Proceeding Paper

AI-Based Misogyny Detection from Arabic Levantine Twitter Tweets [†]

Abdullah Y. Muaad [1,2,*], Hanumanthappa Jayappa Davanagere [1,*], Mugahed A. Al-antari [3,*], J. V. Bibal Benifa [4] and Channabasava Chola [4,*]

[1] Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore 570006, India

[2] IT Department, Sana'a Community College, Sana'a 5695, Yemen

[3] Department of Artificial Intelligence, Sejong University, Seoul 05006, Korea

[4] Department of Computer Science and Engineering, Indian Institute of Information Technology Kottayam, Pala 686635, India; benifa.john@gmail.com

[*] Correspondence: abdullahmuaad9@gmail.com (A.Y.M.); hanumsbe@gmail.com (H.J.D.); en.mualshz@sejong.ac.kr (M.A.A.-a.); channabasavaphd2019@iiitkottayam.ac.in (C.C.)

[†] Presented at the 1st International Electronic Conference on Algorithms, 27 September–10 October 2021; Available online: https://ioca2021.sciforum.net/.

Abstract: Twitter is one of the social media platforms that is extensively used to share public opinions. Arabic text detection system (ATDS) is a challenging computational task in the field of Natural Language Processing (NLP) using Artificial Intelligence (AI)-based techniques. The detection of misogyny in Arabic text has received a lot of attention in recent years due to the racial and verbal violence against women on social media platforms. In this paper, an Arabic text recognition approach is presented for detecting misogyny from Arabic tweets. The proposed approach is evaluated using the Arabic Levantine Twitter Dataset for Misogynistic, and it gained recognition accuracies of 90.0% and 89.0% for binary and multi-class tasks, respectively. The proposed approach seems to be useful in providing practical smart solutions for detecting Arabic misogyny on social media.

Keywords: Arabic language processing; Arabic Text Representation; misogyny detection

1. Introduction

People express their thoughts, emotions, and feelings by means of posts on social media platforms. Recently, online misogyny, considered as a harassment, has increased against Arab women on a daily basis [1,2]. An automatic misogyny-detecting system is necessary for minimizing the prohibition of anti-women Arabic harmful content [2]. People are increasingly using social media platforms such as Twitter, Facebook, Google, and YouTube to communicate their various ideas and beliefs [3]. Misogyny on the internet has become a major problem that has expanded across a variety of social media platforms. Women in the Arab countries, like their peers around the world, are subjected to many forms of online misogyny. This is, unfortunately, not compatible with the values of the Islamic religion or with any other values or beliefs regarding women. Detecting such content is crucial for understanding and predicting conflicts, understanding polarization among communities, and providing means and tools to filter or block inappropriate content [3]. The main challenges and opportunities in this field are the lack of tools, with an absence of resources in the non-English (such as Arabic) dataset [4]. This research aims to develop a deep learning-based accurate approach to limit the misogyny problems. The lack of such studies from an Arabic perspective was an inspiration to investigate and find out practical smart solutions by designing and developing an automatic identification misogyny system [5].

The main contributions of this work are summarized as follows:

- The Arabic text is represented using the word and word-embedding techniques.

- The state-of-art deep learning BERT technique is used to detect Arabic misogyny.

  A comprehensive comparison study was conducted using different machine learning and deep learning techniques to achieve prominent and superior detection results.

2. Related Works

In 2020, Aggression and Misogyny Detection using BERT was proposed for three languages, such as English, Hindi, and Bengali [6]. The proposed model used an attention mechanism over BERT to get the relative importance of words, followed by fully connected layers and a final classification layer, which predicted the corresponding class [6]. The misogyny identification techniques offered satisfactory results, but the recognition of aggressiveness is still in its infancy for some languages [7]. Misogyny detection in the Arabic language is still in its early stages, with only a few important contributions existing [8]. In the last five years, there has been a growth in the number of researchers who are interested in automatic Arabic hate speech detection in social media. In the presented research, Arabic text detection based on Misogyny has been extensively studied. This study starts with a comparative study of the neural network and transformer-based language models that have been applied for Arabic fake news detection [9]. In terms of generalization, AraBERT v02 outperformed all other models evaluated. They advised using a gold-standard dataset annotated by humans in the future, rather than a machine-generated dataset, which may be less reliable [9]. In the same domain of detection, the word2vec model was suggested to detect semantic similarity between words in Arabic, which could assist in the detection of plagiarism. The authors built the word2vec model using the OSAC corpus [10]. Here, the authors focused on creating a successful offensive tweet identification dataset. They quickly constructed a training set from a seed list of offensive words. Given an autonomously generated dataset, they represented a character n-gram and used a deep learning classifier to achieve a 90% F1 score [11]. A single learner machine learning approach and ensemble machine learning approach was investigated for offensive language detection in the Arabic language [12]. In addition to this, a transfer learning method and AraBERT were used for Arabic offensive detection datasets. The results reported an outperformance of Arabic monolingual BERT models over BERT multilingual models. Their results mentioned that there was a limitation by the effects of transfer learning on the performance of the classifiers, particularly for the highly dialectic [13]. With the augmentation of the data to improve text detection, the authors experimented with seven BERT-Based models, and they augmented a task dataset to identify the sentiment of a tweet or detect if a tweet was sarcasm [14]. Their experiments were based on fine-tuning seven BERT-based models with data augmentation to solve the imbalanced data problem. For both tasks, the MARBERT BERT-based model with data augmentation outperformed other models with an increase of the F-score by 15%. Regarding the influence of preprocessing in text detection, a simple but intuitive detection system based on the investigation of a number of preprocessing steps and their combinations was addressed [15]. Here, a comparison between LSVC and BiLSTM classifiers was conducted. The detection of misogyny in Arabic text was presented using the Arabic Levantine Twitter dataset for Misogynistic language (LeT-Mi), which was the first benchmark dataset for Arabic misogyny. They employed an MTL configuration to investigate its effect on the tasks. They presented an experimental evaluation of several machine learning systems, including SOTA systems. The result for accuracy was equal to 88 and presented an approach based on stylistic and specific topic information for the detection of misogyny, exploring several aspects of misogynistic Spanish and English user-generated texts on the Twitter Section (Heading 1) [16]. Finally, an approach based on character level for Arabic text utilizing convolutional neural network (CNN) has been presented to solve many problems, such as difficulties in preprocessing, etc. [17].

3. Proposed Model

3.1. ATDS Architecture

The proposed model for detection of Arabic text from the Arabic Levantine Twitter dataset based on different types of representation and different machine learning and deep learning model has been presented in Figure 1:
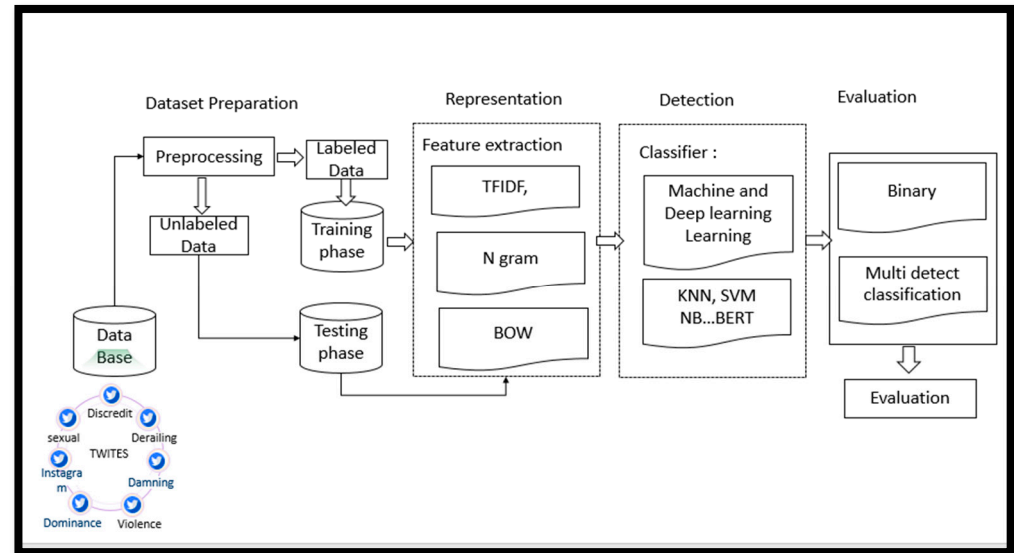
Figure 1. Architecture of the Arabic text detection system (ATDS): Abstract view.

3.2. Pre-Processing

The pre-processing technique is most commonly used for preparing raw data into a specific input data format, which could be useful for machine learning and deep learning techniques. The main purpose of preprocessing is to clean the dataset regarding stop-words, punctuation, poor spelling, slang, and other undesired words abound in text data. This unwanted noise and language may have a negative impact on the recognition performance of the Arabic misogyny detection task. In this work, we eliminated all the non-Arabic words, stop words, and punctuation through the following steps:

(a)   Tokenization

This process was used to convert the Arabic text (sentence) into tokens or words. Tokenized documents can be transformed into sentences, and sentences can be converted into tokens. Tokenization divides a text sequence into words, symbols, phrases, or tokens [18].

(b)   Normalization

The normalization is performed to make all words in the same form, and there are many techniques, such as stemming. We can make normalization by different methods such as regular expressions.

(c)   Stop Word Elimination

In the text preprocessing task, there are numerous terms that have no critical meaning but appear frequently in a document. It refers to words that do not help to increase performance, because they do not provide much information for the sentiment classification task; therefore, stop words should be removed before the feature selection process.

(d)   Stemming

One word can appear in many distinct forms, but the semantic meaning remains the same. Stemming is the process of replacing and removing suffixes and affixes to obtain the root, base or stem word.

(e)   Lemmatization

The goal of lemmatization is the same as stemming: to reduce words to their base or root words. However, in lemmatization, the inflection of words is not simply cut off; rather, it leverages lexical information to turn words into their base forms [19].

3.3. Representation

After Arabic text preprocessing, the data were transformed to be in a specific structure style for representation purposes. To perform this, bag-of-words (BOW) and term frequency-inverse document frequency (TFIDF) were used for data representation with traditional machine learning techniques. For deep learning techniques, we used a new technique called word embedding, in bidirectional encoder representations from Transformers (BERT). Instead of the basic language task, BERT was trained with two tasks to encourage bidirectional prediction and sentence-level understanding [20,21].

3.4. Text Detection

Detection of text and classification to true labeled classes based on their content is known as classification. Several works have been reported here based on text classification using different algorithms as we will explain in part 5. There are many algorithms that have been implemented as follows:

- Passive Aggressive Classifier
  Passive-Aggressive algorithms are a family of Machine learning algorithms that are popularly used in big data applications. Passive-Aggressive algorithms are generally used for large-scale learning. It is one of the online-learning algorithms. In online machine learning algorithms, the input data comes in sequential order, and the machine learning model is updated sequentially, as opposed to conventional batch learning, where the entire training dataset is used at once [20].
- Logistic Regression
  Logistic regression is a statistical model that, in its basic form, uses a logistic function to model a binary dependent variable, although many more complex extensions exist [19].
- Random Forest Classifier
  The term "Random Forest Classifier" refers to the classification algorithm made up of several decision trees. The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest's predictive powers to make accurate decisions [19].
- Linear SVC
  The support vector machine (SVM) classifier is one of the commonly used algorithms for text classification due to its good performance. SVM is a non-probabilistic binary linear classification algorithm, which is performed by plotting the training data in a multi-dimensional space. Then, SVM categorizes the classes with a hyper-plane. The algorithm will add a new dimension if the classes cannot be separated linearly in multi-dimensional space to separate the classes. This process will continue until the training data can be categorized into two different classes [19].
- Decision Tree Classifier
  Decision Trees are also used in tandem when you are building a Random Forest classifier, which is a culmination of multiple Decision Trees working together to classify a record based on a majority vote. A Decision Tree is constructed by asking a series of questions with respect to a record of the dataset we have got [19].
- K Neighbors Classifier
  KNN works by finding the distances between a query and all the examples in the data, selecting the specified number of examples (K) closest to the query, and then voting for the most frequent label (in the case of classification) or averaging the labels (in the case of regression) [19].
- ARABERTv2

AraBERT is an Arabic pre-trained language model based on Google's BERT architecture. AraBERT uses the same BERT-Base configuration [21].

4. Experimental Analysis

4.1. Dataset

The dataset [1] was unbalanced by limiting the number of articles in each specific category, as summarized in Tables 1 and 2.

Table 1. Data Distribution for each class in binary classification.

Type of Misogyny	نوع كراهية النساء	No of Tweets
None	لاشي	3061
misogyny	كراهية النساء	4805

Table 2. Data Distribution for each class in multi-classification.

Type of Misogyny (Clasess Name)	نوع كراهية النساء	No of Tweets
Discredit	تشويه السمعة	2327
Stereo typing and objectification	الكتابة والصياغة المجسمة	290
Damning	اللعنة	256
Threat of violence	التهديد بالعنف	175
Derailing	الخروج عن السكة	59
Dominance	هيمنة	38
Sexual harassment	التحرش الجنسي	17
None	لا كراهية	3388

The author classified his data as mentioned below:

1. Damning (Damn): tweets under this class contained cursing content.
2. Derailing (Der): tweets under this class combined justification of women's abuse or mistreatment.
3. Discredit (Disc): tweets under this class beared slurs and offensive language against women.
4. Dominance (Dom): tweets under this class implied the superiority of men over women.
5. Sexual Harassment (Harass): tweets under this class described sexual advances and sexual nature abuse.
6. Stereotyping and Objectification (Obj): tweets under this class promoted a fixed image of women or described women's physical appeal.
7. Threat of Violence (Vio): tweets under this class had intimidating content with threats of physical violence.
8. None: if no misogynistic behaviors existed.

4.2. Implementation Environment

To perform all experiments in this study, we used a PC with the following specifications: Intel R © Core(TM) i7-6850 K processor with 4 GB RAM and 3.360 GHz frequency. The algorithms such as Passive-Aggressive Classifier, Logistic Regression, Logistic Regression, Random Forest Classifier, K Neighbors Classifier, and linear SVC were implemented herein using Python 3.8.0 programming with Anaconda [Jupyter notebook]. The Python-based ML libraries, such as NLTK, pandas, and sci-kit-learn, were utilized to investigate the performance metrics by the proposed methods; at the same time, TensorFlow and Keras in collab were used to implement ARABERTv2. The results and discussions concerning various techniques incorporated are highlighted in the subsequent sections. The code will

be available in our account in GitHub (https://github.com/abdullahmuaad8, accessed on 8 February 2022).

4.3. Evaluation Metrics

To assess our proposed system, we used the following indices:

The recall was calculated by dividing the number of true positive (TP) observations by the total number of observations (TP + FN).

Specificity was defined as the proportion of true positive (TP) observations to the total positive forecasted values (TP + FP).

F1-score is the weighted average of recall and precision, which means that the F1-score included both FPs and FNs.

Accuracy was defined as the simple ratio of accurately predicted observations to total observations.

The definition formula of all these metrics were defined in [22] as follows:

$$\text{Recall/Sensitivity (Re)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{2}$$

$$\text{F1} - \text{score (F} - \text{M)} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{3}$$

$$\text{Overall accuracy (Az)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \tag{4}$$

where TP, TN, FP, and FN were defined to represent the number of true positive, true negative, false positive, and false negative detections, respectively. To derive all of these parameters, a multidimensional confusion matrix was used.

5. Results and Discussion

The results and discussions concerning various techniques incorporated are highlighted in this section; we describe our experiments on this data. We evaluate the performance of all algorithms on this data. We designed our experiments at two levels (tasks):

1. Misogyny identification (Binary): Tweets' contents were classified into misogynistic and non-misogynistic. This required merging the seven categories of misogyny into the misogyny class.
2. Categories classification (Multi-class): Tweets were classified into eight categories: discredit, dominance, damning, derailing, sexual harassment, stereotyping and objectification the threat of Violence, or non-misogynistic. In addition, we found that Linear SVC outperformed all compared models in terms of generalization for machine learning and BERTv2 for the deep learning technique.

5.1. Binary Classification

The results of the misogyny identification task are shown in Table 3. In terms of accuracy, precision, recall, and F-measure, the Linear SVC model outperformed the others. We also could observe that the model outperformed all the other models, except Random Forest Classifier, which works better in terms of recall. At the same time, we have been used one of the transfer learning tools called ARABERTv2, which provided excellent accuracy, but the time was more when we compared to machine learning.

Table 3. Arabic misogyny detection Evaluation results for binary classification tasks.

Method	Az	Sp	Re	F-M
Passive Aggressive Classifier	81	84	86	85
Logistic Regression	81.50	81	90	86
Random Forest Classifier	62	62	68	76
Linear SVC	83	85	88	86
Decision Tree Classifier	70	74	78	76
K Neighbors Classifier	65	64	98	78
ARABERTv2	90	-	-	-

5.2. Multi Classification

The results of the misogyny identification task are shown in Table 4. In terms of accuracy, the Linear SVC model outperformed the others. According to the results, the typical machine learning Random Forest Classifier model performance was poor. At the same time, we have used one of the transfer learning tools called ARABERTv2, which provided excellent accuracy, but the time was more than machine learning methods.

Table 4. Arabic misogyny detection Evaluation results for multiclass classification tasks.

Method	Az	Sp	Re	F-M
Passive Aggressive Classifier	72	72	72	72
Logistic Regression	69	68	68	68
Random Forest Classifier	40	39	39	39
Linear SVC	74	73	73	73
Decision Tree Classifier	56	56	56	56
K Neighbors Classifier	44	43	43	43
ARABERTv2	89	-	-	-

Finally, we would like to note that the data set was unbalanced. For example, as shown in Table 1, the class Sexual harassment only had 17 comments, which means that learning the pattern for these classes was very limited. As a result, we recommend that the number of comments in this data must increased as a future project.

6. Conclusions

The problem of misogyny has become a major problem for Arab women. In this work, we introduced a model for the detection of misogyny in Arabic text. We performed our work utilizing a dataset called Arabic Levantine Twitter Dataset for Misogynistic. Our results provided excellent accuracy, equal to 83%, using machine learning for detection and classification tasks. This article proves that many open issues need to be handel, starting with the limitation of benchmark dataset, lexicons of Arabic text in general, and especially for the misogyny of women. At the same time, the difficulty of the nature of the Arabic language in morphology and delicacy. Then, the augmentation of data using techniques such as oversampling to solve an unbalance of classes could get better performance. Finally, there is a need to study the correlation between hate speech, misogyny, and the problem of mixed language in future works.

References

1. Farha, I.A.; Magdy, W. Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 11–16 May 2020; pp. 86–90. Available online: https://www.aclweb.org/anthology/2020.osact-1.14 (accessed on 22 February 2021).
2. Mulki, H.; Ghanem, B. Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. *arXiv* **2021**, 154–163. Available online: http://arxiv.org/abs/2103.10195 (accessed on 22 February 2021).
3. Alkhair, M.; Meftouh, K.; Othman, N.; Smali, K. An Arabic Corpus of Fake News: Collection, Analysis and Classification to cite this version: HAL Id: Hal-02314246 An Arabic Corpus of Fake News: Collection, Analysis and Classification. *Arabic Lang. Process.* **2019**. [CrossRef]
4. Jahan, M.S.; Oussalah, M. A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing. 2021. Available online: http://arxiv.org/abs/2106.00742 (accessed on 22 February 2021).
5. Alshalan, R.; Al-Khalifa, H. A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Appl. Sci.* **2020**, *10*, 8614. [CrossRef]
6. Samghabadi, N.S.; Patwa, P.; Pykl, S.; Mukherjee, P.; Das, A.; Solorio, T. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 11–16 May 2020; pp. 126–131. Available online: https://www.aclweb.org/anthology/2020.trac-1.20 (accessed on 22 February 2021).
7. Fersini, E.; Nozza, D.; Rosso, P. AMI @ EVALITA2020: Automatic misogyny identification. *CEUR Workshop Proc.* **2020**, *2765*. [CrossRef]
8. Hengle, A.; Kshirsagar, A.; Desai, S.; Marathe, M. Combining Context-Free and Contextualized Representations for Arabic Sarcasm Detection and Sentiment Identification. 2021. Available online: http://arxiv.org/abs/2103.05683 (accessed on 22 February 2021).
9. Al-Yahya, M.; Al-Khalifa, H.; Al-Baity, H.; Alsaeed, D.; Essam, A. Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity* **2021**, *2021*, 5516945. [CrossRef]
10. Suleiman, D.; Awajan, A.; Al-Madi, N. Deep learning based technique for plagiarism detection in Arabic texts. In Proceedings of the 2017 International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan, 11–13 October 2017; pp. 216–222. [CrossRef]
11. Husain, F. Arabic Offensive Language Detection Using Machine Learning and Ensemble Machine Learning Approaches. 2020. Available online: http://arxiv.org/abs/2005.08946 (accessed on 22 February 2021).
12. Husain, F.; Uzuner, O. Transfer Learning Approach for Arabic Offensive Language Detection System—BERT-Based Model. 2021. Available online: http://arxiv.org/abs/2102.05708 (accessed on 22 February 2021).
13. Abuzayed, A.; Al-Khalifa, H. Sarcasm and Sentiment Detection In {A}rabic Tweets Using {BERT}-based Models and Data Augmentation. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19–20 April 2021; pp. 312–317. Available online: https://www.aclweb.org/anthology/2021.wanlp-1.38 (accessed on 22 February 2021).
14. Lichouri, M.; Abbas, M.; Benaziz, B.; Zitouni, A.; Lounnas, K. Preprocessing Solutions for Detection of Sarcasm and Sentiment for Arabic. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19–20 April 2021; pp. 376–380. Available online: https://www.aclweb.org/anthology/2021.wanlp-1.49 (accessed on 22 February 2021).
15. Frenda, S.; Ghanem, B.; Montes-y-Gómez, M. Exploration of misogyny in Spanish and english tweets. *CEUR Workshop Proc.* **2018**, *2150*, 260–267.
16. Muaad, A.; Jayappa, H.; Al-Antari, M.; Lee, S. ArCAR: A Novel Deep Learning Computer-Aided Recognition for Character-Level Arabic Text Representation and Recognition. *Algorithms* **2021**, *14*, 216. [CrossRef]

17. Alyafeai, Z.; Al-shaibani, M.S.; Ghaleb, M.; Ahmad, I. Evaluating Various Tokenizers for Arabic Text Classification. 2021, Volume 5. Available online: http://arxiv.org/abs/2106.07540 (accessed on 22 February 2021).
18. Kowsari, K.; Meimandi, K.J.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
19. Huang, J. Detecting fake news with machine learning. *J. Phys. Conf. Ser.* **2020**, *1693*, 012158. [CrossRef]
20. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. *arXiv* **2020**, arXiv:2003.00104.
21. Chola, C.; Benifa, J.V.; Guru, D.S.; Muaad, A.Y.; Hanumanthappa, J.; Al-Antari, M.A.; Gumaei, A.H. Gender Identification and Classification of Drosophila melanogaster Flies Using Machine Learning Techniques. *Comput. Math. Methods Medicine* **2022**, *in press*. [CrossRef]
22. Hanumanthappa, J.; Muaad, A.Y.; Bibal Benifa, J.V.; Chola, C.; Hiremath, V.; Pramodha, M. IoT-Based Smart Diagnosis System for HealthCare. In *Sustainable Communication Networks and Application. Lecture Notes on Data Engineering and Communications Technologies*; Karrupusamy, P., Balas, V.E., Shi, Y., Eds.; Springer: Singapore, 2022; Volume 93. [CrossRef]