, a recall of 0,794 and an F1-score of 0,796 in 0.0039 seconds. Then, confusion matrices were plotted for each method as shown in figure 3 below. The "Table.5" represents the results of the classifiers on the validation data.

As shown in “Table. 5” which represents the results of the classifiers on the test data, the decision trees (DT), the gradient boost, the logistic regression and the SVM give a f1-score of 0,707; 0,74; 0,781 and 0,8 respectively.

Model	Accuracy	Precison	Recall	F1-score	Time
Logistic Regression	77,78	79,26	77,78	78,10	0,0029
Support Vector Machine	80	80,20	80	80,06	0,0039
Decision Tree	70,59	74,06	70,59	70,78	0,0040
Gradient Boost	74,07	74,56	74,07	74,00	0,0039

Table 6: Results of different classifiers on the test data

6 Results of different classifiers on the test data.

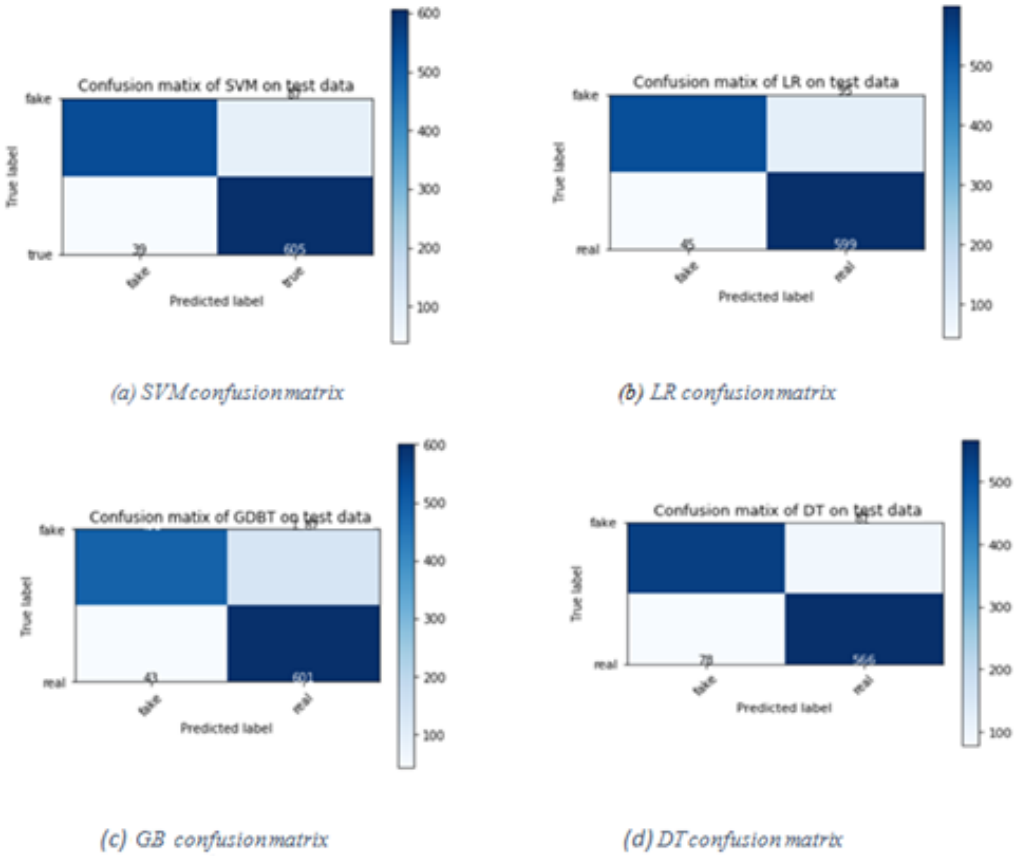

Figure 3: Confusion matrix

We can make the following analogies based on the confusion matrix of the test dataset with SVM:

- i. 542 fake news stories were correctly predicted as fake.
- ii. 605 news articles that are true (real) were correctly predicted as true.
- iii. 87 fake news stories were wrongly predicted as true.
- iv. 39 news articles were wrongly predicted as false.

Also, we can make the following generalizations based on the confounding matrix of the test dataset with logistic regression:

- i. 534 fake news items were correctly predicted as fake.
- ii. 599 news articles that are true (real) were correctly predicted as true.
- iii. 95% of fake news was wrongly predicted as true.
- iv. 45 news articles were wrongly predicted as fake.

Also, we can make the following inferences based on the confusion matrix of the test dataset with Gradient Boost:

- i. 493 fake news items were correctly predicted as fake.
- ii. 601 news articles that are true (real) were correctly predicted as true.
- iii. 136 fake news stories were wrongly predicted as true.
- iv. 43 news articles were wrongly predicted as fake.

Similarly, we can make the following inferences based on the confusion matrix of the test dataset with the Decision Tree:

- i. 536 fake news items were correctly predicted as fake.
- ii. 557 news articles that are true (real) were correctly predicted as true.
- iii. 136 fake news stories were wrongly predicted as true
- iv. 93 fake news stories were wrongly predicted as true.

5. RESULTS AND DISCUSSION

In order to judge the results obtained using our database, we compared them with those obtained with the most widely used state-of-the-art database: "CoVID-19-FNIR DATASET" [7]. It is a set of data collected for the purpose of detecting rumors regarding COVID-19 in the period between February 2020 and June 2020. It contains reputedly fake covid-19 news extracted from Poynter and real information collected from verified Twitter handles. Using the same data partitioning (70% for training, 15% for validation, and 15% for testing), we trained our data using the same models exposed in the previous section. As indicated in “Table.7”, the application of the classification models on the validation data shows that the Support Vector Machine (SVM) gives the best results in terms of accuracy, precision, recall and f1- score as well as in terms of time.

	Model	Accuracy	Precision	Recall	F1-score	Time
COVID19-FINIR	Logistic Regression	89,19	90,25	89,19	89,25	0,0625
	Support Vector Machine	90,16	90,64	90,16	90,18	0,0368
	Decision Tree	86,29	86,51	86,29	86,31	0,0629
	Gradient Boost	87,70	90,07	87,70	87,85	0,0873
COVID19-RM	Logistic Regression	78,68	83,24	78,68	79,44	0,0039
	Support Vector Machine	79,41	80,36	79,41	79,61	0,0039
	Decision Tree	70,59	74,06	70,59	70,78	0,0039
	Gradient Boost	74,26	74,52	74,26	74,20	0,0039
COVID19-FNIR + COVID19-RM	Logistic Regression	86,34	87	86,34	86,39	0,0368
	Support Vector Machine	88,07	88,47	88,07	88,1	0,0257
	Decision Tree	84,22	84,4	84,22	84,24	0,0253
	Gradient Boost	85,95	87,08	85,95	86,04	0,033

Table 7: Comparison of data validation results

By making our COVID-19 RM database undergo the same operations, we keep SVM as the best model, this is generally obtained via all the quality of results measures that we have used. Applying all these models to a general dataset that merges the CoVID19-FNIR and COVID-19 RM datasets also shows that the Support Vector Machine gives an Accuracy of 88,07%, a precision of 0,884, a recall of 0,880 and an F1-score of 0,881 which are the best results obtained when comparing them with other models. About execution terms, decision trees are better with a difference of 0.0004 with SVM. Using the test data from "Table.**??**", we obtained results similar to those found during validation. With CoVID19-FNIR, we get an accuracy of 92,53%, a precision of 90,296, a recall of 0,925 and an F1-score of 0,925, which are the best results, and they are also given by SVM. The latter remains the best in terms of the quality of the results but a longer time compared to logistic regression and decision trees. These results are generalized for the COVID-19 RM set and the extended dataset, with an improvement in all the percentages obtained via the set of model quality measures used, but with a greater time for the three different datasets as well as for the four classification models applied. The figure fig.4. illustrates a graphical comparison of the results obtained by combining the two datasets: CoVID19-FNIR and COVID-19 RM. It can easily be seen that SVM performs better for the proposed model.

From these results, it is easily remarkable that the results obtained with the CoVID19-FNIR dataset are better than those obtained with the global COVID-19 RM dataset. This is due to the fact that it contains diverse data in terms of language. That said, the results obtained with our dataset remain motivating and promising if we expand it by diversifying its content and adding other languages.

	Model	Accuracy	Precision	Recall	F1-score	Time
COVID19-FINIR	Logistic Regression	91,48	92,17	91,48	91,51	0,0291
	Support Vector Machine	92,53	92,96	92,53	92,55	0,0420
	Decision Tree	89,46	89,5	89,46	89,46	0,0260
	Gradient Boost	88,92	90,53	88,93	89,02	0,0511
COVID19-RM	Logistic Regression	77,78	79,26	77,78	78,10	0,0029
	Support Vector Machine	80	80,20	80	80,06	0,0039
	Decision Tree	70,59	74,06	70,59	70,78	0,0040
	Gradient Boost	74,07	74,56	74,07	74,00	0,0039
COVID19-FNIR + COVID19-RM	Logistic Regression	89	89,32	89	89,02	0,0505
	Support Vector Machine	90,1	90,4	90,1	90,12	0,0335
	Decision Tree	85,94	85,94	85,94	85,94	0.0286
	Gradient Boost	85,93	87,03	85,94	86,03	0,0528

Table 8: Comparison of data test results.

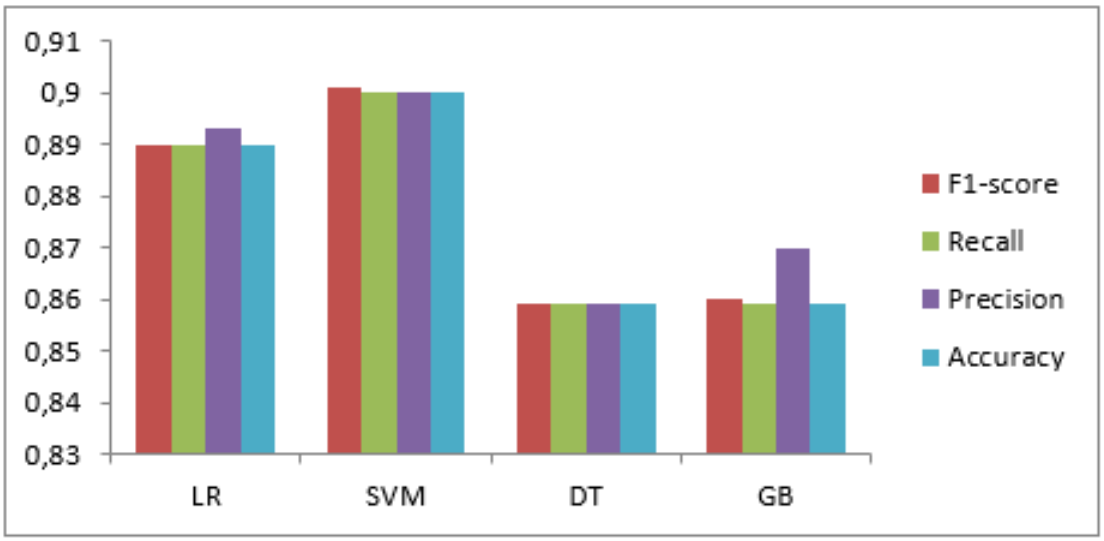

Figure 4: Comparison of all model methods

6. CONCLUSION AND FUTURE WORK.

This article presents a database containing false and real Moroccan news on Covid-19, used for rumor detection. We have collected data from various verified and unverified sources. The proposed work consists of dataset acquisition, preprocessing and classification. We experimented with different machine-learning algorithms. Results show that the SVM performed better than the others methods in terms of accuracy, precision, recall and f1-score. The model developed here uses only a few features and remains on par with other developed models that use a lot of features.

Taking real-time data into account, this model remains much less complex and reliable. From a perspective, our goal of rumor detection will not be restricted by the use of foreign languages, but it will be extended to the Arabic language and the Moroccan dialect as well as the processing of multimedia data in order to increase its efficiency and practicality.

Author Contributions : All authors contributed equally to the development of this manuscript. Conceptualization was done by N. Lamghari (NL), A.Ghazdali (AG) S.Ahamam (SA). All literature reviews and data collection were done by NL and AG. All the experiments were performed by NL and AG. Manuscript writing original draft preparation was done by NL. Review and editing were done by AG. Visualization was carried out by NL, AG and SA.

Conflicts of Interest Statement : The authors have no conflicts of interest to declare.

Funding : No funding was involved in the present work.

Data Availability Statement : The datasets presented in this study can be found in online repositories. The names of the repository/repositories can be found below: .

REFERENCES

[1] Boussahoua, M., Bentayeb, F., Boussaid, O., & Kabachi, N. (2018). A data partitioning optimization approach for distributed data warehouses on column family NoSQL systems. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA) (pp. 54-60). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[2] Boussahoua, M., Boussaid, O., & Bentayeb, F. (2017, August). Logical schema for data warehouse on column-oriented NoSQL databases. In International Conference on Database and Expert Systems Applications (pp. 247-256). Springer, Cham.

[3] Yangui, R., Nabli, A., & Gargouri, F. (2016). Automatic transformation of data warehouse schema to NoSQL data base: comparative study. Procedia Computer Science, 96, 255-264.

[4] Chevalier, M., El Malki, M., Kopliku, A., Teste, O., & Tournier, R. (2015, September). Implementation of multidimensional databases in column-oriented NoSQL systems. In East European conference on advances in databases and information systems (pp. 79-91). Springer, Cham.

[5] Dehdouh, K., Bentayeb, F., Boussaid, O., & Kabachi, N. (2015). Using the column oriented NoSQL model for implementing big data warehouses. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA) (p. 469). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[6] Boumlik, A., Soussi, N. and Bahaj, M. (2020, September). SMART-ETL-MR: Novel ETL Framework For Building Data Warehouse From Big Data Source Using Mapreduce, J. Theor. Appl. Inf. Technol., vol. 98, no. 17, pp. 3449-3460

[7] Scabora, L. C., Brito, J. J., Ciferri, R. R., & Ciferri, C. D. D. A. (2016, April). Physical data warehouse design on NoSQL databases. In Proceedings of the 18th International Conference on Enterprise Information Systems (pp. 111-118).

[8] Mior, M. J., Salem, K., Aboulnaga, A., & Liu, R. (2017). NoSE: Schema design for NoSQL applications. IEEE Transactions on Knowledge and Data Engineering, 29(10), 2275-2289.

[9] Prakash, D. (2019, May). NOSOLAP: Moving from Data Warehouse Requirements to NoSQL Databases. In ENASE (pp. 452-458).

[10] Berkani, N., Bellatreche, L., Khouri, S., & Ordonez, C. (2019). Value-driven Approach for Designing Extended Data Warehouses. In DOLAP.

[11] Khouri, S., Berkani, N., Bellatreche, L., & Lanasri, D. (2019, December). Data Cube Is Dead, Long Life to Data Cube in the Age of Web Data. In International Conference on Big Data Analytics (pp. 44-64). Springer, Cham.

[12] Berkani, N., Bellatreche, L., Khouri, S., & Ordonez, C. (2020). The contribution of linked open data to augment a traditional data warehouse. Journal of Intelligent Information Systems, 55(3).

[13] Strauch, C., Sites, U. L. S., & Kriha, W. (2011). NoSQL databases. Lecture Notes, Stuttgart Media University, 20, 24.

[14] Yangui, R., Nabli, A., & Gargouri, F. (2015, April). Towards Data Warehouse Schema Design from Social Networks. In Proceedings of the 17th International Conference on Enterprise Information Systems-Volume 1 (pp. 338-345).

[15] Dehdouh, K. (2016, September). Building OLAP cubes from columnar NoSQL data warehouses. In International Conference on Model and Data Engineering (pp. 166-179). Springer, Cham.

[16] Llave, M. R. (2018). Data lakes in business intelligence: reporting from the trenches. Procedia computer science, 138, 516-524.

[17] Nebot, V., & Berlanga, R. (2010, March). Building data warehouses with semantic data. In Proceedings of the 2010 EDBT/ICDT Workshops (pp. 1-8).

[18] Derrar, H., Boussaid, O., & Ahmed-Nacer, M. (2015). An objective function for evaluation of fragmentation schema in data warehouse. In Encyclopedia of Information Science and Technology, Third Edition (pp. 1949-1957). IGI Global.

[19] Santos, M. Y., Costa, C., Galvão, J., Andrade, C., Pastor, O., & Marcén, A. C. (2019, June). Enhancing Big Data Warehousing for Efficient, Integrated and Advanced Analytics. In International Conference on Advanced Information Systems Engineering (pp. 215-226). Springer, Cham.

[20] Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016.

[21] Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. International Journal of Information Technology, 10(2), 217-224.

[22] Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. International Journal of Computational Intelligence Research, 13(5), 899-906.

[23] Boumlik, A., Soussi, N., & Bahaj, M. (2018). Automatic Data Modeling Transformation Approach Of Nosql Document And Column Stores To Rdf. Journal Of Theoretical & Applied Information Technology, 96(15).

[24] Verma, R., & Puntambekar, D. (2018). Comparison of partitioning algorithms for categorical data in cluster. Int J Eng Sci, 18701.