Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 3, July 2024, pp: 243-252; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received May 2, 2024; revised May 23, 2024; accepted May 27, 2024; date of publication July 8, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i3.407

How to cite: Halimatus Sa'diah, Mohammad Reza Faisal, Andi Farmadi, Friska Abadi, Fatma Indriani, Muhammad Alkaff, Vugar Abdullayev, "Gender Classification on Social Media Messages Using fastText-base Feature Extraction and Long Short-Term Memory", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 3, pp. 243-252, July 2024.

Gender Classification on Social Media Messages Using fastText-base Feature Extraction and Long Short-Term Memory

Halimatus Sa'diah[1], Mohammad Reza Faisal[1*], Andi Farmadi[1], Friska Abadi[1], Fatma Indriani[1], and Muhammad Alkaff[2,3], and Vugar Abullayev[4]

[1] Department of Computer Science, Lambung Mangkurat University, Banjarbaru, South Kalimantantan, Indonesia
[2] Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
[3] Department of Information Technology, Lambung Mangkurat University, Banjarmasin, South Kalimantantan, Indonesia
[4] Department of Computer Engineering, Azerbaijan State Oil and Industry University, Baku, Azerbaijan

Corresponding author: Mohammad Reza Faisal (e-mail: reza.faisal@ulm.ac.id).

ABSTRACT Currently, social media is used as a platform for interacting with many people and has also become a source of information for social media researchers or analysts. Twitter is one of the platforms commonly used for research purposes, especially for data from tweets written by individuals. However, on Twitter, user information such as gender is not explicitly displayed in the account profile, yet there is a plethora of unstructured information containing such data, often unnoticed. This research aims to classify gender based on tweet data and account description data and determine the accuracy of gender classification using machine learning methods. The method used involves FastText as a feature extraction method and LSTM as a classification method based on the extracted data, while to achieve the most accurate results, classification is performed on tweet data, account description data, and a combination of both. This research shows that LSTM classification on account description data and combined data obtained an accuracy of 70%, while tweet data classification achieved 69%. This research concludes that FastText feature extraction with LSTM classification can be implemented for gender classification. However, there is no significant difference in accuracy results for each dataset. However, this research demonstrates that both methods can work well together and yield optimal results.

INDEX TERMS feature extraction, gender classification, fastText, RNN, LSTM.

I. INTRODUCTION

Social media is currently widely used as a source of interaction for many people [1], one of which is online platforms like Twitter, which provide a space for various individuals to easily share and discuss [2],[3]. Based on this, social media data provides valuable additional information regarding real-time news accessible to the public [4], [5]. This advantage makes social media data useful in gathering up-to-date information [6], [7]. In the meantime, among other social media platforms, Twitter is a popular platform with over 68 million adult users [8],[9] whose genders are not clearly known, thus posing a constraint to directly identify Twitter users, unlike other social media accounts that provide detailed user profile information.

Given the large number of Twitter users in Indonesia, it is clear that they come from various groups. One of these groups is gender. Twitter users can be divided into several groups, one of which is gender. Social media user profiles can be utilized in several fields, namely security, forensic analysis, and commercial domains, but some users may not like to disclose their identities. This factor is why a system that can classify gender is needed. Then, the results of this classification can be utilized by companies or businesses by creating new business strategies to serve their customers [10]. As a result, Twitter only displays the number of followers and account descriptions on the profile page. Therefore, to identify further, it is necessary to recognize through the account description and the form of tweets posted [11].

This social media platform deduces gender from various sources [12], such as usernames, screen names, descriptions,

images, or user-generated content [11]. The gender analysis process can be conducted by applying machine learning capabilities, from the feature extraction stage of the data used to the classification stage.

Based on the research conducted by [13], gender classification was performed using the word2vec feature extraction method and Convolutional Neural Network (CNN) classification, achieving an accuracy of 94.50%. Then, research conducted by [14] employed FastText word embedding for feature extraction and Random Forest classification, resulting in a good accuracy of 70.27%. Furthermore, [15] conducted feature extraction of Twitter data using FastText Embeddings and CNN classification, achieving excellent results with an accuracy of 84.01% by optimizing FastText Embeddings hyperparameters for CNN classification. Additionally, [16] performed text classification using Bi-LSTM and FastText feature extraction with the CBOW architecture model, obtaining an accuracy of 77% [17] also implemented FastText feature extraction for classification, comparing CNN classification results with various feature extraction methods such as FastText, Glove, and Word2Vec, with classification accuracies of 97.2%, 95.8%, and 92.5% respectively, showing the highest accuracy when using FastText feature extraction.

Apart from feature extraction processes affecting classification results, the classification process significantly influences the research outcome. For instance, [18] classified text data using LSTM, achieving an accuracy of 99.56% with 128 LSTM units after ten epochs. Additionally, [19] compared text data classification results from LSTM, CNN, and Simple Neural Network, yielding accuracies of 87%, 82%, and 81%, respectively, indicating LSTM as the superior method for text classification. Moreover, [20] classified novel review data based on positive, negative, and neutral sentiment using LSTM and Naïve Bayes methods, resulting in 72.85% and 67.88% accuracy, respectively, with LSTM outperforming Naïve Bayes.

Considering the superior feature extraction methods from [14]-[17] and the best classification methods from [18]-[20], this research aims to combine the previously best-performing feature extraction method with the best classification method. Therefore, this study aims to combine the feature extraction method that previously yielded the best accuracy with the best classification method, namely using FastText feature extraction with LSTM classification in gender classification. FastText feature extraction has the advantage of being able to handle uncommon words and complex word morphology, as well as its ability to consider subword information in word vector representations. Additionally, the advantage of LSTM classification lies in its complex structure, allowing it to remember long-term information and, thus, be capable of handling long-distance dependencies in sequential data, commonly encountered in text analysis and time series data. Hence, the difference between this and previous research lies in the fact that, as seen in previous studies, there has not yet been found a FastText feature extraction process classified with LSTM, especially for gender classification issues, by combining tweet data and profile description data on Twitter users as in this study [13]. The novelty of this research lies in the combination of methods used for gender classification, which involves utilizing FastText as the feature extraction method and LSTM as the classification method. This is because, based on previous research, there has not yet been found any study that performs gender classification on textual data using this method.

Therefore, the purpose of this research is to identify someone's gender through textual data such as account descriptions and tweets, and also aims to determine the performance of the FastText feature extraction method with LSTM classification when classifying gender. This will certainly provide benefits for future researchers interested in studying gender classification case studies and serve as a reference for selecting the appropriate method.

II. MATERIAL AND METHODS

The research procedure used in this study is as follows. FIGURE 1 depicts the research workflow. The research process begins with data collection, which includes gathering account descriptions and tweets from each account.

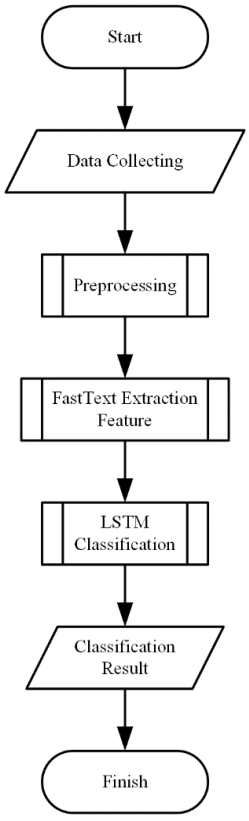

FIGURE 1. Research Workflow

After collecting the data, the next step is preprocessing, which involves labelling, cleaning, stemming, and stop word removal. Following preprocessing, the FastText feature extraction process is performed. Once the feature extraction is

completed, the LSTM classification is carried out for both tweet data, account descriptions, and combined text and description data.

A. DATASET

The dataset used in this study is sourced from research [13], consisting of Twitter data comprising tweets adjusted with the account descriptions based on their Twitter accounts. The data collection process in [13] involves scraping through the Twitter API using search keywords such as "soccer," "game," and "sports." Based on these keywords, several tweet data with similar content from various accounts will be found, and each account's description will be examined as additional data for the labelling process. Subsequently, these data will be collected and labelled based on gender, as proposed in this study. The total number of data used is 2000, with 1000 labelled male and 1000 labeled female. An example dataset can be viewed in TABLE 1.

TABLE 1
Dataset Details

No.	Username	Description	Tweet
1	madeinheavenIV	love is just a dumb luck	*tidak ada pemain yang lebih besar dari sebuah club. tapi ini bukan tentang sepak bola hehe #Barcelona #Egypt*
2	1_yoshino2	IND/ENG She/her Hidup jalan terus tanpa henti	*@bluelockfess Secara emang tahan untuk membuat pemain hebat/ maha karya pak ego, selain itu terikat keinginan kuat tentang sepak bola.*
3	mstrprta	yesterday, today and tommorrow :)	*MPV terus tapi losestreak 11 kali, kena troll terus, rusak ini game*
4	Radithisme	*Sebuah seni tentang perjalan hidup, lewat tulisan. suka tidak suka hidup akan terus berjalan ke depan, meninggalkan kenangan menuju harapan baru.*	*@FWBESS Olahraga merakyat https://t.co/6BrZ0cA59l*
5	sagigirllss	*kadang receh kadang serius*	*Olahraga dulu kak biar kuat menjalani kehidupan https://t.co/reZNQ6pp4Z*

B. PREPROCESSING

The data obtained cannot be directly implemented into the feature extraction process because the data condition still contains many characters that must be corrected first. The preprocessing process begins with labelling the collected data [21], which is divided into two labels: labelling for male and female genders. Next is the cleansing process, which involves removing words with characters that do not contribute to the data analysis [22]. Following that is the stemming process, which converts words with affixes into their base forms [23]. Then, stopword removal is performed to filter out or remove words that do not carry meaning [13]. The process for each preprocessing step can be observed in TABLE 2 as an example.

C. FEATURE EXTRACTION

The preprocessed data will proceed to the feature extraction process [24], [25]. The feature extraction method used is FastText. FastText is an open-source API from Facebook research [26]. This method can quickly and effectively learn word representations. The FastText embedding model follows the skip-gram and vector models, an update from the word2vec models. FastText feature extraction divides words into parts called n-grams, then learns vector representations for each subword. This allows FastText to handle unknown or rare words [27]. Word embedding is a technique for transforming text into vector representations. There are many types of word embeddings, one of them being FastText embeddings. FastText is an extension of Word2Vec embeddings, which were previously used as word embeddings. FastText works by representing each word in several n-grams. This means that each word is broken down into multiple subwords or characters [28].

The FastText feature extraction method has two main advantages. First, this method considers the internal structure of words when learning word representations. Therefore, it is beneficial for morphologically rich languages and infrequent words. Second, FastText works well with n-grams, making it suitable for text research, particularly in gender analysis of language, thus precisely capturing complex human emotions. FastText provides the desired adaptability without sacrificing space and time efficiency by considering the morphological structure of words. FastText offers richer word representations and is capable of handling languages with large and complex vocabularies [29],[30].

D. LSTM

Long Short-Term Memory (LSTM) is a development method of the conventional Recurrent Neural Network (RNN) that only knows one type of network [31]. Long Short-Term Memory (LSTM) is one variation of RNN designed to avoid the problem of long-term dependency in RNNs. LSTM can retain long-term information, which is a major advantage compared to conventional RNNs. Like RNNs, LSTM also consists of recurrent processing modules. The idea behind LSTM is the creation of a pathway connecting the old context to the new context, also known as the cell state, memory cell, or memory pathway. With this pathway, a value from the old context can easily be connected to the new context, if necessary, with minimal modification. One of the advantages of LSTM is its ability to control the information stored or deleted from the cell state through sigmoid gates. These gates allow LSTM to select relevant information to store in long-term memory and ignore irrelevant information [32].

TABLE 2
Example Of Preprocessing Steps Performed

Username	Description	Tweet	Label	Cleansing	Stemming	Stopword Removal
madeinheaven IV	love is just a dumb luck	*tidak ada pemain yang lebih besar dari sebuah club. tapi ini bukan tentang sepak bola hehe #Barcelona #Egypt*	male	*tidak ada pemain yang lebih besar dari sebuah club tapi ini bukan tentang sepak bola hehe Barcelona Egypt*	*tidak ada main yang lebih besar dari buah club tapi ini bukan tentang sepak bola hehe Barcelona Egypt*	*main buah club sepak bola hehe barcelona egypt*

LSTM introduces gate mechanisms: the input gate, forget gate, and output gate [33]. The input gate determines new information to be stored in the cell state and calculates a new value to update the cell state. The forget gate determines which information from the cell state will be discarded, while the output gate determines the output value based on the cell state [22], [34], [35].

The gate structure is implemented using the sigmoid function (σ). The sigmoid value ranges from 0 to 1, indicating how much information is allowed to pass through. The illustration of LSTM can be seen in FIGURE 2 [36]. In the first step, the LSTM unit processes information from the previous memory state through the forget gate to determine which information to forget from the memory state. The equation used is shown in equation (1) [37]:

$$f_t = \sigma\,(W_f . [h_{t-1}, x_t] + b_f) \quad (1)$$

In equation (1) f_t represents the forget gate, W_f is the weight for the *f* gate value, h_{t-1} is the output from the previous LSTM unit (at time t -1), x_t is the input to the current LSTM unit, and b_f is the bias for f. Next, the LSTM unit determines what information to store in memory. On one side, the input gate decides which information to update, and on the other side, the candidate layer influences the tanh vector [38]. The equations used in this step are shown in equations (2) and (3) :

$$i_t = \sigma\,(W_i . [h_{t-1}, x_t] + b_f \quad (2)$$
$$C_t = tanh\,(W_C . [h_{t-1}, x_t + b) \quad (3)$$

Equation (2) i_t represents the input gate, W_i is the weight for the *i* gate value. Next, in equation (3) represents the memory value at time *t*. The hyperbolic tangent function (tanh) produces values between -1 and 1. W_C is the weight or weight matrix used to multiply the input. Then, the LSTM unit combines the two parts above to update the memory status, as shown in equation (4) :

$$C_t = f_t * C_{t-1} + i_t * C_t) \quad (4)$$

Finally, the LSTM unit uses the output gate to control the memory status that needs to be outputted, as shown in equations (5) and (6):

$$o_t = \sigma\,(W_o . [h_{t-1}, x_t + b_o) \quad (5)$$
$$h_t = o_t * tanh\,(Ct) \quad (6)$$

In equation (5) o_t represents the output value at time *t*, W_o is the weight or weight matrix used to multiply the input. Meanwhile, in equation (6) h_t is the output at time *t.* and o_t is the output gate value at time *t*, which controls how much information can be outputted. These equations form the basis of how the LSTM method works, including the input, forget, and output gates [39], [40]. In the LSTM architecture, activation functions, namely Sigmoid and Tanh, are present. The sigmoid function transforms values between -1 and 1 into the range of 0 and 1, while the tanh function regulates values passing through the network always to remain close to -1 and 1. The LSTM model construction in this research implements a sequence model consisting of three layers, which include the Embedding layer, LSTM layer, Dropout layer, and Dense layer. The Embedding layer is a step to convert words into vector form. The approach used to represent word vectors and arrays is in real numbers. In the LSTM layer, there are parameters such as memory units, input shape obtained from word embedding results, and dropout. The dropout parameter is used to prevent overfitting/underfitting, with values typically ranging between 0 and 1. The Dense layer is used to add a fully-connected layer based on the number of classes specified [41].

E. EVALUATION RESULT

Evaluation metrics are parameters used to measure the quality of a model or machine learning algorithm. The evaluation

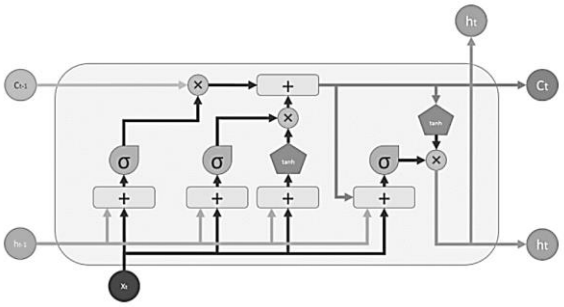

FIGURE 2. The General Architecture of LSTM

metrics used in this study are Accuracy, Precision, Recall, and F1-Score. Each evaluation metric is formulated as follows: Accuracy is the value representing the comparison of True Positive (TP) and True Negative (TN) predictions with the total number of data. The formula used can be seen in the equation (7). Precision is the value representing the comparison of True Positive (TP) predictions with the total

number of data predicted positive. The formula used is shown in equation (8). Recall is the value representing the comparison of True Positive (TP) predictions with the total number of truly positive data. There is a difference between precision and recall, where precision involves the False Positive (FP) variable, while recall involves the False Negative (FN) variable. The formula used is shown in equation (9). F1-Score is the value representing the weighted average comparison of precision and recall. The formula used can be seen in equation (10) [41] :

$$Accuracy = TP+TN / (TP+FN)+(FP+TN) \quad (7)$$

$$Precision = TP / TP+FP \quad (8)$$

$$Recall = TP / TP+FN \quad (9)$$

$$F1\text{-}Score = 2TP / 2TP + FN + FP \quad (10)$$

III. RESULT

This section is dedicated to evaluating the results of the gender classification research based on account description data, tweet data from each account, and the combined tweet and description data. The process involves utilizing the FastText feature extraction and LSTM classification methods.

A. PREPROCESSING RESULT

As an initial step, the preprocessing of the initial data shown in TABLE 1 starts from labeling the data as male or female labels, data cleaning involves removing punctuation or characters from the text, stemming by converting words with affixes into their base form, and performing stopword removal or filtering data to remove data deemed irrelevant. An example of this process can be seen in TABLE 2. Meanwhile, the results of this data preprocessing stage can be observed in TABLE 3.

TABLE 3
Preprocessed Dataset Results

No.	Username	Description	Tweet
1	madeinheavenIV	love is just a dumb luck	*main buah club sepak bola hehe barcelona egypt*
2	1_yoshino2	indeng she her hidup jalan henti	*bluelockfess emang tahan main hebat maha karya ego*
3	mstrprta	yesterday today and tomorrow	*mpv losestreak kali kena troll rusak game*
4	radithisme	buah seni jalan hidup tulis suka suka hidup jalan	*fwbess olahraga rakyat*
5	sagigirllss	kadang receh kadang serius	*olahraga kak biar kuat jalan hidup*

The preprocessed data will undergo further processing, namely the feature extraction process.

B. FEATURE EXTRACTION RESULT

The feature extraction method used in this study is the FastText method, which maps each word into vector form. The feature extraction process begins by mapping each word from the tweet sentences or descriptions. After mapping these sentences into words called n-grams, the FastText algorithm learns each of these words and transforms them into vector or numerical forms. Once all the words have been transformed into vector form, these vectors are arranged into document or sentence form to take the average vector representation of each token within them. This arrangement of each vector is useful for the labelling process of each document.

An example of the FastText feature extraction results for each word can be seen in TABLE 4. Then, the process of arranging words for each vector to be labelled or classified can be observed in TABLE 5.

TABLE 4
FastText Feature Extraction Results

	word	V1	V2	……	V100
0	astroarena	-0.19306	-0.09064	……	-0.1006
1	sports	-0.19306	-0.09064	……	-0.07
2	journalist	-0.19306	-0.09064	……	-0.03294
3	bio	-0.19306	-0.09064	……	-1.65811
4	ler	-0.19306	-0.09064	……	0.261279
5	citizens	-0.19306	-0.09064	……	-0.30094
6	fans	-0.19306	-0.09064	……	-0.63107
7	liverpool	-0.19306	-0.09064	……	0.155988
8	mu	-0.19306	-0.09064	……	-1.85521
9	arsenal	-0.19306	-0.09064	……	-0.88895
10	chelsea	-0.19306	-0.09064	……	0.803139
….	….	….	….	……	….
13305	drunk	-0.19306	-0.09064	……	0.039582

TABLE 5
The Arrangement of Words for Each Vector

	label	W1V1	W1V2	……	W30V100
0	pria	-0.2339	0.057187	……	0
1	pria	-0.29086	-0.22218	……	0
2	pria	0.222892	-0.41852	……	0
3	pria	0.004828	0.006421	……	0
4	pria	1.298398	-1.25113	……	0
5	pria	-0.22319	0.368168	……	0
6	wanita	0.192942	-0.31518	……	0
7	pria	-0.22292	-0.37575	……	0
8	pria	0.428628	-0.10531	……	0
9	pria	0.357637	-0.27472	……	0
10	pria	-0.04764	-0.31127	……	0
….	….	….	….	……	….
1999	pria	-0.59832	-0.49986	……	0

When the feature extraction results are obtained, vector values are obtained in each document or sentence labeled in the system to enter the classification stage later, as shown in TABLE 5, the next step is to classify the data using LSTM.

C. LSTM CLASSIFICATION RESULT

In the input layer process of LSTM, there are several considerations for processing tweet data, description data, or their combination. The word sequence length to be implemented in the classification process varies depending on the data category. The maximum sequence length for tweet data ranges from 12 to 49 words, while for description data, it ranges from 8 to 30 words. On the other hand, the combined tweet and description data have a maximum sequence length ranging from 20 to 79 words. This grouping of word lengths is done to differentiate the classification process, where the LSTM classification will classify gender based on three types

of data: tweet data, account description data, and their combination.

The parameters used for each data type can be seen in TABLE 6. Subsequently, the LSTM classification results for tweet data, description data, or their combination can be observed in TABLE 7, and the accuracy results are depicted in FIGURE 3.

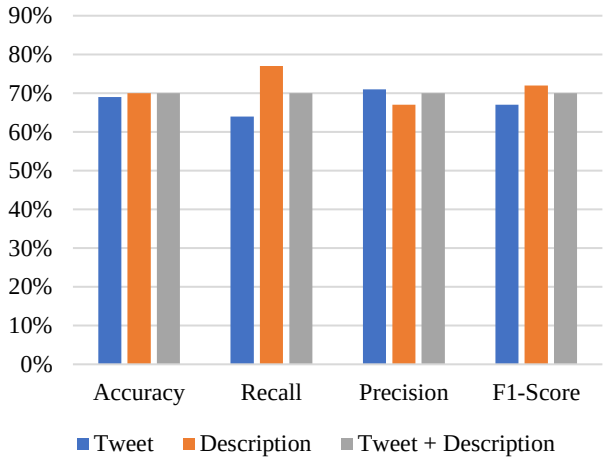

FIGURE 3. Comparison Result LSTM Classification

In TABLE 7, the highest accuracies obtained from testing the parameters for each data implementation, namely tweet data, description data, and the combined tweet and description data, are presented. The parameters used for each data in the classification process are shown in TABLE 6. It can be observed in TABLE 6 that the difference in parameters lies only in the input shape. The variance in input shape for each data occurs because the input shape is adjusted to the number of data used. This is because in the LSTM classification process, the number of words used during classification is determined. This applies to the classification of tweet data, description data, and the combined data of tweet and description. For example, one of the input shapes used for tweet data is (1,4900), which means that each LSTM model used accepts one example at a time, and each example has a sequence length of 4900 tokens. Another input shape is (10,490), indicating that it processes 490 tokens at a time, repeated 10 times.

The determination of input shape is not arbitrary but considers the number of words used. For instance, in tweet data, the maximum result of multiplying the processing time with the data is 4900. Therefore, when these two numbers are multiplied, the resulting input shape must be 4900. This process of determining the input shape also applies to determining the input shape for description data and the combined tweet and description data. The variance in input shape is one of the factors determining the search for the best accuracy in the LSTM classification process. Furthermore, in FIGURE 3, the LSTM classification results for gender data for each dataset, whether tweet data, description data, or the combined data of tweet and description, are shown. It can also be observed in FIGURE 3 that the accuracy results for description data and the combined tweet and description data are the same at 70%, while the accuracy for tweet data is 69%. Thus, it can be concluded that FastText feature extraction and LSTM classification perform better when implemented on description data or the combined tweet and description data.

IV. DISCUSSION

The final result of gender classification using the FastText feature extraction method and LSTM classification in this study is consistent with previous research conducted by [14] - [20], where this study can perform the classification process and achieve good accuracy. This study successfully achieved the highest accuracy of 70% in classifying combined data. Similar to the study by [13], this research used similar data collected based on tweets and profile descriptions from Twitter accounts.

This research indicates that gender classification using social media data from Twitter can be successfully performed. By utilizing the FastText feature extraction method to convert text data into vectors and LSTM classification, good accuracy can be achieved for each type of data used. These findings provide a strong foundation for discussion in line with the previous research objectives.

The classification results for tweet data, description data, and their combination did not show significant differences in accuracy. When tweet data was extracted using FastText and classified using LSTM, an accuracy of 69% was achieved, while implementing the method on description data and the combined data resulted in an accuracy of 70%. Looking at the obtained results, there was only a 1% difference in the accuracy produced.

The difference in accuracy generated is partly due to the difference in the amount of data from each dataset and of course, from the difference in parameters implemented, especially the different input shape parameters in each dataset to adjust to the amount of data used in each dataset. This difference in input shape influences the time of data analysis and the length of data used to obtain the most optimal results. However, because the other parameters have the same value, it is one of the reasons why the accuracy generated does not differ significantly from each dataset, considering that tweet data, description data, and combined data are all text data obtained in the same way. During the classification process, these data are only distinguished by the length of words used for each data. Compared with previous research [10] conducting gender classification, clear differences in the resulting accuracies can be observed. In research [10], gender classification was performed using the word2vec extraction method and classification using CNN with the same research flow. Thus, it is obtained that the data for gender classification in this study is a research study that can be further investigated

TABLE 6
LSTM Classification Parameter

	Alpa	Vector Size	Min Count	Window	Word Embedding Epoch	Input Shape	Layer	Neuron	Batch Size
Tweet	0.01	100	1,3,5	3,5,7	10, 50, 100	(1,4900), (10,490), (14,350), (2,2450), (20,245), (25,196), (4,1225), (5,980), (7,700)	1,2,3	32, 64, 128	32,64
Description	0.01	100	1,3,5	3,5,7	10, 50, 100	(1,3000), (10, 300), (12,250), (15,200), (2,1500), (3,1000), (4,750), (5,600), (6,500), (8,375)	1,2,3	32, 64, 128	32,64
Tweet + Description	0.01	100	1,3, 5	3,5,7	10, 50, 100	[(1,4900), (1,3000)], [(10,490), (10,300)], [(100,49), (100,30)], [(2,2450), (2,1500)], [(20,245), (20,150)], [(25,196), (25,120)], [(4,1225), (4,750)], [(5,980), (5,600)], dan [(50, 98), (50,60)]	1,2,3	32, 64, 128	32,64

TABLE 7
LSTM Classification Result

	Alpa	Vector Size	Min Count	Window	Word Embedding Epoch	Input Shape	Layer	Neuron	Batch Size	Akurasi	Recall	Presisi	F1-Score
Tweet	0.01	100	1	3	50	14,350	3	64	64	69%	64%	71%	67%
Description	0.01	100	1	7	100	8,375	2	32	64	70%	77%	67%	72%
Tweet + Description	0.01	100	1	5	100	[(4,1225), (4,750)]	3	128	64	70%	70%	70%	70%

with other method combinations. Additionally, it provides results that show that the combination of FastText feature extraction and LSTM classification yields quite good results when implemented on text data for gender classification. However, in terms of predicting positive outcomes based on actual data, the highest recall of 77% was achieved when classifying description data. Thus, this method combination can predict data well based on actual data through description data prediction. Furthermore, the highest precision of 71% was obtained when classifying tweet data, and the highest F1-Score value of 72% was achieved when classifying description data. The number of correctly predicted data can be clearly understood in TABLE 8 and FIGURE 4.

TABLE 8
Comparison Result Confusion Matrix

Dataset	Confusion Matrix Result			
	True Positive	False Negative	True Negative	False Positive
Tweet	64	36	74	26
Description	77	23	63	37
Tweet + Description	70	30	70	30

Based on TABLE 8 and FIGURE 4, the data successfully predicted as true positive and true negative are 138 for tweet data and 140 for both description data and the combined tweet and description data. Based on these prediction results, the difference in accuracy between each dataset is not significant, as the number of correctly predicted data is also not significantly different.

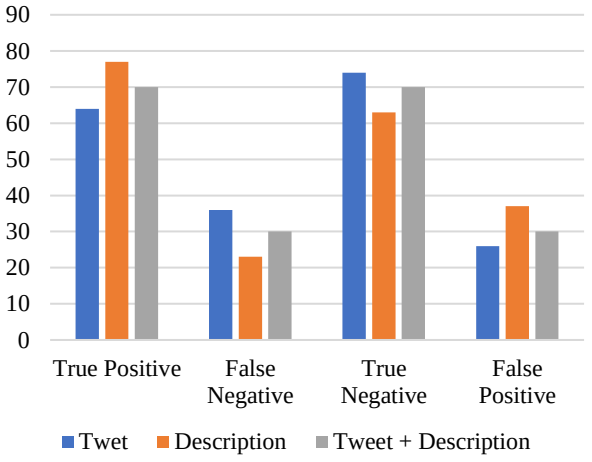

FIGURE 4. Comparison Result Confusion Matrix

The main reason for the accuracy not varying much when implemented on tweet data, description data, and combined data is that, upon closer examination, tweet data and description data share many similarities, with the main difference being the length of the text. Additionally, the distinctive writing style of individuals in their daily tweets and descriptions contributes to the minimal difference in accuracy obtained. Furthermore, the weaknesses of this research lie in the combination of methods used and the limitations in parameter testing. The combination of methods used may not be optimal to achieve the desired results, and limited parameter testing restricts the ability to explore various configurations that may yield better results. These parameters are crucial in determining model performance, in addition to the selection of methods used. Additionally, other stages may need to be added or modified to make the

research results more optimal. This indicates that this research still has room for further development to achieve maximum results. It was considered that research [13] achieved an accuracy of 94.50% on the same dataset using different methods. This could be due to the different combinations of methods implemented on the research data and the varying parameters used, significantly influencing the research results. This study did not address solutions for other steps that could be taken to address this issue, but it serves as a note for future researchers.

Based on the research, it is important to retest the same data using different methods to determine whether or not the accuracy can be improved. This can be done by comparing the results of feature extraction method implementations or comparing them with the classification methods performed. Considering that the data has only been implemented in research [13], there is still room for further development. Additionally, the findings obtained from this research have contributed to knowledge about the methods that can be used for gender classification based on the tweet and account description data through the Twitter social media platform. It also provides information about the steps and processes that can be undertaken when using the FastText feature extraction method combined with the LSTM classification method. This study serves as new knowledge, especially in determining the length of words categorized as tweet data, description data, or their combination. It also informs about the process of labelling vector data resulting from feature extraction.

V. CONCLUSION

This study aims to investigate the FastText feature extraction method and LSTM classification method in gender classification processes. In summary, the findings of the gender classification study using the combination of these two methods resulted in the highest accuracy of 70% when classifying description data and the combination of tweet and description data. Additionally, an accuracy of 69% was achieved when implemented on tweet data alone, indicating that classification on description data or their combination only outperformed tweet data by 1%. This research aligns with previous studies showing that the combination of FastText feature extraction and LSTM classification performs well. This study contributes to utilizing methods in study data. It provides insights into how the FastText feature extraction and LSTM classification methods are applied to tweet data, description data, and their combination.

However, a limitation of this research lies in the combination of methods employed and the accuracy obtained, which only reached 70%. Therefore, there is a need for additional steps to improve accuracy further. Besides adding additional steps, it may also involve combining one of the methods to obtain different accuracy comparisons when implemented with different method combinations.

Based on these results, it is recommended to compare the FastText feature extraction method with other methods such as word2vec or TF-IDF, or to compare the LSTM method with other RNN methods in gender classification. This increases the likelihood of achieving optimal accuracy for this case study. Future researchers will undoubtedly explore comparing method combinations to achieve even better model performance.

VI. ACKNOWLEDGMENT

We express our utmost gratitude to all parties involved in the completion of this research, especially to Lambung Mangkurat University, particularly the Computer Science program, which served as the platform for completing this research, and to King Abdulaziz University, which also contributed by providing ideas and inputs until the successful completion of this research. We also extend our thanks to the team members who have contributed significantly to the collaboration, which undoubtedly played a key role in achieving the results of this research.

REFERENCES

[1] F. Aftab *et al.*, "A Comprehensive Survey on Sentiment Analysis Techniques," *Int. J. Technol.*, vol. 14, no. 6, pp. 1288–1298, 2023, doi: 10.14716/ijtech.v14i6.6632.

[2] M. Cormier and M. Cushman, "Innovation via social media – The importance of Twitter to science," *Res. Pract. Thromb. Haemost.*, vol. 5, no. 3, pp. 373–375, 2021, doi: 10.1002/rth2.12493.

[3] C. J. Powers *et al.*, "Using artificial intelligence to identify emergency messages on social media during a natural disaster: A deep learning approach," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100164, 2023, doi: 10.1016/j.jjimei.2023.100164.

[4] M. R. Faisal, I. Budiman, F. Abadi, D. T. Nugrahadi, M. Haekal, and I. Sutedja, "Applying Features Based on Word Embedding Techniques to 1D CNN for Natural Disaster Messages Classification," *2022 5th Int. Conf. Comput. Informatics Eng. IC2IE 2022*, no. December, pp. 192–197, 2022, doi: 10.1109/IC2IE56416.2022.9970188.

[5] K. Y. Firlia, M. R. Faisal, D. Kartini, R. A. Nugroho, and F. Abadi, "Analysis of New Features on the Performance of the Support Vector Machine Algorithm in Classification of Natural Disaster Messages," *Proc. - 2021 4th Int. Conf. Comput. Informatics Eng. IT-Based Digit. Ind. Innov. Welf. Soc. IC2IE 2021*, no. September, pp. 317–322, 2021, doi: 10.1109/IC2IE53219.2021.9649107.

[6] M. Dou, Y. Wang, Y. Gu, S. Dong, M. Qiao, and Y. Deng, "Disaster damage assessment based on fine-grained topics in social media," *Comput. Geosci.*, vol. 156, no. March, p. 104893, 2021, doi: 10.1016/j.cageo.2021.104893.

[7] Muhamad Fawwaz Akbar, Muhammad Itqan Mazdadi, Muliadi, Triando Hamonangan Saragih, and Friska Abadi, "Implementation of Information Gain Ratio and Particle Swarm Optimization in the Sentiment Analysis Classification of Covid-19 Vaccine Using Support Vector Machine," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 261–270, 2023, doi: 10.35882/jeeemi.v5i4.328.

[8] A. Karami *et al.*, "2020 U.S. presidential election in swing states: Gender differences in Twitter conversations," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, 2022, doi: 10.1016/j.jjimei.2022.100097.

[9] E. Cano-Marin, M. Mora-Cantallops, and S. Sánchez-Alonso, "Twitter as a predictive system: A systematic literature review," *J. Bus. Res.*, vol. 157, no. December 2022, 2023, doi: 10.1016/j.jbusres.2022.113561.

[10] A. S. Zakia, Indriati, and Marji, "Klasifikasi Jenis Kelamin Pengguna Twitter dengan menggunakan Metode BM25 dan K-Nearest Neighbor (KNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 10, pp. 3331–3337, 2020, [Online]. Available: http://j-ptiik.ub.ac.id.

[11] M. Vicente, F. Batista, and J. P. Carvalho, "Gender Detection Of Twitter Users Based On Multiple Information Sources," *ISCTE-IUL Repos.*, no. 351, 2018.

[12] E. Fosch-Villaronga, A. Poulsen, R. A. Søraa, and B. H. M. Custers, "A little bird told me your gender: Gender inferences in social media," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102541, 2021, doi: 10.1016/j.ipm.2021.102541.

[13] F. A. Mubarok, M. Reza Faisal, D. Kartini, D. T. Nugrahadi, and T. H. Saragih, "Gender Classification of Twitter Users Using Convolutional Neural Network," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 1, pp. 79–92, 2023, doi: 10.30812/matrik.v23i1.3318.

[14] Y. Gunawan, J. C. Young, and A. Rusli, "FastText Word Embedding and Random Forest Classifier for User Feedback Sentiment Classification in Bahasa Indonesia," *Ultim. J. Tek. Inform.*, vol. 13, no. 2, 2021.

[15] F. Alfariqi, W. Maharani, and J. H. Husen, "Klasifikasi Sentimen pada Twitter dalam Membantu Pemilihan Kandidat Karyawan dengan Menggunakan Convolutional Neural Network dan Fasttext Embeddings," *e-Proceeding Eng.*, vol. 7, no. 2, pp. 8052–8062, 2020.

[16] Y. V. Aritonang, D. P. Napitupulu, M. H. Sinaga, and J. Amalia, "Pengaruh Hyperparameter pada Fasttext terhadap Performa Model Deteksi Sarkasme Berbasis Bi-LSTM," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2612–2625, 2022, doi: 10.35957/jatisi.v9i3.1331.

[17] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "the Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (Cnn) Text Classification," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 2, pp. 349–359, 2022.

[18] G. S. . Murthy, S. R. Allu, B. Andhavarapu, M. Bgadi, and M. Belusonti, "Text based Sentiment Analysis using Long Short Term Memory (LSTM)," *Int. J. Eng. Res. Technol.*, vol. 9, no. 05, pp. 299–303, 2020.

[19] A. C. M. V. Srinivas, C. Satyanarayana, C. Divakar, and K. P. Sirisha, "Sentiment Analysis using Neural Network and LSTM," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1074, no. 1, p. 012007, 2021, doi: 10.1088/1757-899x/1074/1/012007.

[20] M. A. Nurrohmat and A. SN, "Sentiment Analysis of Novel Review Using Long Short-Term Memory Method," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 3, p. 209, 2019, doi: 10.22146/ijccs.41236.

[21] S. Monika Nooralifa, M. Reza Faisal, F. Abadi, R. Adi Nugroho, J. A. Yani Km, and K. Selatan, "Identifikasi otomatis pesan saksi mata pada media sosial saat bencana gempa," *Kumpul. J. Ilmu Komputer(KLIK)*, vol. 08, no. 2, p. 129, 2021.

[22] M. R. Faisal *et al.*, "LSTM and Bi-LSTM Models For Identifying Natural Disasters Reports From Social Media," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 4, pp. 241–249, 2023.

[23] M. Padhilah *et al.*, "Implementasi Neural Network Multilayer Perceptron Dan Stemming Nazief & Adriani Pada Chatbot Faq Prakerja," *J. Sains Komput. Inform. (J-SAKTI*, vol. 6, no. 2, pp. 671–685, 2022.

[24] M. R. Faisal, R. A. Nugroho, R. Ramadhani, F. Abadi, R. Herteno, and T. H. Saragih, "Natural disaster on twitter: Role of feature extraction method of word2vec and lexicon based for determining direct eyewitness," *Trends Sci.*, vol. 18, no. 23, 2021, doi: 10.48048/tis.2021.680.

[25] M. Khairie, M. R. Faisal, R. Herteno, I. Budiman, F. Abadi, and M. I. Mazdadi, "The Effect of Channel Size on Performance of 1D CNN Architecture for Automatic Detection of Self-Reported COVID-19 Symptoms on Twitter," *2023 Int. Semin. Intell. Technol. Its Appl. Leveraging Intell. Syst. to Achieve Sustain. Dev. Goals, ISITIA 2023 - Proceeding*, no. August, pp. 621–625, 2023, doi: 10.1109/ISITIA59021.2023.10220444.

[26] B. Darmawan, A. Dwi Laksito, M. Resa Arif Yudianto, and A. Sidauruk, "Analisis Perbandingan Ekstraksi Fitur Teks pada Sentimen Analisis Kenaikan Harga BBM," *Krea-TIF J. Tek. Inform.* , vol. 11, no. 1, pp. 53–63, 2023, doi: 10.32832/krea-tif.v11i1.13819.

[27] S. Ghosal and A. Jain, "Depression and Suicide Risk Detection on Social Media using fastText Embedding and XGBoost Classifier," *Procedia Comput. Sci.*, vol. 218, pp. 1631–1639, 2022, doi: 10.1016/j.procs.2023.01.141.

[28] S. A. Shalehah and Y. S. Triana, "Analisa Kinerja RNN Menggunakan FastText Embedding terhadap Ulasan Peduli Lindungi di Masa Covid-19," *Mercu Buana*, pp. 1–20, 2022.

[29] S. Sadiq, T. Aljrees, and S. Ullah, "Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets," *IEEE Access*, vol. 11, no. August, pp. 95008–95021, 2023, doi: 10.1109/ACCESS.2023.3308515.

[30] S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth, and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," *IEEE Access*, vol. 8, pp. 26172–26189, 2020, doi: 10.1109/ACCESS.2020.2971087.

[31] S. Song *et al.*, "Research on a working face gas concentration prediction model based on LASSO-RNN time series data," *Heliyon*, vol. 9, no. 4, 2023, doi: 10.1016/j.heliyon.2023.e14864.

[32] J. Pardede and I. Pakpahan, "Analisis Sentimen Penanganan Covid-19 Menggunakan Metode Long Short-Term Memory Pada Media Sosial Twitter," *J. Publ. Tek. Inform.*, vol. 2, no. 3, pp. 12–25, 2023.

[33] M. Muñoz-Organero, P. Callejo, and M. Á. Hombrados-Herrera, "A new RNN based machine learning model to forecast COVID-19 incidence, enhanced by the use of mobility data from the bike-sharing service in Madrid," *Heliyon*, vol. 9, no. 6, p. e17625, 2023, doi: 10.1016/j.heliyon.2023.e17625.

[34] V. Matoušek, "Application of LSTM Neural Networks in Language Modelling," *Univ. West Bohemia, Fac. Appl. Sci. Dep. Cybern. Univerzitn´ı 22, Plzen, Czech rep*, no. June 2018, 2013, doi: 10.1007/978-3-642-40585-3.

[35] M. R. Faisal *et al.*, "A Social Community Sensor for Natural Disaster Monitoring in Indonesia Using Hybrid 2D CNN LSTM," *ACM Int. Conf. Proceeding Ser.*, no. December, pp. 250–258, 2023, doi: 10.1145/3626641.3626932.

[36] I. Budiman, M. R. Faisal, D. T. Nugrahadi, Muliadi, M. K. Delimayanti, and S. E. Prastya, "Harvesting Natural Disaster Reports from Social Media with 1D Convolutional Neural Network and Long Short-Term Memory," *2023 8th Int. Conf. Informatics Comput. ICIC 2023*, no. January, pp. 1–6, 2023, doi: 10.1109/ICIC60109.2023.10382045.

[37] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *J. Phys. Conf. Ser.*, vol. 2171, no. 1, 2022, doi: 10.1088/1742-6596/2171/1/012021.

[38] W. K. Sari, D. P. Rini, and R. F. Malik, "Text Classification Using Long Short-Term Memory With GloVe Features," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 5, no. 2, p. 85, 2020, doi: 10.26555/jiteki.v5i2.15021.

[39] C. Wang, D. Han, Q. Liu, and S. Luo, "A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM," *IEEE Access*, vol. 7, pp. 2161–2168, 2019, doi: 10.1109/ACCESS.2018.2887138.

[40] A. Ajitha, M. Goel, M. Assudani, S. Radhika, and S. Goel, "Design and development of Residential Sector Load Prediction model during COVID-19 Pandemic using LSTM based RNN," *Electr. Power Syst. Res.*, vol. 212, no. October 2021, p. 108635, 2022, doi: 10.1016/j.epsr.2022.108635.

[41] N. P. S. Wati and C. Pramartha, "Penerapan Long Short Term Memory dalam Mengklasifikasi Jenis Ujaran Kebencian pada Tweet Bahasa Indonesia," *J. Nas. Teknol. Inf. dan Apl.*, vol. 1, no. 1, pp. 755–762, 2022.

BIBLIOGRAPHY

Halimatus Sa'diah is a student at Lambung Mangkurat University who began her education in 2018 in the Department of Computer Science. Her current research field is Data Science. In addition, her final project includes research focused on gender classification based on text data through the X social media platform. The aim of this research effort is to determine gender based on tweets made by individuals or based on account descriptions from the X social media platform.

Mohammad Reza Faisal was born in Banjarmasin. Following his graduation from high school, he pursued his undergraduate studies in the Informatics department at Pasundan University in 1995, and later majored in Physics at Bandung Institute of Technology in 1997. After completing his bachelor's program, he gained experience as a training trainer in the field of information technology and software development. Since 2008, he has been a lecturer in computer science at Universitas Lambung Mangkurat, while also pursuing his master's program in Informatics at Bandung Institute of Technology in 2010. In 2015, he furthered his education by pursuing a doctoral degree in Bioinformatics at Kanazawa

University, Japan. To this day, he continues his work as a lecturer in Computer Science at Universitas Lambung Mangakurat. His research interests encompass Data Science, Software Engineering, and Bioinformatics.

Andi Farmadi is a senior lecturer in the Computer Science program at Lambung Mangkurat University. He has been teaching since 2008 and currently serves as the Head of the Data Science Lab since 2018. He completed his undergraduate studies at Hasanuddin University and his graduate studies at Bandung Institute of Technology. His research area, up to the present, focuses on Data Science. One of his research projects, along with other researchers, published in the International Conference of Computer and Informatics Engineering (IC2IE), is titled "Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," and this research was published in 2021.

FRISKA ABADI finished her bachelor's degree in Computer Science from Universitas Lambung Mangakurat in 2011. Subsequently, in 2016, she obtained her master's degree from the Department of Informatics at STIMIK Amikom, Yogyakarta. Following that, she joined Universitas Lambung Mangakurat as a lecturer in Computer Science. Currently, she holds the position of head of the software engineering laboratory. Her current area of research revolves around software engineering.

Fatma Indriani is a lecturer in the Department of Computer Science, Lambung Mangkurat University. Her research interest is focused on Data Science. Before becoming a lecturer, she completed her undergraduate program in the Informatics Department at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia in 2012. And her latest education is a doctorate degree in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The research fields she focuses on are Data Science and Bioinformatics.

Muhammad Alkaff completed his undergraduate studies in Computer Science at Brawijaya University, Malang. He advanced his education with a master's degree from the Informatics Department at the Sepuluh Nopember Institute of Technology, Surabaya. Joining Universitas Lambung Mangkurat as a lecturer in 2015, he went on to pursue a Ph.D. at the Computer Science Department of King Abdulaziz University in Saudi Arabia in 2022. His research is primarily focused on the areas of Machine Learning and Reinforcement Learning.

Vugar Abdullayev was born in Azerbaijan. received the B.S. degree in Automatics and control of technical systems speciality from the Azerbaijan State Oil and İndustry University (ASOİU), Baku, Azerbaijan, M.S. degree in Manufactury Automation and Informatics speciality from the Azerbaijan State Oil and İndustry University (ASOİU), Baku, Azerbaijan in 2000, and a Ph.D. degree from - Institute of Cybernetics of Azerbaijan National Academy of Sciences in 2005. In 2002-2004 – Dr. Vugar Abdullayev has been expert on IT and Payment systems department in the Azerbaijan Central Bank. In 2004-2012 – Dr Vugar Abdullayev has been an Researcher, head researcher in the Institute of Cybernetics of Azerbaijan National Academy of Sciences, Baku, Azerbaijan. Since 2012, he is an doctor of technical sciences, Associate Professor at Azerbaijan State Oil and Industry University, Department of Computer Engineering. He is author of 85 scientific papers. His researchers related to the study of the cyber physical systems, IoT, big data, smart city and information technologies, cloud computing, computational complexity, machine learning (artificial intelligence), and behavioral sciences computing. He has published 20 book chapters and 10 edited books (calling for book chapters - Taylor and Francis) in healthcare ecosystem.