Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining

Mohammed Abdulhameed Al-Shabi
Taibah University

Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining

M. A. Al-Shabi[1†]
mshaby@taibahu.edu.sa malshabi@yahoo.com
Department of Management Information System,
College of Business Administration,
Taibah University, Saudi Arabia

Summary
With the emergence of social media that includes billions of people from around of the world interacting and sharing data, information, feelings and opinions among themselves on various topics, there is a huge amount of social media data generated by users to explore opinions and analyze emotions. Sentiment analysis aims to classify the polarity of a text based on the writer's opinion, by revealing the positive, negative or neutral sentiments about a particular subject. It is used in marketing, customer service and other fields. State of art approaches for Sentiment analysis are classified in two categories: the first one depends on Machine learning techniques and data mining by training a model on a set of labeled data. Whereas, the second category lexicon-based ones give specific weights for each word according to polarity of the word which it belongs, and thus identify the sentiments by comparing text words with pre-prepared lexicons.
This study relies on the methodology of sentiment analysis based on the lexicon-based, it focuses on five of the most important and well-known lexicons used in the field of sentiment analysis on Twitter data, such as(VADER, SentiWordNet, SentiStrength, Liu and Hu opinion lexicon and AFINN-111).
It provides an assessment of the performance of these lexicons in Twitter polarity classification by comparing the overall classification accuracy and the F1-measure The Results show that the accuracy of classification using Vader lexicon were higher for positive and negative sentiments.

Key words:
Sentiment Analysis, Opinion Mining, Lexicons, Mining Social Media.

1. Introduction

Accompanying the increased use of social media such as Facebook, twitter, and so on, provides rich data and information sources which can be invested and used for mining opinions and user behavior analysis. Sentiment analysis is generally used to determine the writer's emotions about a topic that can express an opinion or emotional state while writing or an intentional feeling communicating it. The classification of polarity is one of the basic processes in emotions analysis, i.e. revealing positive, negative or neutral emotions that the text holds [1].

State of art approaches for Sentiment analysis are classified in two categories:

- depends on Machine learning techniques and data mining by training a model on a set of labeled data.
- Depending on lexicon-based ones give specific weights for each word according to polarity of the word which it belongs, and thus identify the sentiments by comparing text words with pre-prepared lexicons.

In particular, companies are interested in knowing the opinions of their customers regarding its products and services that its provide and exerting a lot of effort and money towards this. Considering that Twitter includes millions of people exchanging opinions and emotions with the assumption that there is a relationship between public opinion and real world events, in this paper we will deal with small texts from Twitter. this paper attempts to evaluate the most important lexicons used in the field of sentiment analysis On Twitter data, including(VADER, SentiWordNet, SentiStrength, Liu and Hu opinion lexicon and AFINN-111). And the following sections describes the previous five lexicons and how they are used, then shows the used data sets, reaching results and recommendations.
We will apply these lexicons to two Twitter data sets(Stanford and Sandars) without pre-processing this data. This is achieved by writing a special investigation for these lexicons, and then discussing the results of applying these lexicons to assess its accuracy in a polar classification task.

2. Related works

The Musto et al. [2] By comparing the widespread lexical resources such as SentiWordNet, WordNet-Affect, MPQA and SenticNet. The effectiveness of the methodology was studied on the Stanford Twitter Sentiment (STS) and SemEval-2013. A tweet-based method has been suggested for smaller segments, based on punctuation and connecting words, and the overall opinion of the tweet is the sum of the degree of opinion of the smaller segments.
SentiWordNet and MPQA performed best while the worst was WordNet-Affect and the results obtained by SenticNet were really interesting since it was the best-performing

Manuscript received January 5, 2019
Manuscript revised January 20, 2019

configuration on STS and the worst performing one on SemEval data. The Rahul et al.[3] a comprehensive overview of the methodology based on lexicons, where a comprehensive survey of studies based on this methodology was presented, in addition to proposing many improvements and presented challenges, limitations and applications for the use of Sentiment Analysis. In Al-Ayyoub et al. [4] developed a lexicon of 120,000 Arabic words to handle tweets written in Arabic, whereby the lexicon-based approach was adopted because of the challenges faced by tweets written in Arabic, such as the complexity of the language and the limited number of research and data for this purpose, which gave very good results. Pollyanna et al. in their research [5] provided a comparison of eight methods of analysis:
SentiWordNet, SASA, PANAS-t, Emoticons, SentiStrength, LIWC, SenticNet, and Happiness Index. The comparison was based on two criteria: agreement and coverage using two data sets: Near-complete Twitter logs and labeled Web 2.0 data, manually tagged. It has developed a new method integrates the seven lexicons except LIWC because of the presence of ownership restrictions, which gave good results. The study focused on polarity (positive and negative) content discovery and launched the iFeel web application to conveniently compare the results of various tools. The study [6] selected four general purpose lexicons based on various methodologies of development: VADER, MPQA, AFINN and SentiStrength. A benchmark dataset [7] they used and also they used the Group-2 and Group-3 portions of the dataset containing 1,600 and 4,000 issue comments extracted from JIRA issue tracking system. The issue comments are manually annotated with six basic emotions (i.e., joy, love anger, sadness, fear, and surprise). To assess the performance of each lexicon in terms of accuracy in detecting the positive, negative and neutral emotions. The results showed that the four lexicons were closely related, while SentiStrength excelled In terms of overall accuracy. AFINN have a better performance in detecting the positive and neutral emotions and the worst performance in the negative. While MPQA was the worst. Jeremy and Tawunra [8] in their study compared the performance of the SentiStrength and SentiWordNet lexicons using the data of the two web forums: Montada and Qawem.
The results showed close accuracy in the classification of positive and negative emotions. It also showed that the Qawem Forum has the highest percentage of negative sentences from the second forum.
Finally, researchers in [9] compared the Brazilian Portuguese LIWC lexicon for Sentiment Analysis with two other Portuguese LIWC lexicon: SentiLex and Opinion Lexicon. Where evaluation was made according to: Accuracy of classification, LIWC Lexicon had a better in positive texts (F-measure = 70.37%) and while SentiLex had a better performance in negative texts (F-measure = 60.25%). And the relationship between each lexicon in terms of the number of income elements of the same type of polarity.
the previous studies shown that the accuracy of the positive classification differs from the negative and neutral accuracy of the same lexicon, and therefore the results cannot be generalized due to the difference in the methodology used in applying these lexicons. In addition to the type of data used and the difference in its composition, the method of handling and processing the data set, as well as various comparison criteria. This paper will evaluation the positive, negative and neutral rating accuracy using untreated Twitter Tweets.

3. Sentiment Lexicons

A lexicon is a set of features such as words and emotions classification for each word. This method of sentiment analysis is based on comparing the words used within the text with one of the Previously prepared lexicons.

3.1 VADER

VADER (Valence Aware Lexicon and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Developed by Hutto and Gilbert [10] to solve the problem of analyzing language, symbols, and style of texts in the field of social media.
Capable of detecting polarity (positive, neutral, negative) in addition to the intensity of emotions in the text. The authors have published the lexicon in Python code as open source. It was built by examining and selecting features from three preset lexicons: Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW) and General Inquirer (GI). In addition to common abbreviations in social media such as WTF and acronyms (LOL), slang and facial expressions such as :-), which denotes a smiling face and indicates a positive emotion. 7,500 features have been tested to be in the lexicon. The following Python script formula shows the polarity classification of a phrase using VADER:

```
#from vaderSentiment import SentimentIntensityAnalyzer
sentences=
["The book was good" #positive example
,"VADER is not smart nor funny." #negative example
,"Today only kinda sux! But I'll get by, lol" #mixed example ]
analyzer = SentimentIntensityAnalyzer()
for sentence in sentences:
scores = analyzer.polarity_scores(sentence)
print(sentence + str(scores))
```

Table 1 shows the result after applying the previous code

Table 1: VEDER Classification

Sentence	***pos***	***compound***	***neu***	***Neg***
The book was good	0.492	0.4404	0.508	0.0
VADER is not smart nor funny.	0.0	-0.7424	0.354	0.646
Today only kinda sux! But I'll get by, lol	0.317	0.5249	0.556	0.127

The compound score was calculated by adding the equivalent values for each word in the lexicon and adjusted according to the rules to be between -1 (most negative) and +1 (most positive).
The pos, neg and neu values give the percentage of text occurring in each row. Their also set a standard threshold for classifying sentences, either negative, neutral or positive, as follows:
1- Positive: compound score> = 0.0.5
2- Neutral: (compound> -0.05) and (compound <0.05)
3- Negative: compound score <= - 0.05

3.2 SentiWordNet

SentiWordNet is a widely used tools in opinion mining, based on a WordNet Lexicon. This lexical dictionary groups called synsets from synonym, adjectives, nouns, verbs and other grammatical classes. SentiWordNet blends three scores with a WordNet dictionary synset to indicate the text's sentiment: positive, negative and objective (neutral). The scores, which are in the values of [0, 1] and add up to 1, are obtained using a semi-supervised machine learning method [11]. Data Format method within the lexicon: a pair (ID, POS) are uniquely defined. An expression from the third version of the WordNet lexicon, the PosScore and NegScore values mean the degree of positive and negative assigned to each expression as follows:
PosScore [0,1] : positivity scale
NegScore [0,1]: Negative Scale
ObjScore [0,1]: Objective scale.
Where the degree of objectivity can be calculated as follows:
ObjScore = 1 - (PosScore + NegScore).

we used SentiWordNet version 3.0, which is available at http://sentiwordnet.isti.cnr.it/. To assign polarity based on this method, we considered the average scores of all associated synsets of a given text and consider it to be positive, if the average score of the positive affect is greater than that of the negative affect. Scores from objective sentiment were not used in determining polarity[10].

We used the lexicon by writing code in a language Python and take advantage of library NLTK.
The following example shows the positive and negative degrees of some expressions within the lexicon:
<very.r.01: PosScore=0.25 NegScore=0.25>
<nice.a.01: PosScore=0.875 NegScore=0.0>
<love.v.02: PosScore=1.0 NegScore=0.0>
<worse.a.01: PosScore=0.0 NegScore=0.75>
<bad.a.01: PosScore=0.0 NegScore=0.625>
<truly.r.01: PosScore=0.625 NegScore=0.0>

3.3 SentiStrength

SentiStrengh is a program that is supported by Softpedia to analyze the emotions, free academic research.
It was produced as part of the CyberEmotions project, with the support of the European Union FP7. It analyzes the sentiments of more than 16,000 texts on the social network in a second, which has an accuracy of the level of human accuracy of texts in English, with the exception of political texts. Expressions and terms used in the lexicon is derived from the lexicon LIWC [12]. The program appreciates the power of positive and negative emotions in short texts according to the following:
-1 (not negative) to -5 (extremely negative)
1 (not positive) to 5 (extremely positive)
To classify the texts using SentiStrength, we placed the texts in a plain text file, one text per line, so the output was a copy of the file containing positive and negative classifications at the end of each line. Table 2 shows the results of program implementation on three lines of a text file, explaining the positive and negative value of each text:

Table 2: SentiStrength Classification Example

Text	***Positive***	***Negative***
using Linux and loving it - so much nicer than windows... Looking forward to using the wysiwyg latex editor!	4	-1
@ruby_gem My primary debit card is Visa Electron.	1	-1
@kirstiealley I hate going to the dentist.. !!!	1	-4

3.4 Liu and Hu Lexicon

The lexicon is composed of two lists of words, the first with a positive classification and the second with a negative classification, prepared by researchers from the Department of Computer Science University of Illinois at Chicago, Hu and Liu [13]. There are words that are very common spelling mistakes in social media content, included in the lexicon. The following code in Python shows how to access negative and positive lexicon words:

```
>>> from nltk.corpus import opinion_lexicon
>>> opinion_lexicon.words()
['2-faced', '2-faces', 'abnormal', 'abolish', . . .]
>>> opinion_lexicon.negative()
['2-faced', '2-faces', 'abnormal', 'abolish', . . .]
>>> opinion_lexicon.positive()
['a+', 'abound', 'abounds', 'abundance', 'abundant', . . .]
```

To find a positive or negative emotions in the texts used, we have written a Python code that compares the words of the entered text with the verified lexicon words within the NLTK Natural Language Toolkit in the Opinion Lexicon.

3.5 AFINN-111

AFINN-111 is a list of English words that are rated and give a value between -5 (negative) and +5 (positive) [14]. Classified manually by Finn Arup Nielsen in 2000-9-2011. There are two versions of the list:

AFINN-111: The latest version containing 2477 words and phrases.

AFINN-96: 1468 unique words and phrases contain 1480 lines, some words are found twice, and are alphabetically unordered. The following in table 3 is part of the AFINN-111 list consisting of two columns , the first represents the word and the second represents the corresponding value, which falls within the field [-5, + 5].

Table 3. AFINN-111 List

Word	*Value*
Lose	-3
Loses	-3
Loser	-3
Losing	-3
Loss	-3
Lost	-3
Lovable	3
Love	3
Loved	3
Loving	2
Lowest	-1

The following Python code demonstrates how to calculate the degree of emotion of a positive statement of degree 3:

```
afinn= dict(map(lambda kv:(kv[1])),[line.split('\t')
for line in open("AFINN-111.txt")]))
score=sum(map(lambda word: afinn.get(word, 0), "
AFINN is very good".lower().split()))
print(score) // 3
```

4. Performance Evaluation

The performance evaluation methodology consists of the steps: selecting the appropriate data sets for the polarity classification task, then performing the classification task by applying the chosen lexicons. Then calculating the accuracy of lexicons-based works.

4.1 Twitter Data Set

Stanford and Sandars Test set were used to test sentiment analysis in Twitter, without performing the data pre-processing stage.

- **Stanford Twitter Sentiment Test Set:**

This collection was built by Go and et al. [15] using (Twitter Search API).Their searched for names that represented people, companies and products. Their methodology used to collect and classify training data on facial expressions, assuming that any tweet containing positive facial expressions such as :) is positive, and any tweet containing negative facial expressions is considered negative. In addition to a set of manually tagged test data consisting of 182 tweets with a positive tag, 177 negative; 139 neutral. The test data is formatted as a csv file with 6 fields:
Tweet polarity (negative = 0, neutral = 2, positive = 4), Tweet ID, Tweet date, recipient, User, Tweet text.

- **Sandars Twitter Dataset Test Set:**

Built by researcher Niek J. Sanders, it consists of 5513 tweets on four topics (Apple, Twitter, Microsoft, Google). It was manually tagged with four tags: positive, negative, normal and unrelated (Neglected).
The Test set is formatted as a csv file with 3 fields(Topic, Tag, Tweet ID). By applying the Python code that accompanies the test set, we get two additional fields for each tweet, (the tweet’s date ,tweet text). Python code implementation requires handling the Twitter API's [16]. By registering the application within Twitter to obtain two pair (secret key) for authorized access, this is known as (OAuth) the authentication method [17][18] .

4.2 Evaluation criterion

To evaluate classification results using previous lexicons, the following measures were focused on:

- $Accuracy = \frac{TP+TN}{P+N}$
- $Precision = \frac{TP}{TP+FT}$
- $Recall = \frac{TP}{P}$
- $F1 - Score = \frac{2 \times precision \times recall}{precision+recall}$

Where TP : the number of positive values classified properly,
TN the number of values negatively classified properly,
FP number of positive values classified incorrectly,
P is the total number of positive values, N is the total number of positive values

4.3 Results Discussion

When using the Stanford dataset, the results showed that the best performance in terms of accuracy was achieved by the VADER lexicon 72%, while the performance accuracy of the sentiStrength, AFINN-111 and Liu-Hu lexicon was close to each other, while the accuracy of the sentiwordnet lexicon fell to the value 53%.

By focusing on the F1-measure affected by FP and FN, the VADER lexicon has excelled In the positive and negative classification, while we note that it has achieved close value with the rest of the lexicons in the neutral classification. AFINN-111 performed best in negative sample rating, while positive and negative rating sentiStrength was close to VADER.

Table 3 illustrates evaluation criteria using the Stanford dataset and Figure 1, summarizes the accuracy of the performance with the Stanford dataset.

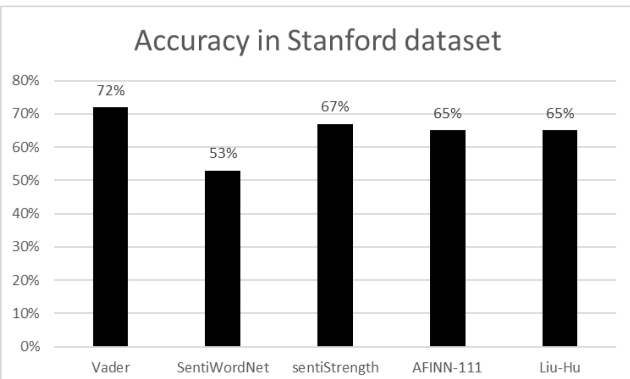

Fig 1 Accuracy of performance in Stanford dataset

When using the Sandars dataset, it is noted that the accuracy of the Vader and AFINN-111 lexicons is close and higher than the accuracy of the remaining three lexicons. It is noted that VADER gives the best result in the positive and negative classification. While outperform both Sentiwordnet and AFINN-111 in the neutral l classification. It is noted that Liu-Hu's performance is low, while the results show SentiStrength's performance classification is better in both negative and neutral more than positive, as shown in Table 4.

The Figure 2 illustrates the accuracy of performance with the Sandars dataset:

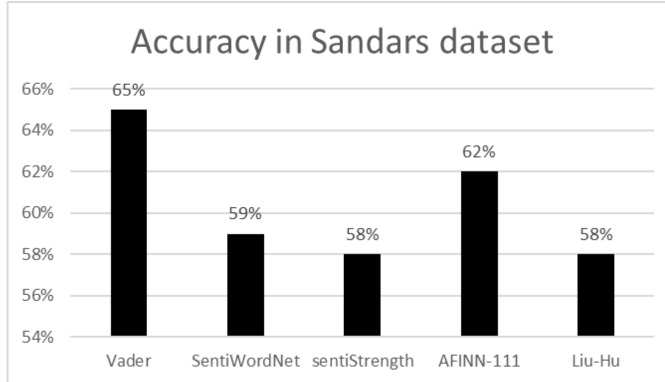

Fig. 2 Accuracy of performance in Sandars dataset

Finally, it becomes clear that VADER's lexicon performance was most accurate using two datasets, while the rest of the lexicons achieved close results. While Li-Hu's lexicon performance was better when using Stanford dataset than with Sandars dataset. The lexicons were used by default and without any modifications, in addition to dealing with the two data sets without going through a pre-processing stage.

5. Conclusion and Future Work

We note that the results of the investigation of the applied sentiment lexicons are good, although the two sets of data used did not undergo pre- process, that is, the original tweets were dealt with.

After the previous comparison, we find that lexicon VADER has a good possibility for the classification of short texts pre-process, positively and negatively and neutral, where it can deal with all cases of text.

The results showed the ability of the lexicon SentiStrength negative classifying more than positive in both data sets. AFINN-111 performed best in classifying negative samples. While the classification accuracy by using the lexicon SentiWordNet less than the rest of the lexicons, despite strong structure in the lexicon. As for Li-Hu lexicon, it achieved different results using two datasets, the performance declined with the use of the Sandars dataset.

In future works, other comparable lexicons can be added, such as Opinion Finder and GPOMS, as their can used in the most important research, especially related to studying Twitter sentiments and linking it to financial stocks.

Other data sets can also be used, in addition to studying the effect of data pre-processing on the accuracy of implementation.

Table 3: Evaluation criteria in the Stanford dataset

Sentiment Lexicon	*accuracy*	*Positive*			*Negative*			*Neutral*		
		P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Vader	72%	0.67	0.81	0.73	0.84	0.64	0.72	0.67	0.7	0.68
SentiWordNet	53%	0.65	0.54	0.59	0.68	0.31	0.42	0.42	0.79	0.54
sentiStrength	67%	0.66	0.76	0.71	0.78	0.63	0.7	0.59	0.62	0.61
AFINN-111	65%	0.72	0.68	0.7	0.89	0.49	0.63	0.49	0.81	0.61
Liu-Hu	65%	0.72	0.66	0.69	0.85	0.55	0.67	0.5	0.78	0.61
P:precision R:recall F:F1-measure										

Table 4: Evaluation criteria in the Sandars dataset

Sentiment Lexicon	*accuracy*	*Positive*			*Negative*			*Neutral*		
		P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Vader	65%	0.49	0.72	0.59	0.67	0.53	0.59	0.84	0.53	0.65
SentiWordNet	59%	0.43	0.43	0.45	0.46	0.42	0.47	0.74	0.71	0.73
sentiStrength	58%	0.29	0.63	0.4	0.45	0.49	0.47	0.82	0.59	0.68
AFINN-111	62%	0.3	0.53	0.38	0.51	0.41	0.45	0.78	0.68	0.73
Liu-Hu	58%	0.3	0.59	0.4	0.44	0.49	0.46	0.79	0.6	0.68
P:precision R:recall F:F1-measure										

References

[1] M Al-Ayyoub, SB Essa, I Alsmadi,"Lexicon-based sentiment analysis of Arabic tweets." IJSNM, vol. 2, no. 2, pp.101-114, 2015.

[2] S Baccianella, A Esuli, F Sebastiani,"Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In Lrec., pp.2200-2204, 2010.

[3] R Rajput, AK Solanki, "Review of Sentimental Analysis methods using lexicon based approach," IJCSMC , vol. 5, no. 2, pp. 159-166, 2016.

[4] M Al-Ayyoub, SB Essa, I Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets," IJSNM, vol. 2, no. 2, pp. 101-114, 2015.

[5] P Gonçalves, M Araújo, F Benevenuto and C. Meeyoung, "Comparing and combining sentiment analysis methods.," in In Proceedings of the first ACM conference on Online social networks, ACM, 2013.

[6] MR Islam, MF Zibran, "A comparison of dictionary building methods for sentiment analysis in software engineering text," in ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), IEE, 2017.

[7] M Ortu, A Murgia, G Destefanis, P Tourani, R Tonelli, M Marchesi, and B Adams, "The emotional side of software developers in JIRA.," in In Proceedings of the 13th International Conference on Mining Software Repositor, ACM, 2016.

[8] T Chalothorn, J Ellman, "Sentiment analysis of web forums: Comparison between sentiwordnet and sentistrength.," in The 4th International Conference on Computer Technology and Development (ICCTD 2012), 2012.

[9] PP Balage Filho, TAS Pardo, SM Aluísio, "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis.," in In Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013.

[10] CJ Hutto, E Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text.," in In Eighth international AAAI conference on weblogs and social media., 2014.

[11] S Baccianella, A Esuli, F Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in In Lrec, 2010.

[12] D Iosub, D Laniado, C Castillo, MF Morell and A Kaltenbrunner, "Emotions under discussion: Gender, status and communication in online collaboration," PloS, e104880., vol. 9, no. 8, 2014.

[13] M Hu, B Liu, "Mining and summarizing customer reviews," in In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004.

[14] FÅ Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," arXiv preprint arXiv:1103.2903, 2011.

[15] A Go, R Bhayani, L Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford 1, no. 12, 2009.

[16] J Bollen, H Mao, X Zeng, "Twitter mood predicts the stock market," Journal of computational science , vol. 2, no. 1, pp. 1-8, 2011.

[17] W Xu, T Li, B Jiang, C Cheng, "Web Mining For Financial Market Prediction Based On Online Sentiments.," in In PACIS,, 2012.

[18] A Karim, MA Adnan, "An OpenID Based Authentication Service Mechanisms for Internet of Things." In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 687-692. IEEE, 2019.

Dr. Mohammed Al-Shabi received his bachelor's degree (B.Sc. Computer Science) from Technology University at Iraq (1997), Postgraduate Master (M. Sc. Computer Science from Putra Malaysia University at 2002), and Ph.D. (Computer Science) from Putra Malaysia University, Malaysia (2006). He is currently an associate professor in the Department of Management Information System, College of Business Administration at Taibah University, Kingdom of Saudi Arabia. Prior to joining Taibah University, he worked in the Faculty of a computer at Qassim University, Saudi Arabia. His research interests include wireless security, cryptography, UML, Stenography Multistage interconnection network, Vehicular Ad-hoc Network-Cloud, Smart and Intelligent computing and Apply Mathematical.