

Evaluating Frequency of words and Word Cloud from Astrological sentiments using NLP

C. N. V. B. R. Sri Gowrinath¹, Dr. Ch. V. M. K. Hari², Prof. P. G. V. D. Prasad Reddy³

¹Department of MCA, CBIT(A), Hyderabad, Telangana, India

gowrinath_mca@cbit.ac.in

²Department of Computer Science, Dr. V. S. Krishna Government Degree College, Visakhapatnam, Andhra Pradesh, India

kurmahari@gmail.com

³Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India

prasadreddy.vizag@gmail.com

ABSTRACT

Article Info

Volume 8, Issue 3

Page Number : 920-928

Publication Issue

May-June-2021

Article History

Accepted : 18 June 2021

Published : 25 June 2021

The identification of interest/disinterest over a notion is having a huge demand in the current competitive data analytical world. For example, the customer preferences in various seasons, approximate visitors to a tourist place based on scenarios like weather and special occasions in the place, and so on. While giving an opinion on any concept, natural language in form of sentences/words/symbols/ratings plays a vital role. Depends upon the context and usage of natural language, captured opinions can be interpreted as either in a positive or negative sense. The terminology used for providing the opinions is used for analysing the data in an easy way. The evaluation of the word frequencies and word cloud are identified accurately, only after a keen analysis of the collected opinions.

The Term-Document Matrix is one of the techniques that identify the frequency of words in each and every document/row in the given dataset, which can be used to generate the word cloud. In this paper to identify the frequency of words from the opinions given by multi-domain personalities on Astrology, distinct Natural Language Processing (NLP) techniques are used. A word cloud can also be generated from the set of words used for the astrological dataset.

Keywords : Natural Language Processing, Astrology, Word Cloud, COVID-19, Knowledge Management System, Parsing.

I. INTRODUCTION

Astrology is a multidisciplinary, time-variant, persistent, addressable, cognitive approach and also one of the oldest domains. The basic interpretation regarding Astrology is that it provides better ideas/solutions/remedies to overcome the hurdles that are facing in daily life. The flip side interpretation is that Astrology is only an illusion but not real. Contemporaneous Astrological predictions are superficial but not scientific and so there is a huge number that negates Astrology. The inference about Astrology is its literature suggests that it should be interpreted as interactions within the Cosmological symmetry between individuals rather than planets and individuals [2]. Similarly, another notion is that an individual can ascertain the needs of Vitamins, also based on Zodiac signs for the health of a human [3].

Astrology is definitely domain-oriented and it consists of technical specifications such as computations, analysis, modelling. The number of clients, as well as astrologers, is also increasing day by day in the current situation. During the COVID-19 situation, many of the astrologers enlightened their knowledge and provided useful information to reduce panic. Based on the captured / historical knowledge, the movement of the planets in astronomy is the heart of situations like COVID-19. It is very important to explore Astrology so that it computes more accurate results by imposing less number of questions [16]. For this, an affluent AI system is required to predict in an efficient way.

For any scenario, relevant data collection is a vital thing and it is a part of opinion mining that relies on vector extraction to classify the relevant features such as term frequency and presence. The collection of the relevant data takes more time than other activities during analysis. The pre-processing of data is crucial for model generation and it decides the accuracy of the model while testing the model. Model generation is not always clear as well as accurate for all situations. Technology like Term-Document Matrix is suitable for the situation where we don't need a specific and complex model to analyse our data. To deal with the identification of interest/disinterest situations, distinguish NLP activities will be used for better analysis. One of the objectives of model generation is Knowledge Discovery (KD). With the exponential growth of data, Big Data classification approaches are used to improve traditional classification methods in knowledge discovery [21]. Advances in Artificial Intelligence resulted in a huge number of applications for analysis and prediction [17].

NLP stands for Natural Language Processing and is frequently using notion for application design as well as research purpose. NLP is an interface between computers and human languages. NLP works in association with distinct activities like speech recognition, language linguistics, statistical models, POS tagging, Machine translation, and others. The natural language statements are generally processed using 5 modules – Lexical Analysis, Syntactic Analysis, Semantic Analysis, SQL transformation, and SQL execution [6]. The present work takes assistance from these modules that were already implemented using the interpreted programming language Python.

In some NLP applications datasets also grouped into training and testing datasets, where the training dataset is used to build a model and the test set is to find out the accuracy of the developed model. Training and testing classification of entire data sets is not possible for large data sets. Hence for

classification of equal or preferable proportions of training and testing, use methods like Random Forest ensemble method or Cluster Sampling [22] or any other relevant. With the technologies like Python, the model building is a built-in functionality, and hence overhead is less. More and more languages are also available in the present digitized market to perform NLP. Out of which some familiar tools are R, Hadoop, Python, Scala. Python is an interpreted and ease-of-use language that is fulfilling the requirements of various clustered clients. In Python, the package (module) nltk (natural language tool kit) is used for Natural Language Processing. This module is having the provisions such as classification, stemming, lemmatization, tokenization of natural languages. Also, the module nltk offers easy-of-interfaces for various corpus and lexical resources. The modules corpus, tree bank, punkt, wordnet, wordcloud, webtext are various sub-modules of nltk.

In the current study, the objectives are to identify the word frequency and word cloud building from the Astrological dataset mined. The insight of dataset is different for different individuals. It means that the same dataset is used for multiple extractions using techniques like skip-gram, CNN [7], health management [1], hotel reviews using classification methods [9], Sentiment Analysis [15], profession classification using AI [17].

II. LITERATURE SURVEY

The notion of Astrology is not bogus and it is a scientific aspect. Every human life or any incident occurrence will be followed based on mathematical patterns such as sinusoidal waves with ups and downs over a specific time period [16]. For example, during the COVID-19 pandemic situation, astrologers are forecasting the end of this pandemic based on planetary position [5]. There is a definite separation between Astrological and non-astrological artifacts. The non-astrological artifacts should not be

overshadowed by the Horoscope chart representation which is a method for pictorial description for Astrological predictions. The clear separation between Astrological and non-astrological artifacts is being interpreted using a good test design known as “eminence” [2].

The Astrological prediction conduction using Natural Language Processing is an efficient technique to evaluate distinct opinions. For some approaches, predictions can happen based on various classification algorithms such as ZeroR, Simple CART, and Decision Tree [17]. For example Stock prices prediction [7]. Even some eminent studies specify that Vitamin development for a human being is also depending on various planets [3]. The data required to process the predictions are captured from various sources such as blogs, social media [19], CSV files [6], and normal text files. The social web is providing new tools to create and share ideas with others, connected in WWW and the information shared is unstructured in general.

Text mining is one of the activities after collecting the relevant data and is done using techniques TidyLexicons for Sentiment Analysis [15]. While collecting the data/opinions from users, a method like attention mechanism [8] is used to identify the close relation between cue word and response word. In general, the data collected initially has noisy data like missing values, outliers, inconsistencies. In some applications, computers learn how to predict and classify from unknown, large, noisy, and complex data sets [17] as well as in some other applications, uncertain values are collected as value for data items and so Decision Tree-like classifiers are used to convert uncertain to certain [20]. But for other datasets, some pre-processing techniques of data should be required before any model generation or analysis.

One of the insights of Natural Language Processing (NLP) is to identify the association words from a given plain text corpus and this word association reveals the psychological connections of humans with a particular perception [8]. The other insights from NLP is Music Emotion classification [18], student's performance, profession satisfaction, movie reviews, text summarization, sentiments of citizens towards various political parties during the pre-election process [19], student's response towards ICT based education [25], Text extraction from hoardings [23] and belief in astrology [11]. NLP is a strong tool in such a way that it can identify the credit transfer for a standard course in different universities without studying the course more than once [13]. Some basic studies supported that astronomy improvising student attitudes towards science and evaluation capability [4]. The attitudes can be identified by their opinion extraction. The opinions not only from students but from all are being identified using Social media interfaces like Twitter [19], Facebook, Instagram.

Word cloud exploration under a text corpus provides a distillation of text down to words along with their frequencies [10] and the concept is pure text summarization. Sentiment analysis can be performed on the exploited Word Cloud using the method like Lexicon based method [12]. While converting the user opinions into tokens/words then to Word Cloud, parsing is an essential trick for successful conversion. The multi-lingual dependency parser is one of the solutions to reduce the feature engineering that is often required for parsing [14]. OCR (Optical Character Recognition) is an approach that is used in a variety of industry and academic applications to recognize text inside images such as scanned documents and photos [24].

III. Proposed Work and Methodology

In recent times NLP tools are producing effective and efficient applications, which are useful for society.

NLP is an interdisciplinary field that consists of domains such as Artificial Intelligence (AI), Information Retrieval Systems (IRS), Big Data Analytics (BDA). It is used to process and analyse voluminous, variety, and versatile natural language data using programming languages like Python, R, Scala in view of technical, domain-oriented, and statistical aspects. Python Keras framework is another module to conduct deep learning techniques for NLP. Based on the insight of analysis, the required dataset can be generated either from scratch or reused the existing one, such as to identify the relationship between prompting and retorting words. The insight and feature extraction from the dataset plays a vital role in NLP and is approximately 60 to 70 % of the overall effort. To identify the absolute insights as well as feature extraction, domain knowledge is very much useful and it represents the type of data to be extracted. Once insight(s) is clear then the analysis intention shifted to feature extraction followed by model evaluation. The generation of models in form of NLP activities is also crucial for processing the natural language data and its effort is about 20%. The remaining effort is conducting predictions and decision-making.

The word frequency is used to evaluate the number of occurrences of each word in the dataset. But the raw datasets consist of more prepositions as well as punctuations and hence removing these also is an important activity. In the raw data set, the information is of sentence form and can be converted into words also referred to as tokens. This process is referred to as Tokenization. The raw data will be pre-processed in any model-building to remove noisy data.

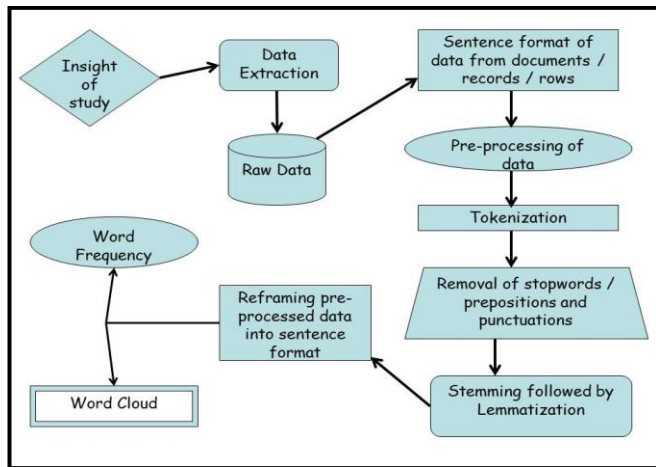


Fig.3.1 Architectural diagram for Word frequency and Word cloud generation

In NLP removing prepositions, punctuations, and stopwords is coming under pre-processing techniques. After pre-processing, Stemming and Lemmatization conducted as part of model building. The word cloud on the Astrology dataset can be generated using the words used most number of times in the dataset. Applying further predictions and generating reports based on the requirement are the next activities. Taking decisions within the business/domain constraints and observing the feedback leads to accuracy identification for the model.

IV. Methodology

- The initial step in the current work is to collect the raw data that consists of Astrological opinions of persons/professionals from multiple domains/occupations. The data collection was done using a Google form, as it stores the submitted opinion automatically in a CSV file. Also, data were collected manually from people who are unable to avail the technology of Google form.
- The collected data is in the form of Excel sheets and might contain noisy data. Here noisy data means – data inconsistencies, outliers, missing values, and duplications.
- Different pre-processing methods are used to clean, transform and load the data. To pre-process the data, different Python in-built

methods under different modules are used. The pre-processing techniques convert data into information.

- The post-processed information will be converted into sentence format from individual records and then tokens (words) generated from sentences.
- Stemming and Lemmatization procedures are used to identify the common root (stem) words for similar types of words. The root words are converted into standardized words using Lemmatization.
- Before identifying the frequency of each word in the dataset, the prepositions, stop words, as well as punctuations, are removed from the data as they are having zero significance in data analysis.
- Tokens after removal of prepositions, stop words and punctuations, are reframed into sentence format to generate the word cloud.
- The word cloud is displayed in the rectangle format that consists of different words used in the collected information. Also, frequencies of few words are represented in the bar graph in descending order.

V. Result Analysis

In this paper Python programming language is used, which is more appropriate than other programming languages for Astrological data analysis. During the current study, distinct results were generated and are useful for further processing. After capturing the data using Google form, the raw information will be undergone different methods such as `open()`, `read_csv()`. The initial outcome was, the record-wise data display in sentence form as shown in the below figure.

From the dataset the data under the column name 'Conclusion' will be displayed in sentence format. The Doc2Vec is an unsupervised learning algorithm that generates vectors from documents/sentences and similarly, Word2Vec generates vectors for words. In

some cases, Ontology learning framework like methods also used to process the natural language and evaluate the required result, as the framework contains Term extraction, Ontology building and Ontology pruning as pre-processing steps. The proportionate result from the word similarity based on domain-specific using Word2Vec is more rather than using any unsupervised learning.

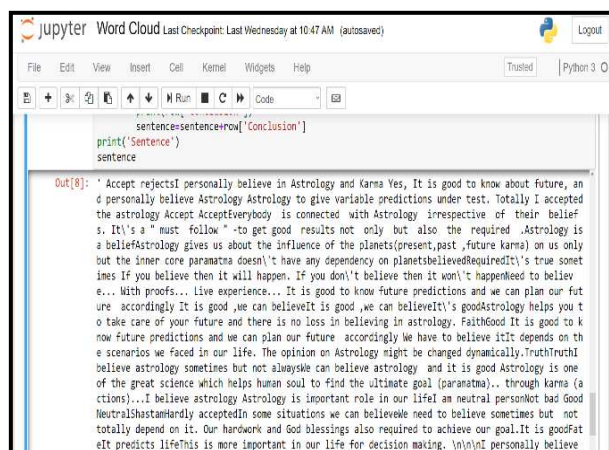


Fig.4.1 Sentence format for record-wise data

The next upshot is converting sentence format into tokens and for this purpose word_tokenize() in python is used. The output from this method is a set of tokens that consists of individual words along with prepositions, stopwords as well as punctuations. For better result analysis the null significant words are filtered and extracted only the relevant information. The filtered data will be extracted for analysis after removing stopwords like 'i', 'me', 'my'; prepositions 'to', 'and', 'by'; punctuations ',', '!', and the same will be represented in the below screenshot. In general, a word has many forms, and hence the process of Stemming and Lemmatization used. At this stage, the data contains a duplication of words and is able to display the word frequency in a better and effective manner.

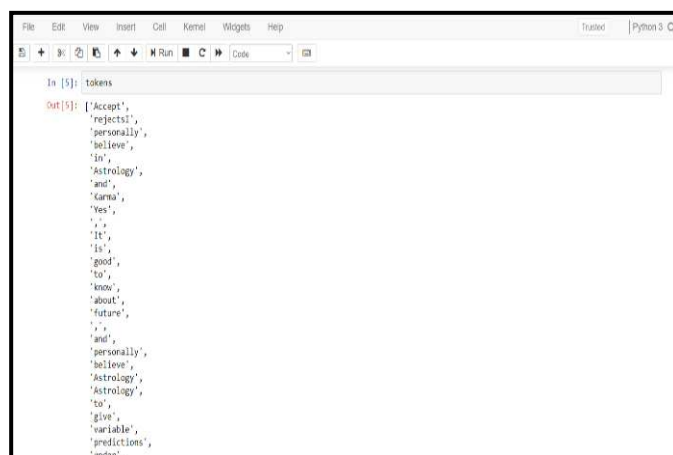


Fig.4.2 List of tokens after pre-processing of raw data
The frequency of the words in the astrological word cloud is shown in the below figure-4 and the diagram shows the frequency in descending order. The figure **Fig.4.3** is representing few words along with their descending order of frequencies.

In the final stage, the duplicate words are removed and only one time a particular word should be considered. For this, the attribute collocation takes 'False' as a value to ignore duplicates. The collocation attribute is built-in in the method WordCloud, which is in the wordcloud module.

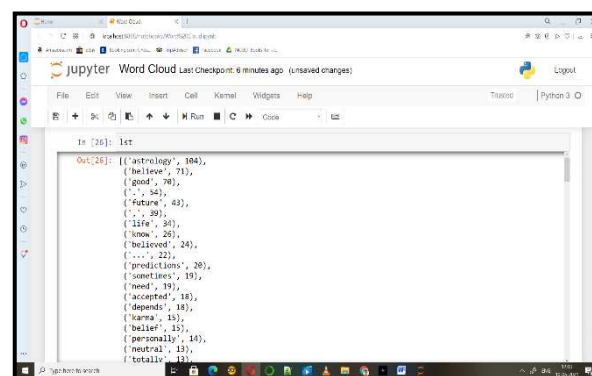


Fig.4.3 Frequency of words in descending order

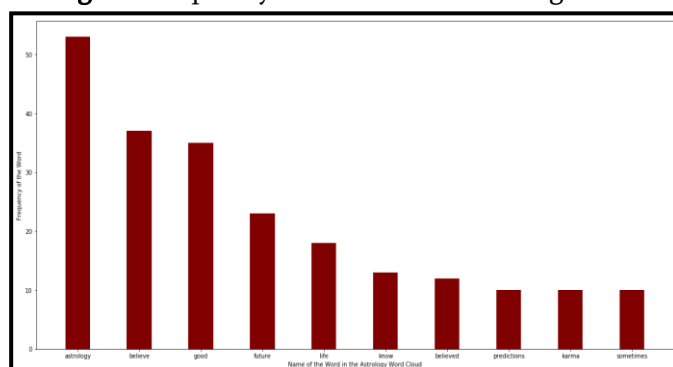


Fig.4.4 Graphical Bar chart representation for frequency of words in descending order



Fig.4.5 Representation of Word cloud for Astrological dataset

The Word Cloud exploration provides central visualization of the collected text corpus and integrates several interactive features into a consistent framework for interactive text analysis. The diagram **Fig.4.5** represents the overall structure for the word cloud of the Astrological database.

VI. Conclusion and Future Enhancements

In the upcoming future, there is a drastic change of people who believe in Astrology and so a definite effective and accurate exploratory AI system is required. One of the reasons behind the familiarity of Astrology is the occurrence of the COVID-19 pandemic. The future enhancement for the current study is to collect more distinct corpus and find out the associated correlation between different words using a Knowledge Management System (KMS). To fulfil this, different algorithms such as Association Rule Mining algorithms, Reading Comprehension (RC) algorithm, Clustering, Classification are to be used. The future scope with the Astrological KMS is to identify the real-time computations behind the predictions and to improve the accuracy of Astrological predictions using distinct Data Science aspects.

VII. REFERENCES

- [9]. T. Ghorpade and L. Ragha, "Featured based sentiment classification for hotel reviews using NLP and Bayesian classification," 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, 2012, pp. 1-5, doi: 10.1109/ICCICT.2012.6398136.
- [10]. F. Heimerl, S. Lohmann, S. Lange and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, 2014, pp. 1833-1842, doi: 10.1109/HICSS.2014.231.
- [11]. C. N. V. B. R. Sri Gowrinath, B. Srinivasa S. P. Kumar, Chilukuri Megh Phani Dutt, "Identification of Astrological belief using Sentimental Analysis by capturing Opinions from cross-domain individuals," International Journal of Recent Technology and Engineering (IJRTE), ISSN : 2277 – 3878, Volume – 8 Issue – 5, January 2020.
- [12]. X. Fei, H. Wang and J. Zhu, "Sentiment word identification using the maximum entropy model," Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010), Beijing, 2010, pp. 1-4, doi: 10.1109/NLPKE.2010.5587811.
- [13]. A. Heppner, A. Pawar, D. Kivi and V. Mago, "Automating Articulation: Applying Natural Language Processing to Post-Secondary Credit Transfer," in IEEE Access, vol. 7, pp. 48295-48306, 2019, doi: 10.1109/ACCESS.2019.2910145.
- [14]. S. Jaf and C. Calder, "Deep Learning for Natural Language Parsing," in IEEE Access, vol. 7, pp. 131363-131373, 2019, doi: 10.1109/ACCESS.2019.2939687.
- [15]. Niharica Choubey, Vilender Kumar, Sanjay Kumar Gupta, "Use of Tidytext Lexicons Approaches for Sentiment Analysis," International Advanced Research Journal in Science, Engineering and Technology, ISSN (Online): 2393 – 8021, Vol. 7, Issue 1, January, 2020.
- [16]. Dr. Pijush Kanti Bhattacharjee, "Scientific Astrological Prediction of Human Life", International Journal of Jyotish Research: 2020; 5(1): 12-16.
- [17]. N. Chaplot, P. Dhyani and O. P. Rishi, "Astrological prediction for profession using classification techniques of artificial intelligence," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 233-236, doi: 10.1109/CCAA.2015.7148378.
- [18]. W. Shi and S. Feng, "Research on Music Emotion Classification Based on Lyrics and Audio," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2018, pp. 1154-1159, doi: 10.1109/IAEAC.2018.8577944.
- [19]. Neetu Narwal, Kavita Pabreja, "Social Media Analytics", CSI Communications, ISSN: 0970 – 647X, Volume No. 42, Issue 5 & 6, pp. 10-14, August – September 2018.
- [20]. Ramesh Ponnala and P. Vasanth Sena, "Induction Decision Trees for Tentative Data", International Journal of Computer Science and Management Research (IJCSMR), Vol 2, Issue 10, pp. 3463 – 3478, October 2013, ISSN: 2278 – 733X.
- [21]. M. Ramchander and P. Krishna Prasad, "Big Data Classification approaches towards improving traditional classification algorithms", Alochana Chakra Journal, Volume IX, Issue VI, June 2020, pp. 3040-3048.
- [22]. Lakshmi Sreenivasa Reddy D, "Cluster Sampling to improve Classifier accuracy for Categorical data", International Journal of Applied Engineering Research, ISSN:0973-4562, Volume 14, Number 13 (2019), pp.2995-3002.
- [23]. D. Jayaram, J. Shiva Sai, CRK Reddy, V. Kamakshi Prasad, "Text Extraction from Hoardings by Hybrid Model", International Journal of Engineering and Advanced

Technology (IJEAT), ISSN: 2249-8958, Volume-9, Issue-4, April-2020, pp. 73-80.

- [24].M. Kalidas, Madhu Bajaj, "Assessment of Optical Character Recognition Techniques for Hindi Language", International Journal of Innovative Research in Science, Engineering Technology, Volume 8, Issue 12, December 2019, pp. 11717-11726, doi:10.15680/IJIRSET.2019.0812042.
- [25].Dr. B. Indira, Dr. D. Lalitha Devi, "Students response towards ICT based education among various schools and colleges in Hyderabad – a case study", International Journal of Scientific Research, Volume-8, Issue-12, December-2019, ISSN No. 2277 (print), doi:10.36106/ijsr.

Cite this article as :

C. N. V. B. R. Sri Gowrinath, Dr. Ch. V. M. K. Hari, Prof. P. G. V. D. Prasad Reddy, " Evaluating Frequency of words and Word Cloud from Astrological sentiments using NLP ", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 8, Issue 3, pp.920-928, May-June-2021. Available at
doi : <https://doi.org/10.32628/IJSRST2183196>
Journal URL : <https://ijsrst.com/IJSRST2183196>

Authors



C.N.V.B.R.SRI GOWRINATH received the M. Tech. Degree in Computer Science and Engineering from JNTUK affiliated college in 2013, currently working as a faculty in the department of MCA in CBIT, Hyderabad. His research interests are Artificial Intelligence and Data Mining.



Dr. CH.V.M.K.Hari is Head of the Department of Computer Science in Dr.V.S.Krishna Government Degree College(A), Visakhapatnam. He is having two decades of experience in Computer Science. He served as Assistant Professor in Nagarjuna University and AdiKaviNannaya University. He published more than 40 research papers, of which a majority are in impact journals. He presented so many papers at various International Conferences.



Prof. P.V.G.D. Prasad Reddy is a passionate teacher, an expert in the field of Computer Science & Information Technology, and an academic administrator for over three decades.

He has held the positions of Vice-Chancellor (FAC), Rector, Registrar, Head of the Department, and numerous other Administrative Positions in the University. He had been on committees of several public institutions under assignments of great sensitivity and responsibility. He published more than 195 research papers, of which a majority are in impact journals. He is one of the cited researchers in the field of Information Technology and Data Sciences with an h-index of 16 and i-10 index of 28. He has handled various IT consultancy works for Government organizations and other clients.