



# Social Network User Profiling With Multilayer Semantic Modeling Using Ego Network

Tamil Selvi P., Michael Job College of Arts and Science for Women, India


 <https://orcid.org/0000-0002-3737-3025>

Kishore Balasubramaniam, Dr. Mahalingam College of Engineering and Technology, India

 <https://orcid.org/0000-0003-1918-9774>


Vidhya S., PKR Arts College for Women, India

Jayapandian N., Christ University, India

 <https://orcid.org/0000-0002-7054-0163>


Ramya K., PA College of Engineering and Technology, India

Poongodi M., Division of Informational and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

 <https://orcid.org/0000-0001-6801-6138>

Mounir Hamdi, Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Godwin Brown Tunze, Mbeya University of Science and Technology, Tanzania

 <https://orcid.org/0000-0002-7991-9962>

## ABSTRACT

Social and information networks undermine the real relationship between the individuals (ego) and the friends (alters) they are connected with on social media. The structure of individual network is highlighted by the ego network. Egocentric approach is popular due to its focus on individuals, groups, or communities. Size, structure, and composition directly impact the ego networks. Moreover, analysis includes strength of ego – alter ties degree and strength of ties. Degree gives the first overview of network. Social support in the network is explored with the “gap” between the degree and average strength. These outcomes firmly propose that, regardless of whether the approaches to convey and to keep up social connections are evolving because of the dispersion of online social networks, the way individuals sort out their social connections appears to remain unaltered. As online social networks evolve, they help in receiving more diverse information.

## KEYWORDS

Dimension Reduction, DoctoVec, Genism, Kmean, LSTMs, Matplotlib, Meanshift, PCA, Python, RNN, Scikitlib, Vectors, Visualizaton

DOI: 10.4018/IJITWE.304049

\*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

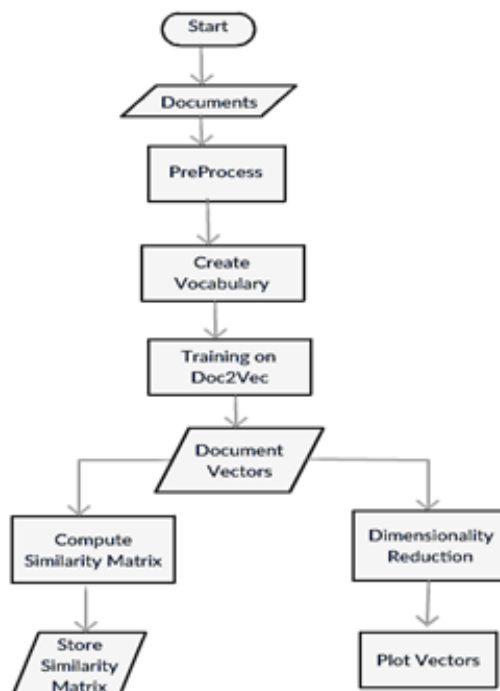
## INTRODUCTION

Nowadays, the way people express their opinions and views is highly changed by the internet Deng (2021); Arnaboldi et al. (2012). Online forums, Product reviews website, social media, blog post are the interactive platforms where the user inform, express, influence others Guidi et al. (2020); Rafi and Shaikh (2013). Social Media websites like Facebook, Instagram, Twitter and YouTube produce a huge amount of the data which can be in the form of posts, comments, images, tweets and videos respectively. This huge data involves different semantic dimensions for a given dataset resulting in multiple views, contradicting opinions, rational opinion and manipulating views. So, to manage this vagueness of the opinions of people from different backgrounds on the same topic. Hence, the core for this research paper is to come up with the powerful searching and clustering concept Gunaratna (2016); Siddiqui and Islam (2019) which can deal with the different background of people and estimate the similarity in their profiles. To predict their similarity in opinions, views and interests that is how similar their personality. This can be done by content matching of their profile instead of keyword match approach Hsu et al. (2020); Stolz and Schlereth (2021).

## DESIGN OF SYSTEM

Given the dataset of the documents of people from various background taken shown in further sections with figures. The first step is data preprocessing followed by creating the vocabulary. The dataset is trained on doc2vec. The numeric representation of doc is challenging task. Doc2vec helps to represent word to vectors using the module of skip gram and Bag-Of-Words model. Reduction of the vector formed from this step is done called dimensionality reduction for forming similarity matrix. The similarity matrix reveals how similar the profiles are.

Figure 1. Architecture-proposed model



The Manuscript is organized as Section II gives the Literature work related to the ego networks in social network analysis, user profiling. Section III, gives the research proposed work, which gives the methodology, which includes the Data preprocessing and cleaning process, Model selection and Training process, Section IV, describes about the Experimental setup and Visualization of results and observation. Section V, gives the conclusion and Future works.

## LITERATURE SURVEY

Weak Ties Theory - In weak ties theory, strong ties that can be used for any person. Whereas in network may some weak ties and need more information to be fully communicated with that node. This theory says that the more weak, the more connected to the world and are more likely to receive important information about ideas, threats and opportunities in time to respond to them. Weak ties are tend to have to be dynamic and innovative. Strong ties are more disjointed and chaotic than weak ties. Some weak relationships to friends of friends are less beneficial than weak ties to strangers since the information and subsequent connections are likely to be similar to those of your friends. As any enterprise or any social network like facebook, instagram they have millions of data and to know how many alters are actively connected to ego or focal actor. Mhamdi and Hamdi (2004); Pun and Hamdi (2002).

Basically Dunbar's number is a suggest some cognitive limit so that the idea of the number of people with whom they maintained stable social relationships like "How often they send data to each other", "How often they are in contact of each other". As it's necessary so that an individual knows with whom each person is connected to him/her/enterprise and how each person relates to every other person. Dunbar theory is depend on our day to day life such that it explained with the number of people you would not feel embarrassed about joining or telling anything like what happened in class etc. It's very useful to assign any work to better alter, the alter having social contact with whom the work is concerned for better functioning. On the periphery, this theory can be used to guess the number also includes past colleagues or are they high school friends, or probability of them to meet again Xia et al. (2009); Wang et al. (2014a); Xia et al. (2007).

In network analysis, have lots of data and that data can be coming for any social media but that data doesn't give any information regarding time at what something happened Lin et al. (2012a); Xia and Hamdi (2006).

1. This is one of the problem for any digitalize company to have enough storage to store the data that coming from any form or links. So if the company have lots of data so it will be very cumbersome for organization to deal with such large number of data. For a limited period, it is used a simple technique like filtering the network into a smaller size, which allows us to compute the network in the quickest time feasible while preserving graph complexity. Xia and Hamdi (2006); Ma and Hamdi (2000); Hamdi and Lee (1997).
2. . Application Heterogeneity - This is also a big challenge for any network analysis organization is maintaining the application heterogeneity. The different graph applications and different network type it needs different kinds of inputs and different kind of formulas or function to interpret them. Since a consequence, it may become a problem for any organization, as graph indexing techniques must be limited to a certain kind of application in order to prevent becoming a problem. As a consequence, a company or other organization may restrict their application to companies that are somewhat similar to their own. Xia and Hamdi (2008); Wang et al. (2014b).
3. The data which are stored in the network that mostly comes from digital world but one problem lies here that the companies can have teleconference data, that is not always added into our dataset. In other ways all the telephonic conversation or face to face conversation is not usually added up, Lin et al. (2012b); Zedini et al. (2017); Wang et al. (2014c). The literature survey shows various

challenges and the possible solutions that can be consider as future work to do for conducting SNA in any organization or small enterprise for optimally functioning for any analysis.

Considered the links and ties that people can have and the type of relation people can have like relationships, including social, financial, and human property . As again there is some challenges exists and strength according to our network and our dataset. Mainly our techniques gets differ according to our network such as for small network or for huge network. Here, it is included some point that can be great for the organization for the manuscript.

## PROPOSED SYSTEM

### Methodology

#### *Data Cleaning and Pre-Processing*

Data Preprocessing – The documents to be converted to vectors, the data had to be converted to a form which could be easily read by the python script. The data that was given to me contained the profiles of the artists in the following format: Name, Area (Musician, Dancer etc.) and then their profiles. The tab in between these three was added manually so that a python script could read the given as a tab delimited file. Furthermore, another python script is written to abstract the name, the label and the profile from the data file and was stored in three different lists.

In this research work, it is compared the effective size and efficiency of samples to the ego network using ego measures. The effective size of the ego network as shown in Figure 2 is equal to the total number of alters that each ego has minus the average number of connections that each alter possesses ( $E_s$ ). Equation 1 may be used to determine the effective size of an undirected ego network with radius 1 in its simplest form. Efficiency ( $E_e$ ) is defined as the proportion of “non-redundant” connections between the ego and its surroundings. The word “efficiency” is an abbreviation for the term “effective network size” (Equation 2). As a consequence, it is an effective technique for comparing ego networks of different sizes.

Figure 2. Go-centric network analysis

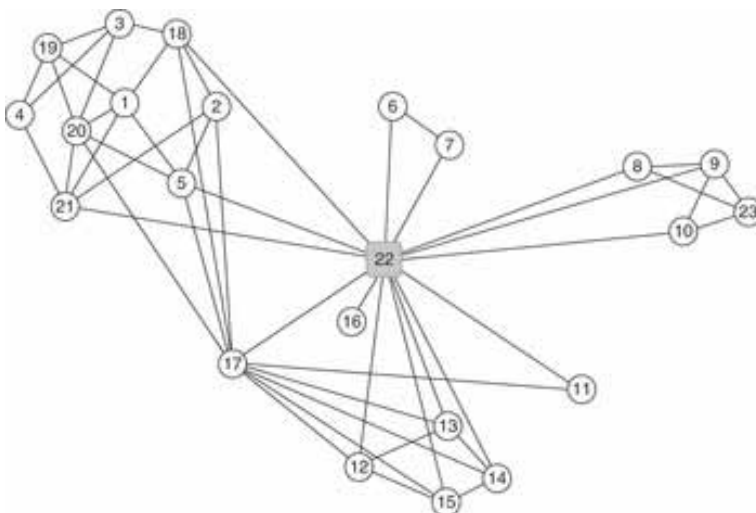




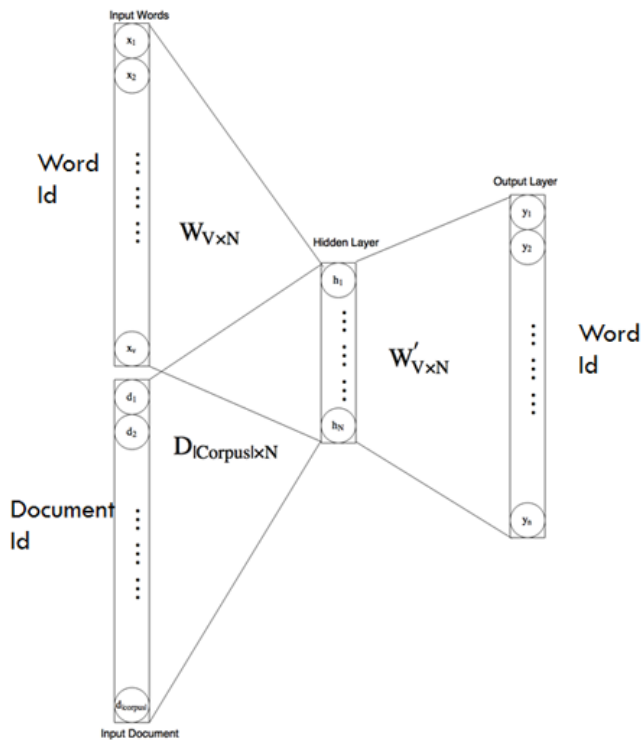
Table 1. Labels of different profiles and their corresponding count in the database

Labels	Count
Singer	705
Musician	583
Comedian	37
Comperer	64
DJ	18
Actor	122
Dancer	306
Party Organizer	32

Table 2. Model Selection and Training

Sno.	Iterations	Alpha	Win Size	Min Count	Training Time (mins)
1	30	0.0007	10	10	17
2.	20	0.002	10	10	10
3.	30	0.0025	7	5	23
4.	35	0.002	5	5	31
5.	15	0.0025	4	5	5

Figure 4. Doc2Vec Feature Vector 1



$$E_c = \frac{E_s}{A_n} (\text{target word} | \text{Context Words}, \text{Document Context})$$

At the end of the training process, you will have word embeddings,  $W$  and document embedding  $D$  for documents in the training corpus.

The optimization problem is not any different from the training problem.

$$E_c = \frac{E_s}{A_n} (\text{target word} | \text{Context Words}, \text{Document Context} = \text{testdoc})$$

One may however, choose to keep  $W$  and  $W^j$  fixed and learn variable  $D$  as document embedding.

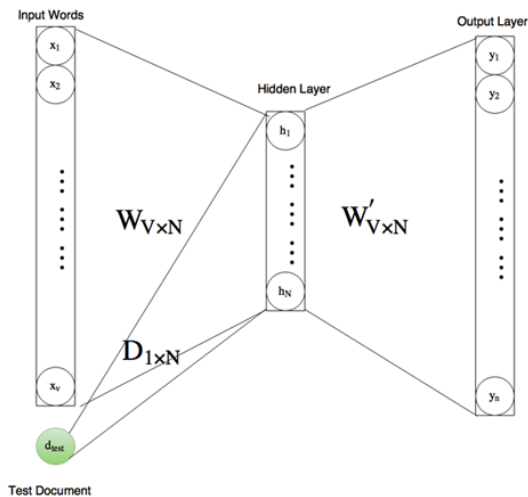
The Doc2Vec model was used to convert each document to a vector. All the profiles were extracted and a vocabulary/bag of words was created which contained all the words in the documents having a frequency greater than the count specified while training. Thus, given each document, it was converted to 300 dimensional vector. This was done by considering the average of the word vectors of the words present in that particular document. Based on this vector, similarity was calculated. The dimensions were chosen as 300 because it is a standard value. Moreover, pertained vectors for English words are also available for the given dimension.

## EXPERIMENTAL SET UP

Language and Tools needed for the implementation

- Python 3.6
- Libraries used:
- Scikit-Learn
- Matplotlib
- Gensim's Doc2Vec.

Figure 5. Doc2Vec Feature Vector 2



Python is an open source interpreter based programming language. It was built in the 1980's and is used for a lot of

Applications from powering Instagram to building video games to testing microprocessors at Intel to machine learning applications. Python has a design philosophy that is based on code readability and hence the code is much cleaner and has a lot more influence of English language as compared to languages as C or Java. A huge amount of freely available libraries allows a programmer to do whatever he requires, enabling Python to be one of the most used languages in the world at the moment.

Scikit-learn is an open source, widely popular machine learning library for Python. It features various classes and method to implement various models, based on the problem statement. Along with models, Scikit-learn also offers a plethora of tools to pre-process and manipulate data, making it one of the well known libraries for Python.

Matplotlib is a Python 2D plotting library that produces high-quality figures in a variety of hardcopy and interactive formats on a variety of platforms. Matplotlib is a Python library that may be used in scripts, the Python and IPython shells, the jupyter notebook, web applications, servers, and four GUI toolkits.

Doc2vec is an unsupervised technique for generating vectors from phrases and paragraphs in documents. The technique is based on a software called word2vec, which creates vectors from words. The vectors produced by doc2vec may be used to find phrases, paragraphs, or documents that are similar. doc2vec phrase vectors are independent of word order, unlike sequence models like RNN, which capture word sequence in generated sentence vectors.

## VISUALIZATION

Once vectorized, the documents were reduced down from 300 dimensions to 2 dimensions. To decrease the data's dimensionality, t-Distributed Stochastic Neighbor embedding was employed (t-SNE). It's a nonlinear dimension reduction technique for encoding high-dimensional data into a two- or three- dimensional space that may be shown in a scatter plot. Available components included Principle Component Analysis (PCA) however t-SNE was selected over pca as it displayed better clusters.

## OBSERVATION

The cosine similarity is applied on the reduced vectors and the distance matrix uses Euclidean distance two calculate the distance between two vectors as shown in Figure 6.

The system returns the Euclidean value calculated from the distance matrix shown is Figure 7

The top similar profiles are those which have minimum distance between them, and more of cosine similarity. For further clustering the top similar profiles based on their score t-SNE is preferred over principle component analysis because of better clustering result. K-means clustering is used for clustering the profiles.

## RESULTS AND DISCUSSION

The clustering and content based similarity check model clusters all the respective data in dataset into appropriate classes. Grouping together people of same interest, hobby or passion in same cluster. Improving the relevancy of the documents by increasing the precision and recall value.

After applying the clustering to the reduced vector from the dataset, the following plot is achieved where people of similar profile are clustered together differentiated from those of different interest and personality.



Figure 6. Cosine Similarity

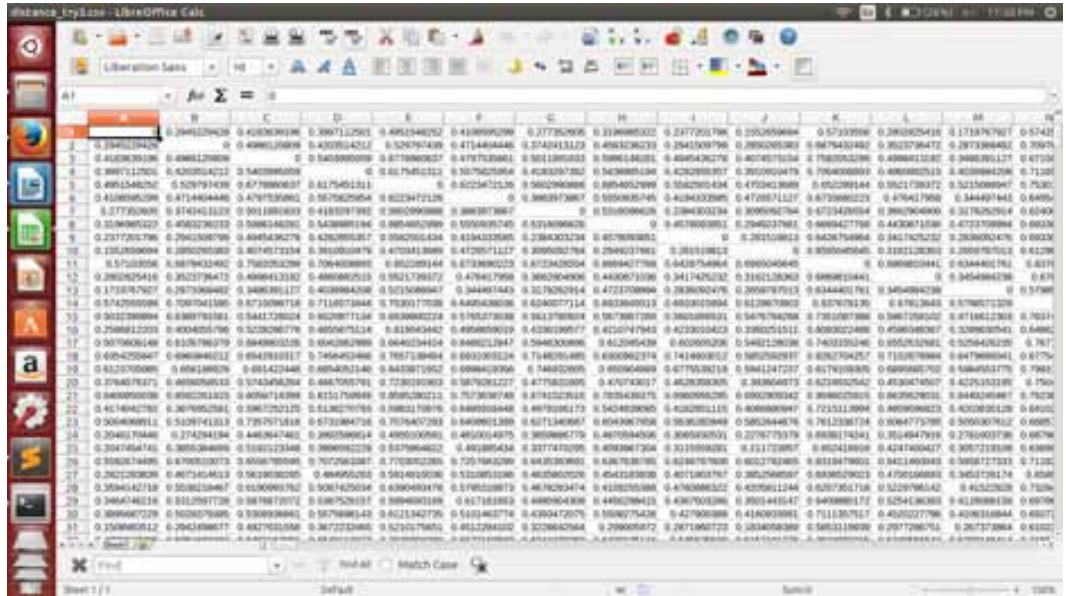
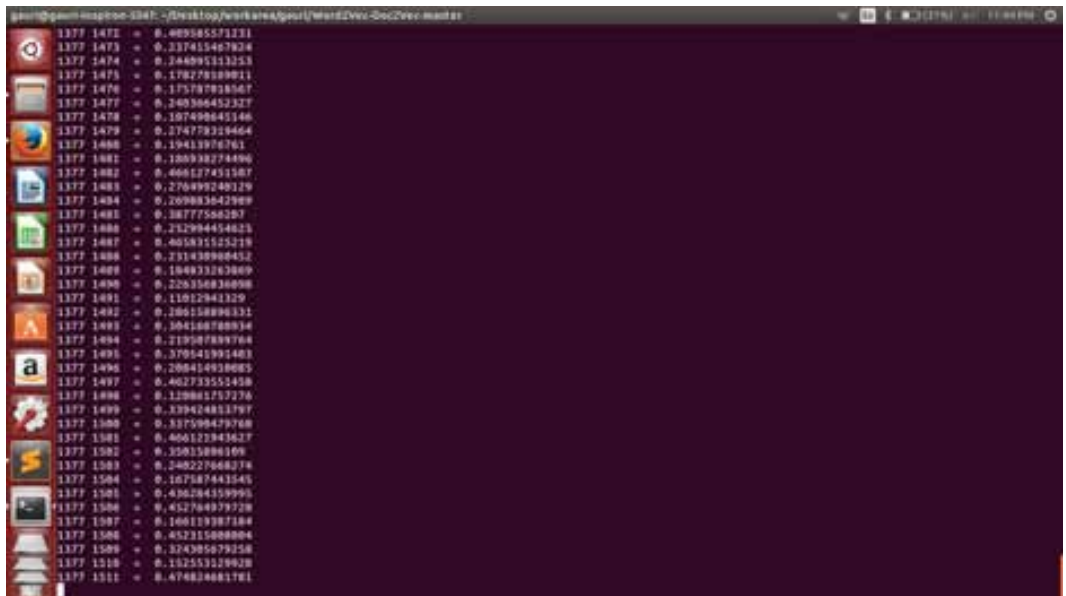


Figure 7. Similarity value



## CONCLUSION AND FUTURE WORK

Social and information networks, as well as their effect on relationships, erode the genuine connection between people (ego) and their online friends (alters). The ego network exemplifies the organization of the individual network. The egocentric perspective is popular because it places the emphasis on the person, group, or community. Effective factors include the size, structure, and makeup of ego

Figure 8. Top similar profile

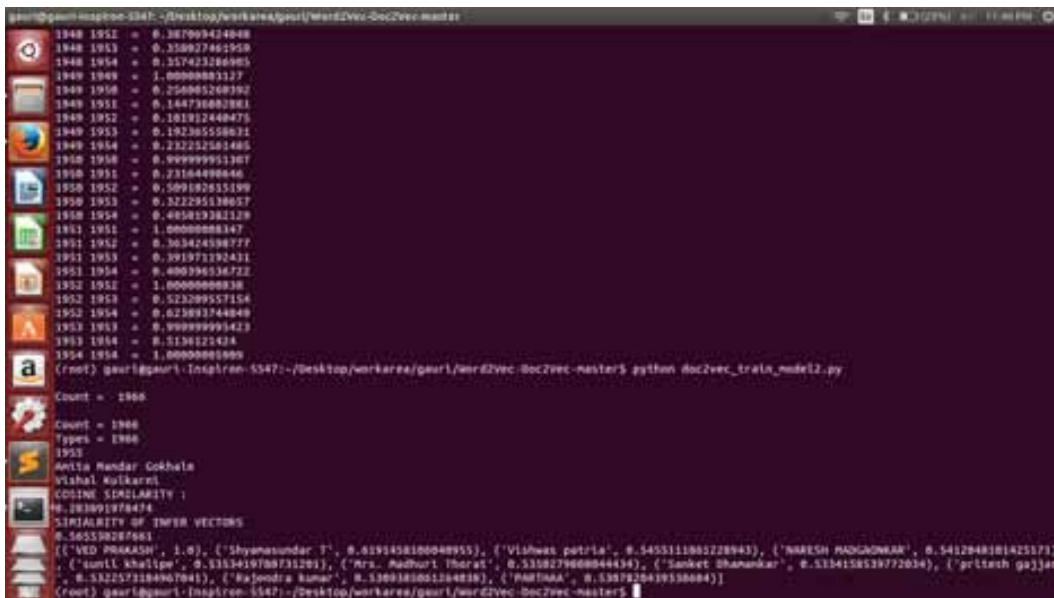
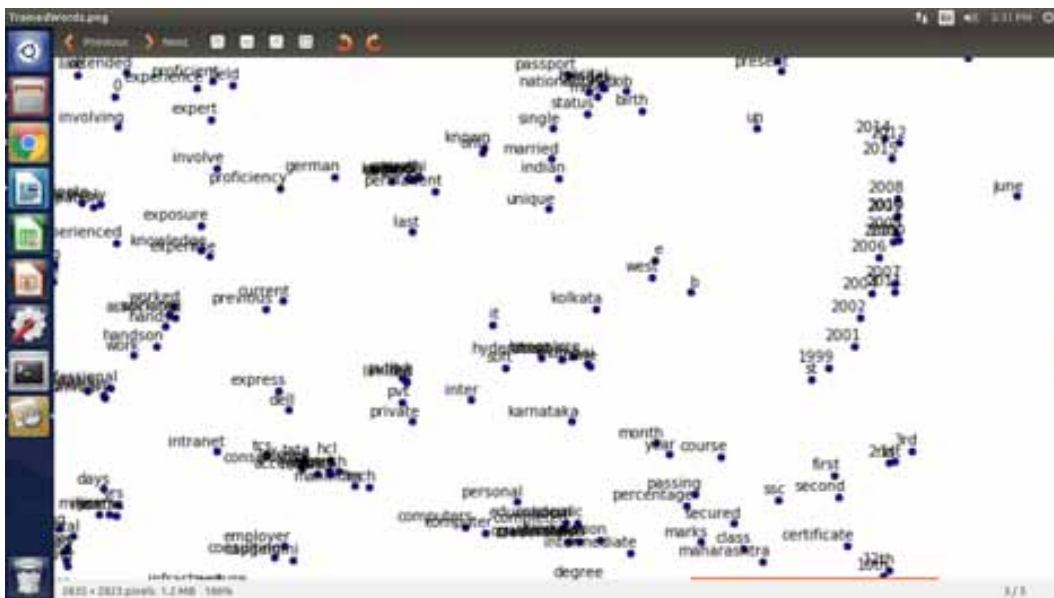


Figure 9. Word embedding

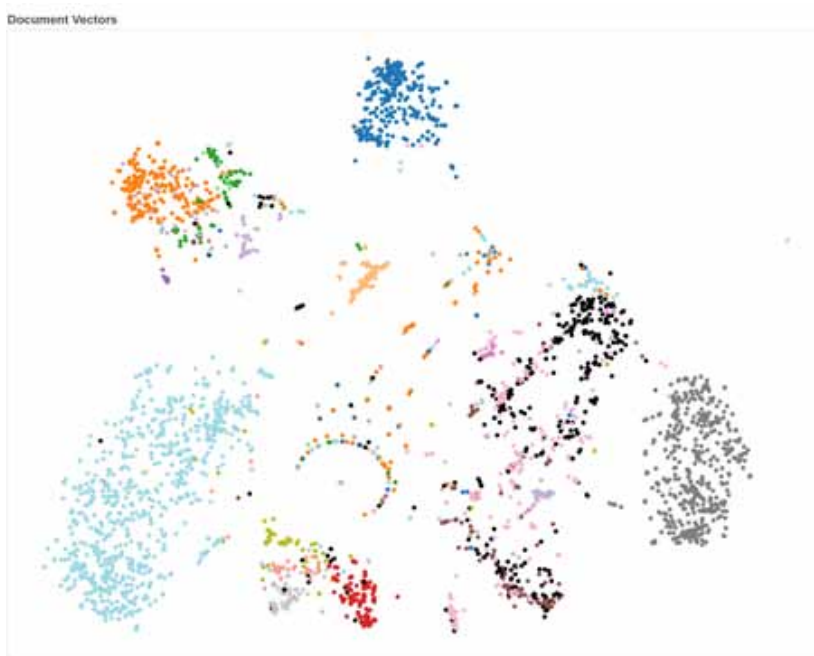


net- works. Additionally, the study examines the strength of connections and the degree to which ego–alter interactions exist. The degree offers an overview of the network in an introductory manner. The gap between the network’s degree of social support and its average strength is investigated. These findings demonstrate unequivocally that, although the manner in which individuals engage and maintain social connections is changing as a result of the emergence of Online Social Networks, the way in which people organize their social relationships seems to remain unchanged. With the growth

Figure 10. Doc embedding



Figure 11. Clustering



of online social networks, it is becoming easier to access a greater range of information. Any social media network is recommended based on the shared characteristics of the profile. In the majority of instances, this method is used. The keyword strategy is ineffective. Utilize the content similarity method rather than searching for and generating recommendations based on keywords. As a result, the system's precision and accuracy will improve, and people's methods for deciphering changes in their social networks (friends linked) have remained constant. In this research work, doesn't append other physical factors like enterprise location, social proximity access, social recommendation, social search, cyber security, anomaly detection, fraud detection. That can also be added later or in our future work and can increase the efficiency of our network analysis.

## REFERENCES

- Arnaboldi, V., Conti, M., Passarella, A., & Pezzoni, F. (2012). Analysis of ego network structure in online social networks. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 31–40. doi:10.1109/SocialCom-PASSAT.2012.41
- Bourouis, S., Band, S. S., Mosavi, A., Agrawal, S., & Hamdi, M. (2022). Meta-Heuristic Algorithm-Tuned Neural Network for Breast Cancer Diagnosis Using Ultrasound Images. *Frontiers in Oncology*, 12, 834028.
- Dhanaraj, R. K., Ramakrishnan, V., Poongodi, M., Krishnasamy, L., Hamdi, M., Kotecha, K., & Vijayakumar, V. (2021). Random Forest Bagging and X-Means Clustered Antipattern Detection from SQL Query Log for Accessing Secure Mobile Data. *Wireless Communications and Mobile Computing*, 2021.
- Dhiman, P., Kukreja, V., Manoharan, P., Kaur, A., Kamruzzaman, M. M., Dhaou, I. B., & Iwendi, C. (2022). A Novel Deep Learning Model for Detection of Severity Level of the Disease in Citrus Fruits. *Electronics*, 11(3), 495.
- Hamdi, M., Bourouis, S., Rastislav, K., & Mohmed, F. (2022). Evaluation of Neuro Image for the Diagnosis of Alzheimer's Disease Using Deep Learning Neural Network. *Frontiers in Public Health*, 35.
- Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., ... & Hamdi, M. (2022). Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques. *Computational and Mathematical Methods in Medicine*, 2022.
- Maurya, S., Joseph, S., Asokan, A., Algethami, A. A., Hamdi, M., & Rauf, H. T. (2021). Federated transfer learning for authentication and privacy preservation using novel supportive twin delayed DDPG (S-TD3) algorithm for IIoT. *Sensors*, 21(23), 7793.
- Muniyappan, A., Sundarappan, B., Manoharan, P., Hamdi, M., Raahemifar, K., Bourouis, S., & Varadarajan, V. (2022). Stability and numerical solutions of second wave mathematical modeling on covid-19 and omicron outbreak strategy of pandemic: Analytical and error analysis of approximate series solutions by using hpm. *Mathematics*, 10(3), 343.
- Poongodi, M., Bourouis, S., Ahmed, A. N., Vijayaragavan, M., Venkatesan, K. G. S., Alhakami, W., & Hamdi, M. (2022). A Novel Secured Multi-Access Edge Computing based VANET with Neuro fuzzy systems based Blockchain Framework. *Computer Communications*.
- Poongodi, M., Hamdi, M., & Wang, H. (2022). Image and audio caps: automated captioning of background sounds and images using deep learning. *Multimedia Systems*, 1-9.
- Poongodi, M., Hamdi, M., Gao, J., & Rauf, H. T. (2021, December). A Novel Security Mechanism of 6G for IMD using Authentication and Key Agreement Scheme. In *2021 IEEE Globecom Workshops (GC Wkshps)* (pp. 1-6). IEEE.
- Poongodi, M., Hamdi, M., Malviya, M., Sharma, A., Dhiman, G., & Vimal, S. (2022). Diagnosis and combating COVID-19 using wearable Oura smart ring with deep learning methods. *Personal and ubiquitous computing*, 26(1), 25-35.
- Poongodi, M., Hamdi, M., Sharma, A., Ma, M., & Singh, P. K. (2019). DDoS detection mechanism using trust-based evaluation system in VANET. *IEEE Access*, 7, 183532-183544.
- Poongodi, M., Hamdi, M., Varadarajan, V., Rawal, B. S., & Maode, M. (2020, July). Building an authentic and ethical keyword search by applying decentralised (Blockchain) verification. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 746-753). IEEE.
- Poongodi, M., Malviya, M., Hamdi, M., Rauf, H. T., Kadry, S., & Thinnukool, O. (2021). The recent technologies to curb the second-wave of COVID-19 pandemic. *Ieee Access*, 9, 97906-97928.
- Poongodi, M., Malviya, M., Hamdi, M., Vijayakumar, V., Mohammed, M. A., Rauf, H. T., & Al-Dhlan, K. A. (2022). 5G based Blockchain network for authentic and ethical keyword search engine. *IET Commun.*, 16(5), 442-448.
- Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., & Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 13(1), 16-27.

- Poongodi, M., Nguyen, T. N., Hamdi, M., & Cengiz, K. (2021). Global cryptocurrency trend prediction using social media. *Information Processing & Management*, 58(6), 102708.
- Poongodi, M., Sharma, A., Hamdi, M., Maode, M., & Chilamkurti, N. (2021). Smart healthcare in smart cities: wireless patient monitoring system using IoT. *The Journal of Supercomputing*, 77(11), 12230-12255.
- Poongodi, M., Vijayakumar, V., Al-Turjman, F., Hamdi, M., & Ma, M. (2019). Intrusion prevention system for DDoS attack on VANET with reCAPTCHA controller using information based metrics. *IEEE Access*, 7, 158481-158491.
- Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES. *Malaysian Journal of Computer Science*, 132-148.
- Ramesh, T. R., Vijayaragavan, M., Poongodi, M., Hamdi, M., Wang, H., & Bourouis, S. (2022). Peer-to-peer trust management in intelligent transportation system: An Aumann's agreement theorem based approach. *ICT Express*.
- Rathore, M. S., Poongodi, M., Saurabh, P., Lilhore, U. K., Bourouis, S., Alhakami, W., ... & Hamdi, M. (2022). A novel trust-based security and privacy model for Internet of Vehicles using encryption and steganography. *Computers and Electrical Engineering*, 102, 108205.
- Rawal, B. S., Manogaran, G., & Hamdi, M. (2021). Multi-Tier Stack of Block Chain with Proxy Re-Encryption Method Scheme on the Internet of Things Platform. *ACM Transactions on Internet Technology (TOIT)*, 22(2), 1-20.
- Rawal, B. S., Manogaran, G., & Poongodi, M. (2022). Implementing and Leveraging Blockchain Programming.
- Rawal, B. S., Manogaran, G., Singh, R., Poongodi, M., & Hamdi, M. (2021, June). Network augmentation by dynamically splitting the switching function in SDN. In *2021 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1-6). IEEE.
- Sahoo, S. K., Mudligiriyappa, N., Algethami, A. A., Manoharan, P., Hamdi, M., & Raahemifar, K. (2022). Intelligent Trust-Based Utility and Reusability Model: Enhanced Security Using Unmanned Aerial Vehicles on Sensor Nodes. *Applied Sciences*, 12(3), 1317.

*P. Tamil Selvi is an Assistant Professor in the Department of Information Technology and obtained Doctoral degree in Bharathiar University in the area of Networks Communications and Security. She has more than 10 years of teaching and 6 years of research experience. She has published more than 30 papers in various international journals and conferences. Her area of interest is Networks Security, Big data Analytics, Cloud Computing and IoT.*

*Kishore Balasubramanian has more than 17 years of academic experience in imparting Engineering Education. He had his Bachelor's Degree in Electronics & Instrumentation from Bharathiar University, India, Master's Degree in Applied Electronics from Anna University, India and Ph.D (Information and Communication Engineering) from Anna University, India. His research interests include Medical Image Processing and Computer Vision. He is an active reviewer in many SCI and Scopus indexed journals, conferences and editor in several scientific international journals. He has authored three books in the field of Analog Electronics and has published papers in international and national journals. He has received grants from CSIR, DRDO (Government Funding agencies) for conducting Faculty Development Programmes, Workshops and Conferences. He is a member of ISTE, IRED and IAENG. Presently he is working as an Assistant Professor (Senior Scale) in the Department of EEE at Dr. Mahalingam College of Engineering & Technology, Pollachi, India.*

*Jayapandian N. has received his PhD from Anna University, Chennai. He has completed M.E.(CSE) from Kongu Engineering College, Erode, Tamilnadu at 2009. He has completed his B.Tech.(IT) from Institute of Road and Transport Technology, Erode, Tamilnadu at 2006. He is active life Member of ISTE. He is currently doing his research in Cloud Computing in Anna University, Chennai. Currently, he is working as Assistant Professor in the Department of Computer Science & Engineering at Christ University, Bangalore. In his 12 years of teaching experience and one year of Industry Experience. His research interests are Grid Computing and Cloud Computing. He has published in 4 book chapters, 25 International Journal articles, 60 international and National Conferences.*

*Mounir Hamdi is the Founding Dean of the College of Science and Engineering at Hamad Bin Khalifa University. He is an IEEE Fellow for contributions to design and analysis of high-speed packet switching, which is the highest research distinction bestowed by IEEE. Before joining HBKU, he was Chair Professor and founding member of the Hong Kong University of Science and Technology (HKUST), and the Head of the Department of Computer Science and Engineering. In 1999 to 2000 he held visiting professor positions at Stanford University and the Swiss Federal Institute of Technology. He received the B.S. degree in Electrical Engineering - Computer Engineering minor (with distinction) from the University of Louisiana in 1985, and the MS and the PhD degrees in Electrical Engineering from the University of Pittsburgh in 1987 and 1991, respectively. His area of research is in high-speed wired/wireless networking in which he has published more than 360 publications, graduated more 50 MS/PhD students, and awarded numerous research grants. In addition, he has frequently consulted for companies and governmental organizations in the USA, Europe and Asia. He is also a frequent keynote speaker in International Conferences and Forums, is/was on the Editorial Board of more than 10 prestigious journals and magazines, and has chaired more than 20 international conferences and workshops. In addition to his commitment to research and academic/professional service, he is also a dedicated teacher and quality-assurance educator. He received the best 10 lecturer award and the distinguished engineering teaching appreciation award from HKUST. He is frequently involved in higher education quality assurance activities as well as engineering programs accreditation all over the world.*