Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

Natural Language Processing: Text Categorization and Classifications

Mona Nasr, Andrew karam, Mina Atef, Kirollos Boles, Kirollos Samir, Mario Raouf
Department of Computer science,Helwan University, Egypt, Cairo
m.nasr@helwan.edu.eg,karamandrew6@gmail.com, minaatef@fci.helwan.edu.eg,kirloesboles98@gmail.com, kirosidhom@gmail.com, raoufmario3@gmail.com

ABSTRACT

There are huge data from unstructured text obtained daily from various resources like emails, tweets, social media posts, customer comments, reviews, and reports in many different fields, etc. Unstructured text data can be analyzed to obtain useful information that will be used according to the purpose of the analysis also the domain that the data was obtained from it. Because of the huge amount of the data the human manually analysis of these texts is not possible, so we have to automatic analysis. The topic analysis is the Natural Language Processing (NLP) technology that organizes and understands large collections of text data, by identifying the topics, finding patterns and semantic. There two common approaches for topic analysis, topic modeling, and topic classification each approach has different algorithms to apply that will be discussed.

Keywords - **Natural Language Processing; Topic Classification; Topic Modeling; Text Categorization**

Date of Submission: June 18, 2020 Date of Acceptance: Aug 31, 2020

I. INTRODUCTION

Daily a huge amount of unstructured text data is obtained from various resources such as social media posts, tweets, and articles. These unstructured texts can be analyzed to extract information about trending topics, products, events, and reactions, etc. This analyzed information is very important and useful, for example allowing businesses to improve their strategies and the decision-making process. Another resource of daily unstructured texts is Emails that can be analyzed to prevent spam. There are many different resources as mentioned social media posts, tweets, articles, emails also customer feedback, reviews, and reports, etc. There are many examples and applications of analyzing the data that will be useful according to the data's domain and the goal of the analysis.

In the rest of the paper, Section II defines the related work which worked on this topic. Section III defines some models that work on topic modeling and topic evaluation. In section 4 define the conclusion.

II. RELATED WORK

1. [1] surveyed a suite of algorithms for managing the document. The used model is topic modeling, for finding the theme `pattern`, after that examine the documents related to this pattern. They proved that topic modeling can handle and apply to massive of data and can adapt with many kinds of data. They have used for finding the theme in social networks, images, and genetic data. They used Latent Dirichlet Allocation `LDA`, and use ``Seeking Life's Bare "Genetic" Necessities `` for determining the number of genes an organism. They highlighted the important words with different type to use it as shown in Fig my label1. They worked by defining a topic as label of words, or on the another meaning makes a distribution over a vocabulary, for example if A topic has words about as with high probability as shown in Fig 1, Fig 2. They worked with this step: first choose distribution over the topics randomly, then in the document, for each word choose randomly a topic from the distribution over topics, then choose randomly a word from the distribution over vocabulary.

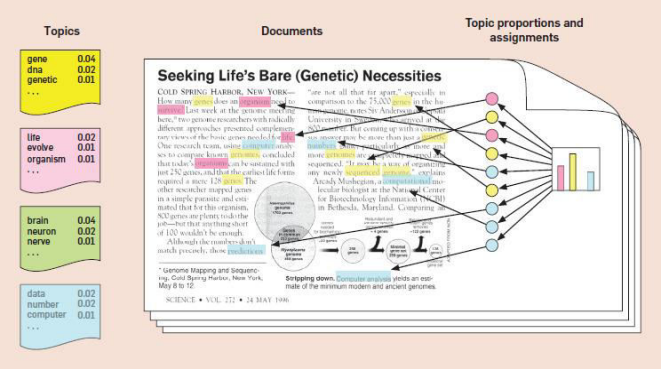

Figure 1. How every topic contains words with high probability, with highlight

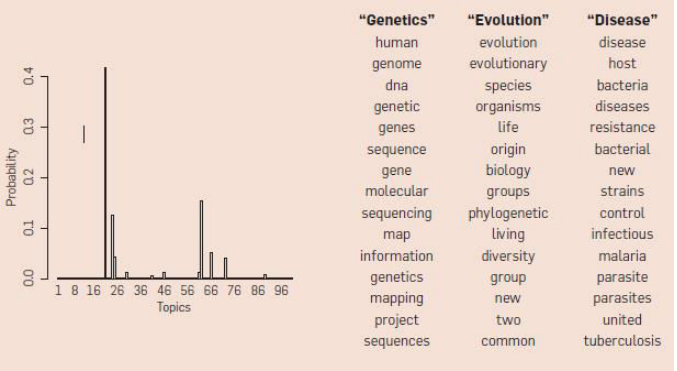

Figure 2 The figure shows every topic distribution

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

2. [2] Make the survey about the topic modeling in the text mining. These models are Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Correlated Topic Model (CTM). And survey topic evolution model, which are dealing with time factor. The topic evolution model is Topic over Time (TOT), Dynamic Topic Models (DTM), Multiscale Topic Tomography, Dynamic Topic Correlation Detection, and Detecting Topic Evolution in scientific literatures and so on. As they defined that the topic is a probability distribution over the words. Which that the topic model is a generative model where worked on the documents. They created a new document by choosing a distribution of words over the topics. Then each word on this document will assign to the topic or the on the other meaning will assign to the best distribution. Then, draw a word from that topic (distribution).

3. [3]Built model to classify text specially News text, because the number of articles on internet is in permanent increasing. That model classifies news based on topic modeling using Latent Dirichlet Allocation (LDA) method (we explain it in later section). The model change text from space vector model to topic vector firstly by LDA, and then use the output from LDA as input train_x of softmax regression and its label as train_y. Then it can get good accuracy of classification. The model is worked as shownin Fig 3To evaluate this model they use 3 ways the precision, recall and F1-Measure. Precision is the detection percentage of related documents and all documents. Recall is the ratio between the number of correct predictions for a category and the number of real documents in the same category. F1-Measure is the key index of the experimental result and harmonic mean of previous methods as shown in Fig 4The data consists from 20 newsgroup data exists in sklearn package and available on kaggle written in the last few years has 20 different topics like: religion, graphics, computer hardware they test on three topics each topic has 1000 sample and the result showed in Fig 5

4. [4]Utilized medical data (reports) that are stored as electronic health records (EHRs) for prospective patients, so that can provide better clinical decision-making. This study provides approach that firstly preprocessing the data: removing all un-useful features like Medical record numbers,frequent words and stop words. Secondly, they applied LDA topic modeling algorithm so that they can determine the topics of clinical reports, LDA was applied using Stanford Topic Modeling Toolbox (TMT). Thirdly using the topics distributions for each topic that are produced from topic modeling to represent them as topic vectors instead of Bag of Words that is less compact. Finally run three types of classifications: Supervised

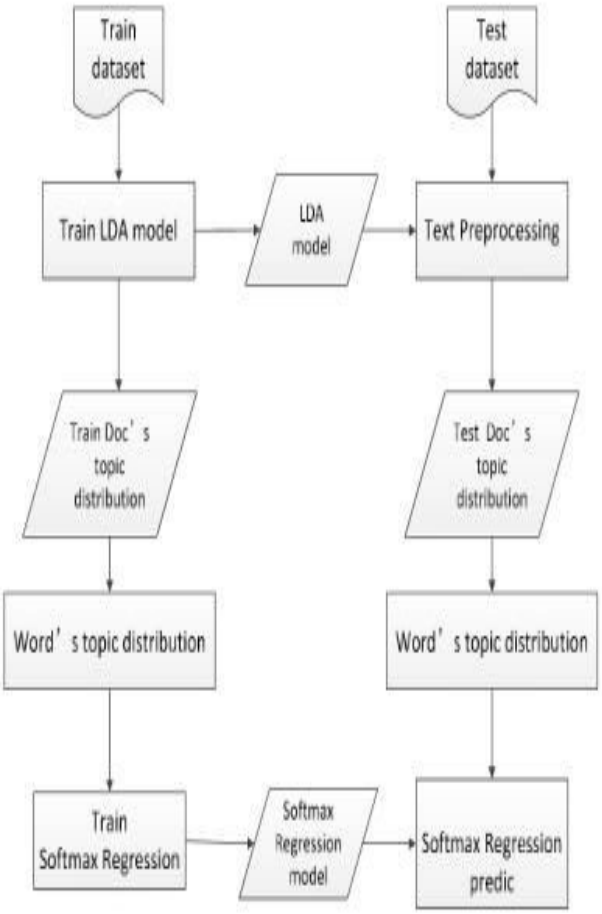

Figure 3The model

$$recall = \frac{\sum_{u_i \in U} TP(c_i)}{\sum_{u_i \in U} T(c_i)} \quad (4)$$

$$precision = \frac{\sum_{u_i \in U} TP(c_i)}{\sum_{u_i \in U} L(c_i)} \quad (5)$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision} = \frac{2\sum_{u_i \in U} TP(c_i)}{\sum_{u_i \in U} R(c_i) + \sum_{u_i \in U} T(c_i)} \quad (6)$$

The paper makes $TP(c_i)$ to represent the number of correct prediction of class i . The $L(c_i)$ represents the return number of documents of class i by classification model and $T(c_i)$ represents the real document number of class i .

Figure 4Equation

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

All classes in All Dataset		
Macro-F1-Measure: 0.84		
Class 1	**Class 2**	**Class 3**
Recall: 0.96	Recall: 0.79	Recall: 0.81
Precision: 0.95	Precision: 0.78	Precision: 0.82
F1-easure:0.95	F1-easure:0.78	F1-Measure:0.81

Figure 5. The result of experiment

Classification using Machine learning model SVM ,Aggregate Topic Classifier(ATC) depending on composing between representative topic vector for each class with averaging their corresponding topic distributions in dataset , Binary Topic Classification(BTC) such that two topics can analyzed as unsupervised classification after get high probability topic assigned as predicted class for each document and note which topic corresponds to which class was found by checking predicted class proportions. Fig 6

5. With the increasing growth of the amount of data(in our case the text data), [5]collected and stored it becomes more and more harder to process that it now almost impossible to process the data manually.so the need to an automated text classifier is increasing. And the solution for this problem is to develop a Topic-classification algorithm that can help classifying the topics of a given document efficiently in the manner of accuracy and time needed to classify. But there are many challenges to accomplish this one of these challenges is the required computation time is very long that

Algorithm	Precision	Recall	F-score
Baseline	76.6	87.5	81.7
BTC	88.6	73.4	77.7
ATC	96.1	96.0	96.1
Topic vectors	96.6	96.7	96.6
Raw Text	96.4	96.3	96.3

Figure 6 Overall classification performance

makes fitting topic algorithm hard thing to do. But the benefits of developing the topic classifier outcomes the challenges to develop it because there is more application for topic models, for example text-recommendation systems, spam filtering, computational biology analyses, and many more applications. And previously there was many approaches to the topic model algorithm, some of these approaches is Probabilistic Latent Semantic Analysis (PLSA) and the current widely used approach Latent Dirichlet Allocation (LDA). Both approaches based on the word's probability distribution in the document, and that every document is a mixture of topics. The proposed approaches denoted as Topic Mapping consists of four steps where the first two steps have the single purpose to denoise the word network.

- Preprocessing: using a stemming algorithm to return the words to its stem so that words and its plural is not considered a distinct word and the same for different tenses of a verb. And removing standard list called "stop words" that does not provide useful topic information.
- Pruning of connections: calculating the similarity of every pair of words that appear together in a document or more with a null model to check if the two words appearance are connected to each other's appearance.
- Clustering of words : assuming that topics in a document will belong to a community or more of words using an algorithm for community detection like Info map algorithm to determine the number of communities in the document unsupervised then the user won't have to guess the number of topics in a document the community detection will do that .
- Topic-model estimation: using locally optimized PLSA-like likelihood to get the final estimate of model probability.

III. METHODOLOGY

In this section we defined some of methods can use on topic modeling and topic classification. This illustrate is inspired from [2]

1. Topic Modeling

1.1. Latent Semantic Analysis (LSA)[6]

The goal of LSA is creating thing can from it compute the similarity between the tests and catch the highest related words. It makes vector-based representation represent the text. The LSA is working with these steps:

- Get a set of text, then divide it by documents.
- Make concurrence matrix for documents and terms, With mention the document, terms and dimensional value for terms and dimensional vector for the documents with this symbols x,y,m,and n.
- Calculated each cell.
- Compute all the diminutions by using SVD model.

LSA uses singular value decomposition (SVD). SVD is a method make reconfigure and calculate the diminutions of vector space by using a matrix. The goal of LSA is finding the meaning of the text.

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

1.2. Probabilistic Latent Semantic Analysis (PLSA)[7]

PLSA is appeared to fix some disadvantages is founded on LSA, by using generative model.

PLSA is a method that can process indexing of documents for counting factor data analysis based on a statistical model `aspect model`. The model of PLSA is shown in Fig **7**The goal of using PLSA is identifying and distinguishing between the context of the words. Without use dictionary. Characteristics of PLSA: PLSA can handle the polysemy of words.

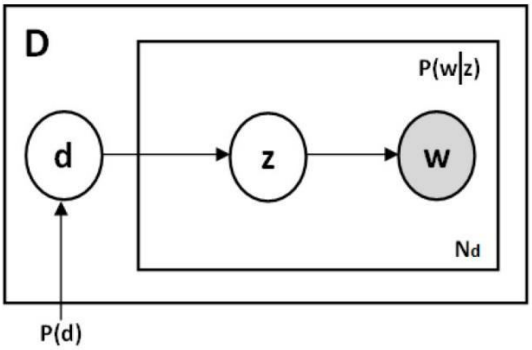

Figure 7The model of PLSA

1.3. Latent Dirichlet Allocation (LDA)[1]

With growing number and the content of articles, blogs and literature has posed a challenge to the data mining researchers specifically, whose worked text analysis to get a new meaningful from the data and visualize it, so the needed of LDA coming. LDA is coming to improve the way of that the oldest model LSA and PLSA work, by mixture the previous model. LDA is a generative model based on statistical (Bayesian) topic models. It works by try to mimic the writing process, it try to generate a document on the topic already given. LDA has different models like, supervised topic models, latent Dirichlet co-clustering, temporal text mining, author- topic analysis, and LDA based bioinformatics. The LDA work with these steps:

- Every document is been a mixture of topics, which that each topic is probability distribution (discrete). Each distribution determines the word is more approximate to the closer topic.
- Determine that the document is bag-of-words (BOW) with no structure
- LDA consider `D` documents as BOW over `K` latent topics, each of this topic is the distribution of word `w`. Fig 8 illustrate more about how LDA model is working

LDA has some negative characteristics:

- LDA cannot find a good representation for the relationship between the topics.
- The programmer should manually remove the stopwords.

1.4. Correlated Topic Model (CTM)[1]

CTM is a statistical model. The goal of CTM is discovering the topics in a group of documents. CTM using logistic distribution, and the CTM is depending on LDA.

Characteristics of CTM:

- Using the logistic for make relation between the topics
- Allows occurrences on the other topics

2. Topic Classification

Classification Machine learning models that are used for topic classification.

Training data should be transformed to vectors so be able to extract feature. That features should be tagged with labels then we can use some models such as:

2.1. Naive Bayes[8]

Simple algorithm based on Bayes' Theorem. Naive Bayes correlates the probability of words with the probability of that text given specific topic.

$c = \text{argmax}P(x1,x2,...,xn|c)P(c)$

- c is the class
- x is representation of document

2.2. Support Vector Machines

SVM separates these vectors into the given classes (topics). Then for classification new text SVM victories it and determine the side of the vector. Shown in Fig 8 inand Fig 9

2.3. Deep Learning

The main networks are used Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

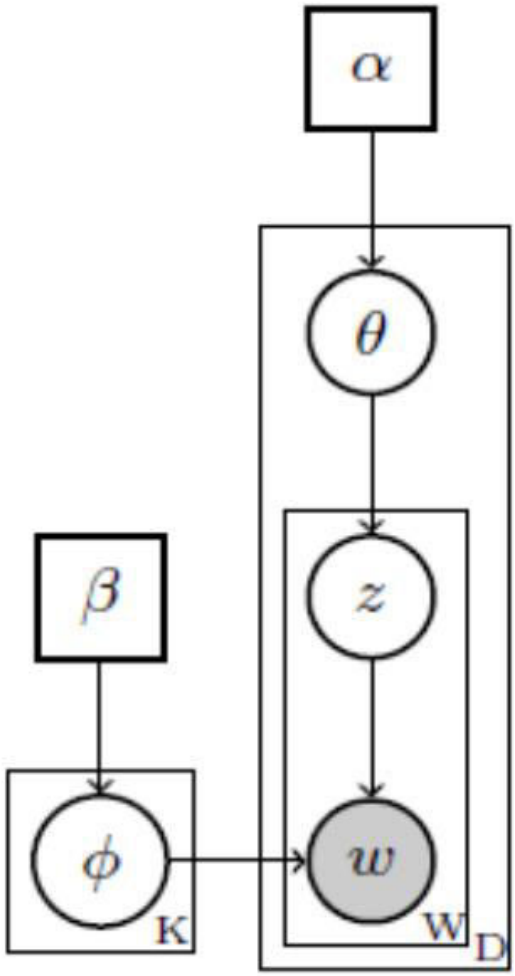

Figure 8 Graphical representation of LDA model

3. Topic Evaluation

With add the time as the factor with text.

3.1. Topic Over Time (TOT)[9]

The TOT model topics and changes are made over time by considering both word repetition pattern and time. We can say that the TOT is a continuous distribution, because of working over the time.

TOT model generates this distribution:

- Multinomial distribution over topics, this makes by sampling the topic from the dirichelt. And the words are generated from the multinomial distribution.
- Beta distribution of each topic, this distribution generates time stamp of the document's
- Also, create a narrow time distribution topic, and if the pattern of a strong word repetition, it will create a broad time distribution.

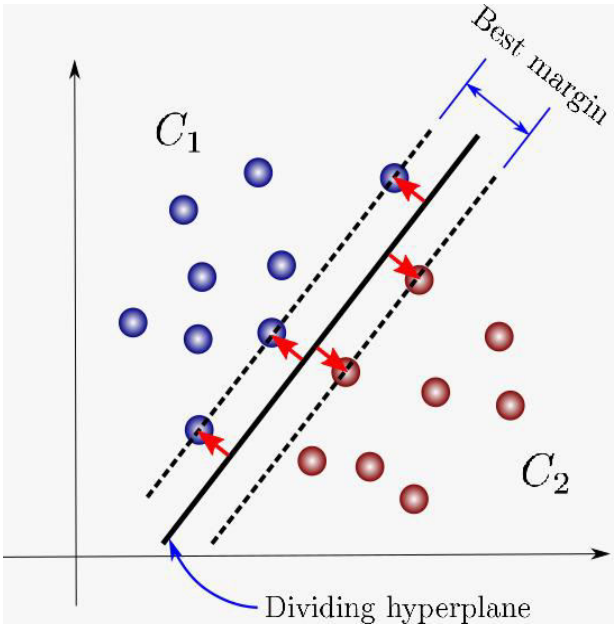

Figure 9simple SVM model

3.2. Dynamic Topic Models (DTM)

The goal of DTM is guessing the topic distribution at different epochs. The DTM is using Gaussian instead of Dirichelt. Also, it can capture the topic over the time.

IV. APPLICATION

In this section we described some of application based on the previous models.

1. Topic Modeling

1.1 LSA

LSA able to model human conceptual language. LSA can work with large dataset, but for more efficient work with small size of dataset Technical dataset and IR [10]

- predictor of query-document topic similarity judgments
- simulation of agreed upon word-word relations and of human vocabulary test synonym judgment
- simulation of human choices on subject-matter multiple choice tests
- predictor of text coherence and resulting comprehension
- simulation of word-word and passage-word relations found in lexical priming experiments
- predictor of subjective ratings of text properties, i.e. grades assigned to essays
- predictor of subjective ratings of text properties, i.e. grades assigned to essays
- Document clustering in text analysis
- Recommender systems

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

- For small group of text, get the best similarity
- Building user profiles.
- Stemming

1.2 PLSA

- Fisher Kernels "Discriminative setting"
- Information retrieval and filtering
- Image Classification.

1.3 LDA

LDA is widely used in many fields[11]

- linguistic science
- political science
- medical and biomedical
- geographical and locations
- Software engineering and topic modeling
- social network and microblogs
  Crime prediction/evaluation
- Role Discovery: Social Network Analysis (SNA).
- Emotion Topic.
- Automatic essay grading
- Anti-Phishing `email`

1.4 CTM

- JSTOR archive
- automatic recommendation systems

2. Topic Evaluation

2.1 TOT

- Topic evaluations with add the time

2.2 DTM

- Biological APP, which contains text.

V. CONCLUSION

DAILY a huge amount of unstructured text data are obtained from various resources such as social media posts, tweets, articles, emails also customer feedback, reviews, and reports, etc.

This unstructured text data can be analyzed o extract meaningful and useful information according to the data's domain and the goal of the analysis. The topic analysis is the Natural language processing (NLP) technology that organizes and understands large collections of text data, by identifying the topics, finding patterns, and semantic. There are two common approaches to topic analysis (topic modeling and topic classification). Topic classification is classifying the text according to labels already predefined than can predict new text with the predefined label. SVM an example algorithm to apply Topic classification as it is a supervised classification problem.

Topic modeling is classifying the text according to bag-of-words (BOW) extract from the text or the dictionary already defined before by the programmer. Topic modeling is working by learning the patterns of words and learning how this word is connected to learn the patterns, then generate the same pattern to use in a new topic. LSA, LDA, and PLSA are algorithms to apply topic modeling as it clustering problem.

VI. REFERENCES

[1] D. M. Blei, "Probabilistic Topic Models," *Commun. ACM,* p. 8, 2012.

[2] R. A. a. K. Alfalqi, "A Survey of Topic Modeling in Text Mining," *International Journal of Advanced Computer Science and Applications,* p. 6, 2015.

[3] Z. L. a. W. S. a. M. Yan, "News text classification model based on topic model," *IEEE/ACIS,* p. 5, 2016.

[4] E. a. C. H.-A. Sarioglu, "Topic Modeling Based Classification of Clinical Reports," in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, Bulgaria, 2013.

[5] A. a. S. M. a. W. J. a. A. D. a. K. K. a. A. L. Lancichinetti, "High-Reproducibility and High-Accuracy Method for Automated Topic Classification," *Physical Review X,* p. 5, 2015.

[6] P. Foltz, "Latent Semantic Analysis for Text-Based Research," *Behavior Research Methods,* pp. 197-202, 1996.

[7] D. Tian, "Research on PLSA model based semantic image analysis: A systematic review," *Journal of Information Hiding and Multimedia Signal Processing,* pp. 1099-1113, 2018.

[8] P. a. D. S. Kaviani, "Short Survey on Naive Bayes Algorithm," *International Journal of Advance Research in Computer Science and Management,* 2017.

[9] X. a. M. A. Wang, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[10] T. K. a. D. S. Landauer, "Latent semantic analysis," *Scholarpedia,* p. 4356, 2008.

[11] H. a. W. Y. a. Y. C. a. F. X. a. J. X. a. L. Y. a. Z. L. Jelodar, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications,* p. 11, 2019.

Volume: 12 Issue: 02 Pages: 4542-4548 (2020) ISSN: 0975-0290

[12] D. a. N. A. a. J. M. Blei, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research,* pp. 601-688, 2001.

[13] D. a. L. J. Blei, "A correlated topic model of Science," *The Annals of Applied Statistics,* 2007.

[14] D. a. L. J. Blei, "Dynamic Topic Models," in *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[15] Mona Nasr, Omar Farouk, Ahmed Mohamedeen, Ali Elrafie, Marwan Bedeir, Ali Khaled, Benchmarking Meta-heuristic Optimization, *International Journal of Advanced Networking and Applications (IJANA), Volume 11 Issue 6 Pages: 4451-4457 (2020).*

[16] Farrag, M., Nasr, M., A Proposed Algorithm to Detect the Largest Community Based on Depth Level, *International Journal of Advanced Networking and Applications (IJANA), Volume 09, Issue 02, Sep - Oct 2017 issue, pp. 3362-3375.*

[17] Mona Nasr, Rana Osama, Rana Osama, Nouran Mosaad, Nourhan Ebrahim, Adriana Mounir, Realtime Multi-Person 2D Pose Estimation, *International Journal of Advanced Networking and Applications (IJANA), Volume 11 Issue 6 Pages: 4501-4508 (2020).*