Arabic cyberbullying detecting using ensemble deep learning technique

Mohammed Shihab Ahmed[1], Saif Muhannad Maher[1], Mokhalad Eesee Khudhur[2]
[1]Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Tikrit, Iraq
[2]Salah Al-Din Education Directorate, College of Computer Science and Mathematics, Tikrit, Iraq

Article Info

Article history:

Received Apr 12, 2023
Revised Jul 13, 2023
Accepted Aug 13, 2023

Keywords:

Deep learning
Feature extraction
Long short-term memory
Term frequency inverse document frequency

ABSTRACT

There has been a huge growth in recent years interest in studies on abusive language and cyberbullying detection due to its effects on both individual victims and societies. Hate speech, bullying, racism, aggressive content, harassment and other forms of abuse have all significantly increased as a result of Facebook, Instagram, and other social media platforms (SMPs). Since there is a significant need to detect, control, and prohibit the circulation of offensive content on social networking sites, we undertook this study to automate the identification of abusive language or cyberbullying. Arabic data set is balanced and will be used in the offensive detection process. Recently, ensemble machine learning has been used to increase the effectiveness of categorization models. Arabic detection is more precise given that each spatial feature text can make references to every other contextual piece of information. The authors utilized a model that merged convolutional neural network (CNN) with bidirectional long short-term memory (Bi-LSTM) and inverse document frequency gated recurrent unit (GRU) in a hybrid fashion without any post-processing. Our work outperformed every other publicly released cutting-edge ensemble model in the specifications of the official deep learning challenge. The findings indicate that the three-layer inverse document frequency long short-term memory (LSTM) classifier surpassed other classifiers in accuracy with a score of 92.75% compared to different algorithms.

Corresponding Author:

Mohammed Shihab Ahmed
Department of Computer Science, College of Computer Science and Mathematics, Tikrit University
Tikrit, Iraq
Email: Mohammad.shihab@tu.edu.iq

1. INTRODUCTION

The development of technology for information and communication (ICT) has made it easier to obtain content online and has boosted international contact between online communities. However, people can submit comments and voice their opinions without any limits thanks to the existence of anonymity and false online identities. As violent conduct and hate speech proliferate today on social media platforms (SMPs) like the internet, Twitter, and Facebook numerous safety risks and vulnerabilities are created [1], [2]. Profanity, such as shouting or swearing, has proliferated in ordinary informal discussions and on social media. Abuseful language can be violent, offensive, aggressive, or filled with hatred. Machine learning and data mining techniques offer tools for studying complex, massive, and continually changing social media data [3], [4]. In recent years, using these methods on social media has given rise to fresh insights on human interaction and behavior. There are numerous definitions of cyberbullying, some of which include cyberstalking [5]. Cyberstalking includes "stalking other people's information to produce a false charge, monitoring, identity theft, threats, and causing data destruction or alteration" [6]. However, some academics [7], [8] consider internet tracking to be a form of online bullying. With the rapid development of social networks, cyberbullying has evolved [9]. As a result, cyberbullying is now widely recognized as a severe threat to both national and international health, and the US centers for disease control (CDC) and prevention have issued public health advisories on the subject [1], [6].

Anonymity refers to circumstances in which victims are unaware of the identity of assaulters, thereby escalating emotions of helplessness and annoyance [10]. According to publicity, cyberbullying occurs on public social networking sites where films, images, or messages can be seen by anyone. According to a prior study, students acknowledge that the most serious instances of cyberbullying involve interactions that involve a sizable and visible audience [10], [11] asserts that cyberbullying poses a severe risk to social media users because it makes victims more prone to issues including low self-esteem, fear, anxiety, wrath, and even suicide. According to data compiled by [12], one in three youngsters receive threats, and 25% of internet users say they have been bullied online. Additionally, 36% of these social media users will have experienced bullying by the year 2020, according to the 145 million active users who tweet on Twitter every day about a variety of issues [13]. Although it is challenging to avoid online abuse on social media, some clever ideas could help. Models machine learning and deep learning are obvious when the need for automatic recognition of cyberbullying content is taken into consideration. The remainder of the work is structured as follows: a thorough assessment of the literature on detecting cyberbullying is included in section 2. The analysis and operation of the suggested ensemble algorithm are thoroughly explained in section 3. The results of the suggested strategy are detailed in section 4. The work in section 5 is concluded. which presents a thorough analysis of the findings and the future scope.

2. RELATED WORK

Machine learning techniques are employed to automatically detect cyberbullying because it has been extremely difficult for researchers to identify it in the Arabic environment. Additionally, it aids in the rapid resolution of problems and the creation of a safe and secure virtual world environment for government organizations. The writers distinguished between offensive speech and hate speech in [14]. Hate speech (HS) is defined as words used to show enmity toward a particular individual or group based on certain important characteristics such as sexual orientation, gender, or race and many sorts of impairments. Hate speech is intended to denigrate or discredit the target. In contrast, offensive communication is always intended to offend the listener. An method for detecting cyberbullying in English-language textual data is presented in a research study [15]. It is regarded as a pioneering study with numerous citations. They divided the assignment into smaller text-classification challenges involving delicate subjects and gathered 4,500 textual comments on divisive YouTube videos. Using broad and narrow feature sets, this work implemented multiclass classifiers, J48 binary, support vector machine (SVM), and naive bayes (NB). One of the most popular strategies for combating cyberbullying on social media is [16], in which a system is created to integrate Arabic twitter streaming application programming interface (API) for tweet collection and then to categorize Twitter users depending on whether they have used any of these phrases in their tweets. They also included a sizable collection of user comments that had been ordered deleted because they violated the rules of the Aljazeera Arabic news site. Their study describes an automated method for developing and growing a list of offensive words, as well as a large corpus of user comments that are interpreted for the purpose of identifying foul and abusive language. In order to identify name-calling harassment using the log odds ratio, they introduced hand-authoring syntactic rules and identified the role that insults and obscenities played in the process.

A supervised learning classifier is created to assist decision-makers who monitor the majority's response to traumatic stimuli and find objectionable information on Twitter. To enhance the outcomes, it integrates probabilistic, rule-based, and spatial classifiers [17]. There are insufficient language resources to do sentiment analysis on the vast collection of Arabic tweets [18]. Another study looked at how to monitor social networking sites and identify bullying by using lists of Arabic swear words [19] that established a model framework and was conducted in arab countries. They have features that are based on tweets, profiles, and social graphs, and these features produced accurate forecasts. The experiments that confirmed their strategy also assessed it using various sets of tweets and attributes to determine the bare minimum of tweets and features needed to attain the best performance from the classifier. They tested their technique using a range of learning algorithms, including J48 classifiers, SVM, and NB, and discovered that NBs produced the best results, with an 80% accuracy rate. In order to detect offensive words in internet conversations.

Alakrot *et al.* [20] and Rachidi *et al.* [21] tested many ML classifiers on a dataset of Arabic-language YouTube comments using various feature extraction and selection procedures. In feature ranking, they employed tree-based ensemble algorithms, extra trees classifier, logistic regression with L1 regularization (LR-L1), singular-value decomposition (SVD), and recursive feature elimination (RFE). They trained SVM, NB, decision tree (DT), logistic regression, and five additional standard ML classifiers (LR) using the obtained features [22] compiled a compilation of hate and insulting statements in Arabic. The sample contains 5,340 tweets with varying hatred annotations.

They explored numerous feature extraction algorithms, including ML and deep learning techniques, inside a supervised classification framework. They also trained a variety of classifiers using 2-class, 3-class, and 6-class datasets. The convolutional neural network (CNN) using mBert features (abusive vs. normal) was the most effective model for two-class categorization, with an F1-score rate of 87.03% [23], meticulously prepared a Roman Urdu micro text, including mapping slang after tokenization and the production of a (Roman Urdu) slang-lexicon. After that, the data underwent additional processing to address the metadata. non-linguistic aspects, and encoded text formats. Following the preprocessing phase, comprehensive testing with the (RNN-LSTM, RNN-BiLSTM, and CNN) models was carried out. To produce the comparison study, the efficacy and accuracy of the models were tested using a range of indicators. RNNLSTM and RNN-BiLSTM produced the best results on Roman Urdu text. Aldhyani [24] developed a BiGRU-CNN sentiment classification model for identifying cyberbullies, which consists of a complete connect layer, a classification layer, an interest technique layer, and a convolution layer.

3. METHOD

3.1. Text pre-processing

The dataset from [16] was utilized in our tests. The collection is made up of 32K removed comments from the Arabic news website Aljazeera.net. According to the community rules and guidelines website, any comment that is deemed to be sensitive, racist, sexist, advocating violence, personal attack, irrelevant, or advertising is deleted by the channel moderators. Three crowd flower employees marked the remarks as obscene, rude, and clean. The annotation produced 5,653 clean, 533 vulgar, and 25,506 foul remarks in modern standard Arabic (MSA) and other dialects. By deleting extraneous material before feature extraction, preprocessing is a vital natural language processing (NLP) step that improves the converter basis' quality before it is provided to the classifier. The Arabic language processing toolkit (ALPT) was employed in this investigation to achieve this.

3.1.1. Cleaning data

To enhance the quality of text data and guarantee that statistical analysis is done correctly, clean data and UTF-8 encoding are utilized. Special characters like @, percent (%), &, URLs, words in foreign languages, emoji, and extra spaces must be deleted from the data before moving on to the relevant analysis. UTF-8 encoding is then applied to the text in order to get the best data.

3.1.2. Tokenization

This is a crucial stage in the natural language processing process during this procedure, words are separated from one another in the text of documents by using spaces, these tokens could be letters, numbers, or symbols. Vijayarani and Janani [25] and Verma *et al*. [26] are used to divide the sentence into words. After that, it will be easier for us to process the words after they have been separated separately. The Tokenization steps as: i) Eliminate all numbers; ii) Eliminate all symbols like: !, ?, @, *, #, $, %, &, ( ), [ ], { }, <, >, =, !=, +, -, _, ,, ;", ', \ and.

3.1.3. Elimination of stop words

Stop words are words which are regularly occurring and may be identified as any word which doesn't have any remarkable importance or any word that doesn't give any importance in the classification process. The number of stop words is more than 400 words, therefore, eliminating them results in reducing dimensionality too much [27], [28]. Stop words, such as “من” “الى” “و” and others, are frequently employed in written work. These terms don't actually signify anything significant because they don't aid in differentiating between two articles, these terms don't actually mean anything significant.

3.1.4. Lemmatization

This process unifies diverse letter variations by changing all of the characters' case, either to upper- or lower-case and by getting rid of all numerals and symbols in contrast to stemming, lemmatization restricts the word to a word that already exists in the language. The normalized step is as: convert all words to lowercase.

3.1.5. Stemming

It involves simplifying a term to its simplest form. In order to remove various suffixes from words, decrease the number of words that must have matching stems, and conserve memory space and time, one technique is to reduce words to their stems [28]. An english token's stem is produced by removing all of its suffixes; for instance, "connected, connecting, connection, connections" all become "connect," as shown by [29]. Table 1 explains all of the document pre-processing procedures that were taken.

Table 1. Document pre-processing steps

1st step	Pre-processing
Input	Document set (T), Stop word list (SWL), symbol and number list (SNL)
Output	Token set for each document (S).
	Begin
Step1	Read the documents.
Step2	Separate each word from the next based on the amount of space available (tokens).
Step3	Remove all numbers and symbols tokens.
Step4	All remaining tokens should be converted to lowercase.
Step5	Remove stop words.
Step6	Porter stemming the remaining tokens executed
Step7	Construct a set of tokens for each document.
	END

3.2. Feature extraction

The feature extraction procedure is critical for selecting the best feature set method. The authors employed the (Keras tokenizer) in this experiment. Keras tokenizers are classified as follows: This class enables the processing of a text corpus by converting each text into either a vector with a coefficient for each token that can be binary depending on word count, based on term frequency inverse document frequency (TF-IDF), or a sequence of integers, each of which represents the index of a token in a dictionary. Machine learning classifiers will be applied to the dataset in this project to improve accuracy utilizing a range of text feature selection techniques. The TF and TF-IDF can be used to obtain text characteristics. The frequency of each word in the dataset is displayed by the TF. In addition to weighing each word in the dataset, the TF and TF-IDF were combined to obtain the inverse document frequency [30]. His is a key stage in document categorization that improves a text classifier's scalability, efficiency, and accuracy [31]. As a result, after pre-processing the text of documents, the properties are acquired from them.

3.2.1. Term frequency

TF is determined for each property in the document by determining how often each property occurs. This method demonstrates the significance of a feature in only one document [31], [32]. The TF is computed using (1).

$$TF(term)=\frac{occurrences\ of\ a\ particular\ term\ in\ the\ document}{total\ number\ of\ terms\ in\ the\ whole\ document} \quad (1)$$

3.2.2. IDF stands for inverse document frequency

Unlike the TF, which assesses the value of a feature in a single document, the IDF is a popular property weighting procedure. It evaluates the importance of a certain asset in a group of documents. The IDF is based on the assumption that if a property exists in just a small percentage of the documents in a collection, it will most likely be an effective discriminator of those documents [31], [32]. The IDF is calculated using (2).

$$IDF=log\ (\frac{total\ number\ of\ the\ document\ in\ the\ whole\ corpus}{number\ of\ document\ that\ contains\ a\ term}) \quad (2)$$

3.2.3. Term frequency-inverse document frequency

This computation is performed by comparing the "relative frequency" of characteristics present in a given text to the inverse proportion of that attribute in the training set [33]. his algorithm primarily determines the level of importance of a certain feature in a specific document [31], [32]. In (3) is used to calculate the TF-IDF.

$$A_{ij=TF_{ij}} \times log\,\frac{N}{DF_i} \quad (3)$$

Where A ij is the term (word) i's weight in document j, N is the total number of documents in the set, TF (ij) is the term I's frequency in document j, and DF i is the term I's frequency in the set's documents.

4. OUR WORK

In this study, the authors suggested a method for detecting cyberbullying that employs AI-based classifiers and language-independent characteristics. In light of this, our method may be used to train and classify articles written in any language. The technique is based on ^long short term memory (LSTM)^ ^CNN^, bidirectional long short-term memory (Bi-LSTM), and gated recurrent unit (GRU) models that have been trained using data from a distributed feature space [17], such as word or character embedding. We'll describe the various steps we took to create our cyberbullying detection algorithm and the rationale behind each one. At the same time, evaluate the results of the investigations and compare them to those in the academic literature. As shown in the Figure 1 explain the steps of the methodology of our proposed methods.

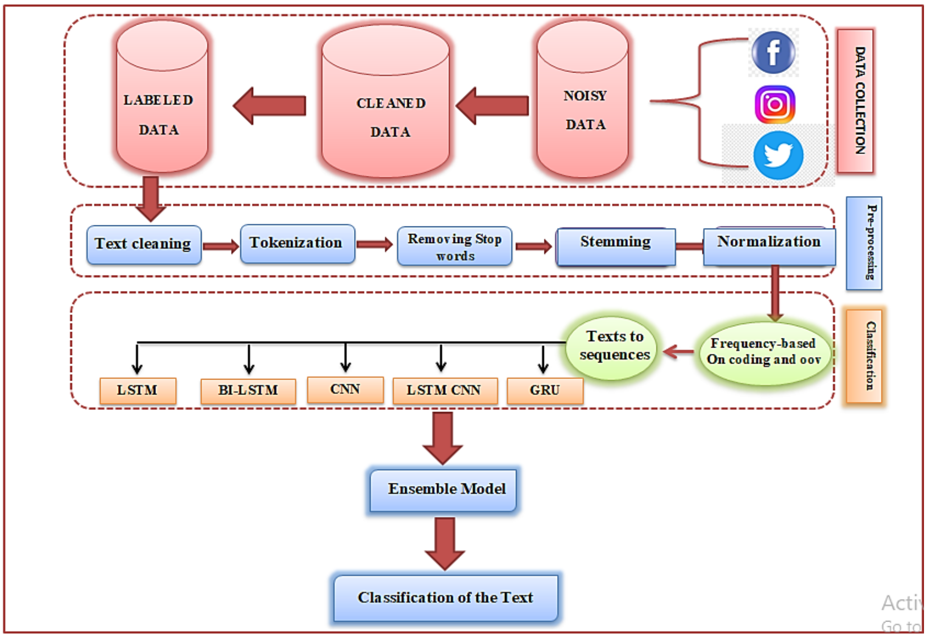


Figure 1. Framework for Arabic fake news detection

Our efforts to identify cyberbullying are based on a variety of choices, including the classification model, the use of an intelligent classifier, the selection of the most efficient algorithm, and the development of precise assessments of correctness. This complex strategy can be simply expanded to encompass new kinds of information and new languages by concentrating on universal qualities. The proposed model needs to be trained on more than just english in order to be effective in a situation where multiple languages are present. This process flowchart illustrates how we got at our end result (shown in green), which is representative of the eight stages in this approach. The selected lexical qualities are organized into four groups, while the extracted characteristics are presented in grey (shown in orange). The authors started by looking at the qualities and datasets that have previously been utilized in the literature and then picked the one that was the greatest fit for this purpose. In the steps that followed, through cleaned up the data by eliminating duplicates and outliers and by isolating linguistic characteristics. All of the obtained characteristics were normalized such that they fell within the range [0,1] before being used in a smart classifier. The intelligence algorithms used in this study (random forest, logistic regression, NBs, SVM, and DT) were chosen based on the best findings provided in the literature. The last steps included adjusting the smart models' settings to optimize performance and determining the metrics, all while comparing our results to those of the literature's previous efforts.

5. ACHIEVEMENT AND DISCUSSION

The dataset, learning strategies, and evaluation standards employed in the proposed work are described in this section. The datasets used varies in size and format. Detecting cyberbullying through the use of AI-based classifiers and language-independent characteristics. The chosen learning method predicts defects in the various types of datasets used in this work. The proposed system's findings were obtained using LSTM, CNN, Bidirectional Bi-LSTM, and GRU. We employed several processes to create system models (pre-processing, feature extraction, proposed feature selection, and modified models based on the three-layer

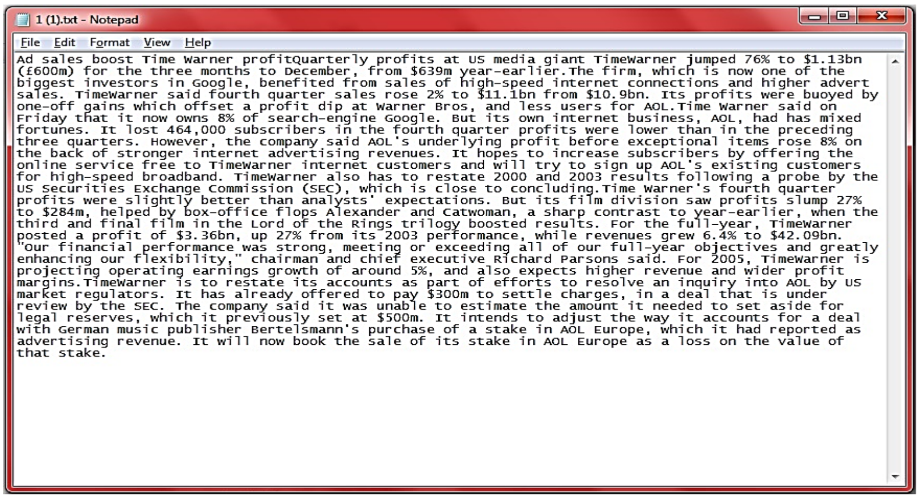
LSTM classifier). All experiments are run on a machine with 2.67 GHz CPU, 6 GB of RAM, 500 GB HDD space, and the Windows 10 64 bit operating system.

5.1. Dataset description

This step is more important in the classification operation; hence a large number of documents are employed to train and test the classifier. In our tests, the dataset was utilized. The collection is made up of 32K removed comments from the Arabic news website Aljazeera.net. According to the community rules and guidelines website, any comment that is deemed to be sensitive, racist, sexist, advocating violence, personal attack, irrelevant, or advertising is deleted by the channel moderators. Three crowd flower employees marked the remarks as obscene, rude, and clean. The annotation generated observations in MSA and other dialects in 5,653 clean, 533 vulgar, and 25,506 nasty words.

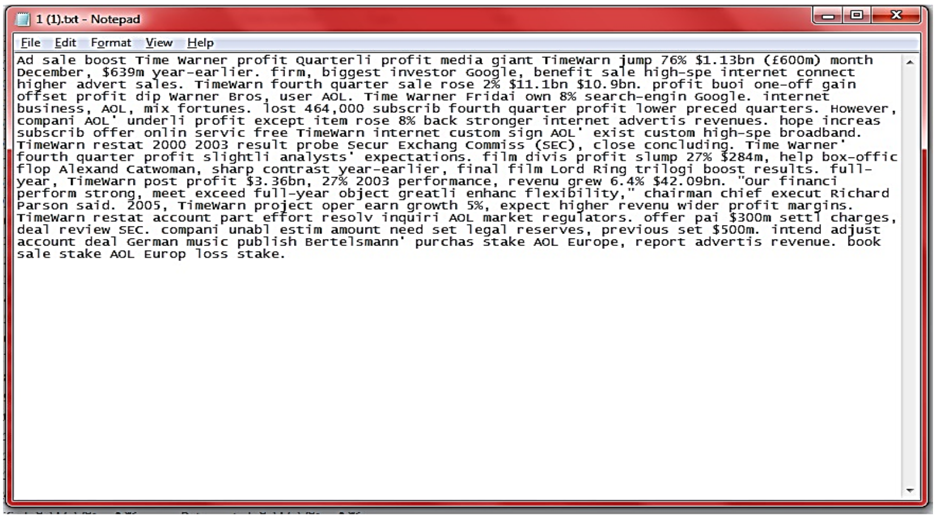
5.2. Pre-processing

The first step in our proposed work enter the dataset to pre-processing. This is a very important step in our work that will help us remove all unwanted words, punctuation marks, commas, symbols and return some words to their origin. This step is done by entering data in five stages, each stage has a role in improving and processing data. Figure 2 shows the raw dataset screen before applying pre-processing and Figure 3 shows the dataset screen after applying pre-processing on our dataset.



```
Ad sales boost Time Warner profitQuarterly profits at US media giant TimeWarner jumped 76% to $1.13bn
(£600m) for the three months to December, from $639m year-earlier.The firm, which is now one of the
biggest investors in Google, benefited from sales of high-speed internet connections and higher advert
sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by
one-off gains which offset a profit dip at Warner Bros, and less users for AOL.Time Warner said on
Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed
fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding
three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on
the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the
online service free to TimeWarner internet customers and will try to sign up AOL's existing customers
for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the
US Securities Exchange Commission (SEC), which is close to concluding.Time Warner's fourth quarter
profits were slightly better than analysts' expectations. But its film division saw profits slump 27%
to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the
third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner
posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn.
"Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly
enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is
projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit
margins.TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US
market regulators. It has already offered to pay $300m to settle charges, in a deal that is under
review by the SEC. The company said it was unable to estimate the amount it needed to set aside for
legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal
with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as
advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of
that stake.
```

Figure 2. Raw dataset screen



```
Ad sale boost Time Warner profit Quarterli profit media giant TimeWarn jump 76% $1.13bn (£600m) month
December, $639m year-earlier. firm, biggest investor Google, benefit sale high-spe internet connect
higher advert sales. TimeWarn fourth quarter sale rose 2% $11.1bn $10.9bn. profit buoi one-off gain
offset profit dip Warner Bros, user AOL. Time Warner Fridai own 8% search-engin Google. internet
business, AOL, mix fortunes. lost 464,000 subscrib fourth quarter profit lower preced quarters. However,
compani AOL' underli profit except item rose 8% back stronger internet advertis revenues. hope increas
subscrib offer onlin servic free TimeWarn internet custom sign AOL' exist custom high-spe broadband.
TimeWarn restat 2000 2003 result probe Secur Exchang Commiss (SEC), close concluding. Time Warner'
fourth quarter profit slightli analysts' expectations. film divis profit slump 27% $284m, help box-offic
flop Alexand Catwoman, sharp contrast year-earlier, final film Lord Ring trilogi boost results. full-
year, TimeWarn post profit $3.36bn, 27% 2003 performance, revenu grew 6.4% $42.09bn. "Our financi
perform strong, meet exceed full-year object greatli enhanc flexibility," chairman chief execut Richard
Parson said. 2005, TimeWarn project oper earn growth 5%, expect higher revenu wider profit margins.
TimeWarn restat account part effort resolv inquiri AOL market regulators. offer pai $300m settl charges,
deal review SEC. compani unabl estim amount need set legal reserves, previous set $500m. intend adjust
account deal German music publish Bertelsmann' purchas stake AOL Europe, report advertis revenue. book
sale stake AOL Europ loss stake.
```

Figure 3. Dataset screen after apply pre-processing

The results of applying five classifiers to the dataset are provided in Table 1. The classifiers are evaluated using the measures of precision, accuracy, F1 scores, and recall. During training, the study's parameters for accuracy, precision, recall, and F1 score are tracked. The accuracy metric is used to evaluate classification models. Informally, accuracy refers to how many right predictions a model makes. To calculate, divide the total number of guesses by the number of correct projections [34]–[38]. A result when the model correctly predicts a positive goal is known as a true positive (TP). When the model accurately predicts the negative target, it is said to be a true negative (TN). A false positive (FP) happens when the model predicts the positive objective erroneously, while a false negative (FN) happens when the model predicts the negative objective incorrectly. Precision aims the respond to the following query: How many right identifications were truly correct [35]. The goal of recall is to provide an answer to the following question: what percentage of real positives were successfully tagged [38].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1\ score\ = \frac{2*TP}{2TP+FP+FN} \tag{7}$$

On our dataset, Table 2 illustrates the performance of five classifiers with TF-IDF features. A closer look at Table 2 reveals that LSTM had the greatest results in terms of accuracy measurements, with 82.75%, outperforming other Algorithms. As shown in the Figure 4 the LSTM with three layers is the best methods between Figure 5. F1 score of models, Figure 6. The precision of models and Figure 7. Recall of the models.

Table 2. Shows the experimental findings for recall, precision, and the F1-measure for five classifiers

Algorithm	No of layers	Accuracy	Precision	Recall	F1 score
LSTM	One layer	0.81	0.8475	0.814683	0.811238
BI_LSTM	One layer	0.8045	0.808571	0.800239	0.804387
GRU	One layer	0.8025	0.826903	0.82043	0.823655
CNN	One layer	0.8165	0.819331	0.808645	0.813957
CNN BI_LSTM	One layer	0.8143	0.801002	0.822801	0.811774
LSTM	Two layer	0.827	0.847334	0.824471	0.825417
BI LSTM	Two layer	0.8155	0.807593	0.825165	0.816296
CNN	Two layer	0.813	0.802785	0.83243	0.812186
GRU	Two layer	0.823	0.81251	0.838008	0.825087
CNN BI_LSTM	Two layer	0.8049	0.857917	0.834497	0.809173
LSTM	Three layer	0.8275	0.841538	0.811521	0.826292
BI_LSTM	Three layer	0.8165	0.809412	0.825165	0.817222
GRU	Three layer	0.812	0.804179	0.837031	0.815115
CNN	Three layer	0.8165	0.811211	0.817952	0.81457
CNN BI_LSTM	Three layer	0.8145	0.807396	0.824773	0.816004

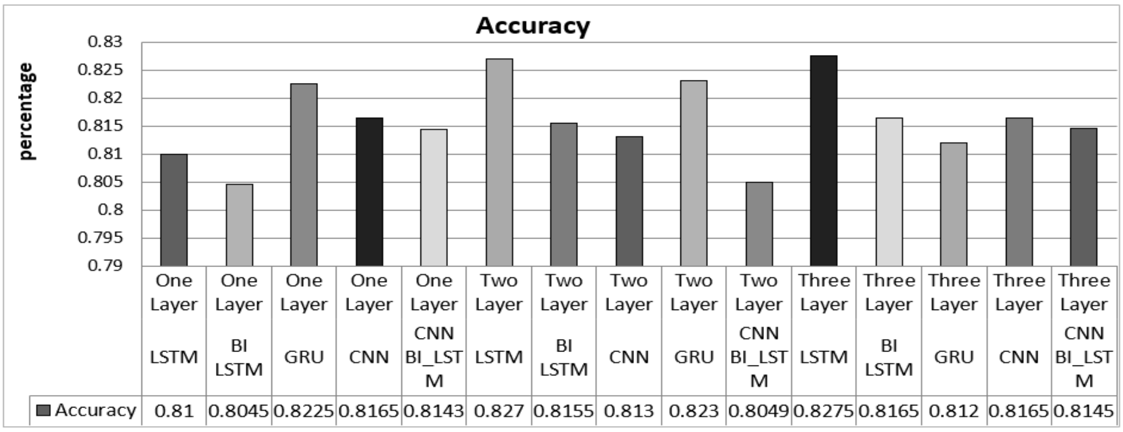
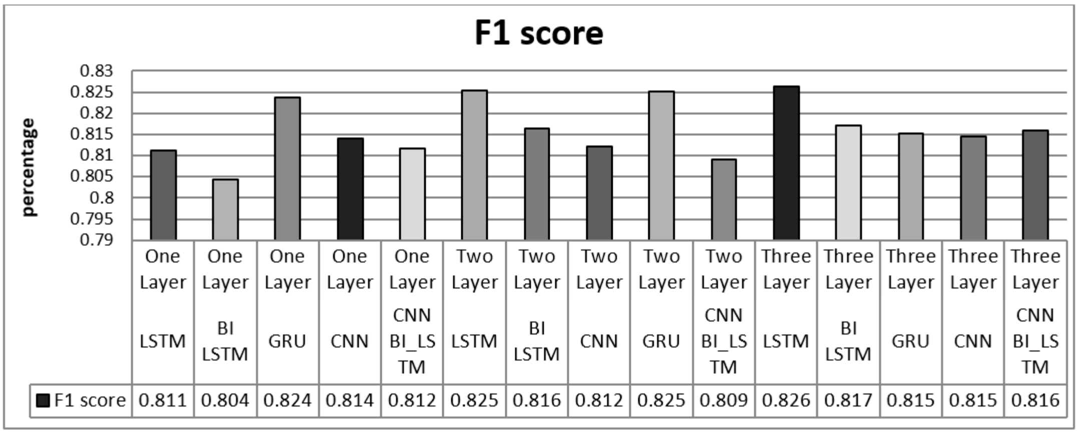
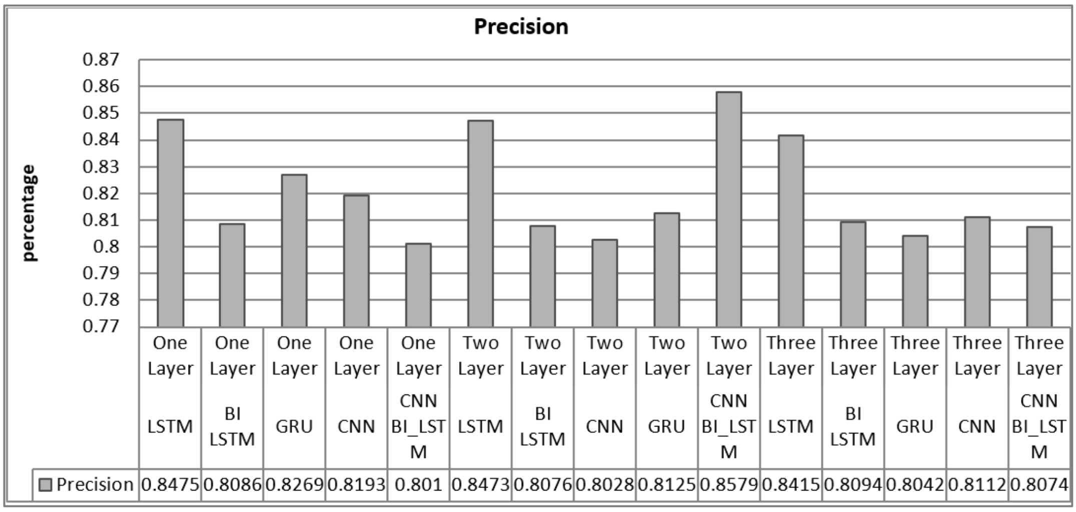


Figure 4. Models accuracy



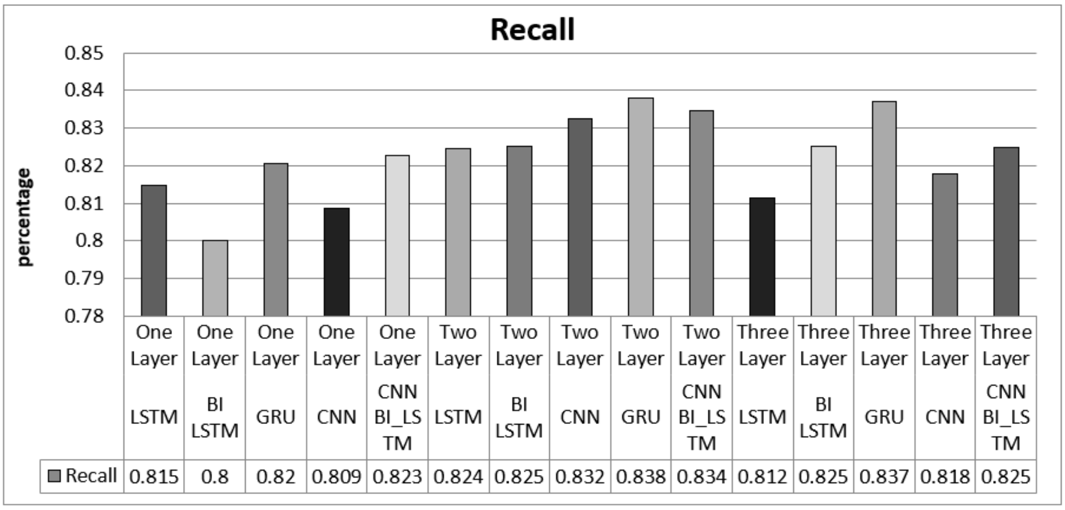
	One Layer LSTM	One Layer BI LSTM	One Layer GRU	One Layer CNN	One Layer CNN BI_LSTM	Two Layer LSTM	Two Layer BI LSTM	Two Layer CNN	Two Layer GRU	Two Layer CNN BI_LSTM	Three Layer LSTM	Three Layer BI LSTM	Three Layer GRU	Three Layer CNN	Three Layer CNN BI_LSTM
■ F1 score	0.811	0.804	0.824	0.814	0.812	0.825	0.816	0.812	0.825	0.809	0.826	0.817	0.815	0.815	0.816

Figure 5. F1 score of models



	One Layer LSTM	One Layer BI LSTM	One Layer GRU	One Layer CNN	One Layer CNN BI_LSTM	Two Layer LSTM	Two Layer BI LSTM	Two Layer CNN	Two Layer GRU	Two Layer CNN BI_LSTM	Three Layer LSTM	Three Layer BI LSTM	Three Layer GRU	Three Layer CNN	Three Layer CNN BI_LSTM
▨ Precision	0.8475	0.8086	0.8269	0.8193	0.801	0.8473	0.8076	0.8028	0.8125	0.8579	0.8415	0.8094	0.8042	0.8112	0.8074

Figure 6. Precision of models



	One Layer LSTM	One Layer BI LSTM	One Layer GRU	One Layer CNN	One Layer CNN BI_LSTM	Two Layer LSTM	Two Layer BI LSTM	Two Layer CNN	Two Layer GRU	Two Layer CNN BI_LSTM	Three Layer LSTM	Three Layer BI LSTM	Three Layer GRU	Three Layer CNN	Three Layer CNN BI_LSTM
■ Recall	0.815	0.8	0.82	0.809	0.823	0.824	0.825	0.832	0.838	0.834	0.812	0.825	0.837	0.818	0.825

Figure 7. Recall of the models

Due to comparable work in Arabic text detection By contrasting our findings with related work done in the Arabic language [17], we could see that LSTM achieved the best accuracy, at 72%, while RF came in at 68%. This little discrepancy may be explained by the different methods for machine learning. Table 3 explains from 4 classifiers on 2 tests [17] (recall, precision, and F1-measure).

Table 3. Shows the recall, precision, and F1-measure scores for four classifiers on two tests [17]

Algorithms	Accuracy	Precision	Recall	F1 score
MNB	49	8	25	38
LSVC	44	79	22	35
LR	58	56	59	58
RDG	61	72	49	58
BNB	48	63	32	43
RF	68	91	54	68
KNN	67	88	56	61
LSTM	72	91	6	72

5. CONCLUSION

The majority of currently used method offered english-language solutions for identifying cyberbullying, but none addressed cyberbullying in Arabic. The current study's objective was to assess how well LSTM and machine learning approaches could identify cyberbullying and harassment in tweets and comments and determine whether the detection of cyberbullying can be improved by using embedding TF-IDF in machine learning and deep learning techniques. In this study, we separate out the distinguishing features of cyberbullying and non-cyberbullying headlines. In this way, we may categorize cyberbullying headlines. The results show that the accuracy of the three layer LSTM classifier outperformed other classifiers with 92.75%. Find stories about cyberbullying in languages with limited resources, such There are still many issues and knowledge gaps with the Arabic language that need to be addressed, which is a novel and crucial topic for future research. Future work will concentrate on using sentiment analysis and synthetic approaches to identify cyberbullying in Arabic. In addition, by including additional deep learning models into the proposed framework and the list of cyberbullying keywords, future work will aim to offer a more precise assessment of the detection of (cyberbullying-harassment in Arabic texts). Our objective is to increase the precision of deep learning algorithms across all SMPs to protect everyone, especially children, from becoming a victim of cybercrime.

REFERENCES

[1] E. van der Walt, J. H. P. Eloff, and J. Grobler, "Cyber-security: identity deception detection on social media platforms," *Computers and Security*, vol. 78, pp. 76–89, Sep. 2018, doi: 10.1016/j.cose.2018.05.015.

[2] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B. W. On, "Aggression detection through deep neural model on Twitter," *Future Generation Computer Systems*, vol. 114, pp. 120–129, Jan. 2021, doi: 10.1016/j.future.2020.07.050.

[3] A. Rahman, M. Sadat, and S. Siddik, "Sentiment analysis on twitter data: comparative study on different approaches," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 13, no. 4, pp. 1–13, Aug. 2021, doi: 10.5815/ijisa.2021.04.01.

[4] S. M. R. K. Al- Jumur, S. W. Kareem, and R. Z. Yousif, "Predicting temperature of Erbil City applying deep learning and neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 22, no. 2, p. 944, 2021, doi: 10.11591/ijeecs.v22.i2.pp944-952.

[5] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: review of an old problem gone viral," *Journal of Adolescent Health*, vol. 57, no. 1, pp. 10–18, Jul. 2015, doi: 10.1016/j.jadohealth.2015.04.011.

[6] N. Tarmizi, S. Saee, and D. H. A. Ibrahim, "Detecting the usage of vulgar words in cyberbully activities from Twitter," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 3, pp. 1117–1122, Jun. 2020, doi: 10.18517/ijaseit.10.3.10645.

[7] H. Vandebosch and K. Van Cleemput, "Defining cyberbullying: a qualitative research into the perceptions of youngsters," *Cyberpsychology and Behavior*, vol. 11, no. 4, pp. 499–503, Aug. 2008, doi: 10.1089/cpb.2007.0042.

[8] G. S. O'Keeffe *et al.*, "Clinical report - The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, Apr. 2011, doi: 10.1542/peds.2011-0054.

[9] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2012, pp. 656–666.

[10] R. Slonje and P. K. Smith, "Cyberbullying: another main type of bullying?: personality and social sciences," *Scandinavian Journal of Psychology*, vol. 49, no. 2, pp. 147–154, Apr. 2008, doi: 10.1111/j.1467-9450.2007.00611.x.

[11] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, Aug. 2016, pp. 186–192, doi: 10.1109/ASONAM.2016.7752233.

[12] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.

[13] X. Luo, “Efficient english text classification using selected machine learning techniques,” *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/j.aej.2021.02.009.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/icwsm.v11i1.14955.

[15] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the detection of textual cyberbullying,” *AAAI Workshop - Technical Report*, vol. WS-11-02, no. 3, pp. 11–17, Aug. 2011, doi: 10.1609/icwsm.v5i3.14209.

[16] H. Mubarak, K. Darwish, and W. Magdy, “Abusive language detection on Arabic social media,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 52–56, doi: 10.18653/v1/w17-3008.

[17] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” *Policy and Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015, doi: 10.1002/poi3.85.

[18] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, “A systematic review of text classification research based on deep learning models in Arabic language,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 6629–6643, Dec. 2020, doi: 10.11591/IJECE.V10I6.PP6629-6643.

[19] E. A. Abozinadah, A. V. Mbaziira, and J. H. J. Jones, “Detection of abusive accounts with arabic tweets,” *International Journal of Knowledge Engineering-IACSIT*, vol. 1, no. 2, pp. 113–119, 2015, doi: 10.7763/ijke.2015.v1.19.

[20] A. Alakrot, M. Fraifer, and N. S. Nikolov, “Machine learning approach to detection of offensive language in online communication in Arabic,” in *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, MI-STA 2021 - Proceedings*, May 2021, pp. 244–249, doi: 10.1109/MI-STA52233.2021.9464402.

[21] R. Rachidi, M. A. Ouassil, M. Errami, B. Cherradi, S. Hamida, and H. Silkan, “Classifying toxicity in the Arabic Moroccan dialect on Instagram: a machine and deep learning approach,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 31, no. 1, pp. 588–598, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp588-598.

[22] S. Alsafari, S. Sadaoui, and M. Mouhoub, “Hate and offensive speech detection on Arabic social media,” *Online Social Networks and Media*, vol. 19, p. 100096, Sep. 2020, doi: 10.1016/j.osnem.2020.100096.

[23] A. Dewani, M. A. Memon, and S. Bhatti, “Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data,” *Journal of Big Data*, vol. 8, no. 1, p. 160, Dec. 2021, doi: 10.1186/s40537-021-00550-7.

[24] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, “Cyberbullying identification system based deep learning algorithms,” *Electronics (Switzerland)*, vol. 11, no. 20, p. 3273, Oct. 2022, doi: 10.3390/electronics11203273.

[25] S.Vijayarani and R. Janani, “Text mining: open source tokenization tools-an analysis,” *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, Jan. 2016, doi: 10.5121/acii.2016.3104.

[26] T. Verma, R. Renu, and D. Gaur, “Tokenization and filtering process in rapidMiner,” *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 16–18, Apr. 2014, doi: 10.5120/ijais14-451139.

[27] H. Saif, M. Fernandez, Y. He, and H. Alani, “On stopwords, filtering and data sparsity for sentiment analysis of Twitter Conference or Workshop Item On stopwords, filtering and data sparsity for sentiment analysis of Twitter,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014, pp. 810–817, [Online]. Available: http://lrec2014.lrec-conf.org/en/.

[28] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, “Preprocessing techniques for text mining -an overview,” *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[29] M. A. Otair, “Comparative analysis of arabic stemming algorithms,” *International Journal of Managing Information Technology*, vol. 5, no. 2, pp. 1–12, May 2013, doi: 10.5121/ijmit.2013.5201.

[30] S. Ghosal *et al.*, “A weakly supervised deep learning framework for sorghum head detection and counting,” *Plant Phenomics*, vol. 2019, Jan. 2019, doi: 10.34133/2019/1525874.

[31] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, “A novel multivariate filter method for feature selection in text classification problems,” *Engineering Applications of Artificial Intelligence*, vol. 70, pp. 25–37, Apr. 2018, doi: 10.1016/j.engappai.2017.12.014.

[32] M. Bounabi, K. El Moutaouakil, and K. Satori, “Text classification using fuzzy TF-IDF and machine learning models,” in *ACM International Conference Proceeding Series*, Oct. 2019, pp. 1–6, doi: 10.1145/3372938.3372956.

[33] Z. Erenel and H. Altnçay, “Nonlinear transformation of term frequencies for term weighting in text categorization,” *Engineering Applications of Artificial Intelligence*, vol. 25, no. 7, pp. 1505–1514, Oct. 2012, doi: 10.1016/j.engappai.2012.06.013.

[34] S. W. Kareem and M. C. Okur, “Pigeon inspired optimization of bayesian network structure learning and a comparative evaluation,” *Journal of Cognitive Science*, vol. 20, no. 4, pp. 535–552, Dec. 2019, doi: 10.17791/jcs.2019.20.4.535.

[35] S. W. Kareem and M. C. Okur, “Falcon optimization algorithm for bayesian network structure learning,” *Computer Science*, vol. 22, no. 4, pp. 553–569, Nov. 2021, doi: 10.7494/csci.2021.22.4.3773.

[36] S. W. Kareem, “A nature-inspired metaheuristic optimization algorithm based on crocodiles hunting search (CHS),” *International Journal of Swarm Intelligence Research*, vol. 13, no. 1, pp. 1–23, Jul. 2022, doi: 10.4018/IJSIR.302616.

[37] S. Ismael, S. Kareem, and F. Almukhtar, “Medical image classification using different machine learning algorithms,” *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 14, no. 1, pp. 133–145, Jun. 2020, doi: 10.33899/csmj.2020.164682.

[38] R. S. Hawezi, F. S. Khoshaba, and S. W. Kareem, “A comparison of automated classification techniques for image processing in video internet of things,” *Computers and Electrical Engineering*, vol. 101, p. 108074, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108074.

BIOGRAPHIES OF AUTHORS

Mohammed Shihab Ahmed is an assistant teacher at Tikrit University's College of Arts in Iraq. He graduated with a B.Sc. in Computer Science from Tikrit University in 2007 and an MSc in 2018 from Turkey's Altinbas University. He can be contacted at email: Mohammad.shihab@tu.edu.iq.

Saif Muhannad Maher is an assistant teacher at Tikrit University's Cisco Networking Academy in Iraq. In 2008, he graduated with a B.Sc. in Computer Science from Tikrit University, and in 2018, he earned an M.Sc. from Al-Esra University in Jordan. He can be contacted at email: saif muhannad1985@tu.edu.iq.

Mokhalad Eesee Khudhur is an assistant teacher in Tikrit, Iraq's Saladin Education Directorate. At Tikrit University, he earned a B.Sc. in Computer Science. 2018: MSc degree from Turkey's Altinbas University. He can be contacted at email: mokhalad2018@gmail.com.