iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** III **Month of publication:** March 2017

DOI: http://doi.org/10.22214/ijraset.2017.3002

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue II, March 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Efficient Deep Learning Approach for Dimensionality Reduction using Micro blogs from Big data

Mr.M.Vengateshwaran[1], Mrs.C.Ramyapriyadarsini[2], Ms.N.Valarmathi[3]
[1]M.E., Asst. Professor & HOD(CSE), [2,3]M.Tech Asst. Professor(CSE)
Mass Group of Institutions Kumbakonam

Abstract: ***Nowadays Information Technology plays a vital role in every aspects of the human life. Now a world, the huge amount of stored information has been enormously increasing day by day which is generally in the unstructured form and cannot be used for any processing to extract useful information. Exploring potentially useful information from huge amount of textual data produced by micro blogging services has attracted much attention in recent years. An important preprocessing step of micro blog text mining is to convert natural language texts into proper numerical representations. Due to the short-length characteristics of micro blog texts, using term frequency vectors to represent micro blog texts will cause "sparse data" problem. Finding proper representations of micro blog texts is a challenging issue. In this project, we apply deep learning networks to map the high-dimensional representations of micro blog texts to low-dimensional representations. To improve the result of dimensionality reduction, we take advantage of the semantic similarity derived from two types of micro blog specific information, namely the retweet relationship and hash tags. Two types of approaches, including modifying training data and modifying the training objective of deep networks, are proposed to make use of micro blog-specific information. To improve the efficiency we implement the system in Hadoop. In addition to that to make services effective. To achieve the scalability and efficiency with help of map reduce framework in a big data environment.***

Keywords:- Microblogs,Deeplearning,Text mining, semantic similarity,Bigdata.

I. INTRODUCTION

Deep learning is an set of an algorithm in machine learning that attempt to model high level abstractions in data by using model architecture composed of multiple nonlinear transformation. deep learning is a part of broader family of machine learning methods based on learning representation of data. Eg. Image –represented in many ways i.e. vector of pixels. Deep learning is an approach and an attribute to learning, where the learner uses higher order cognitive skills such as the ability to analyze, synthesize, solve problems &thinks meta cognitively in order to construct long term understanding. It involves the Critical analysis of new ideas, linking them to already known concepts, and principles so that this understanding can be used for Problem solving in new, unfamiliar contexts. A central idea of deep learning is referred to as greedy layer wise unsupervised pre-training, which is to learn a hierarchy of features one level at a time. The features learning process can be purely unsupervised, which can take advantage of massive unlabeled data. The feature learning is trying to learn a new transformation of the previously learned features at each level, which is able to reconstruct the original data. The greedy layer wise unsupervised pre-training is based on training each layer with an unsupervised learning algorithm, taking the features produced at the previous level as input for the next level.

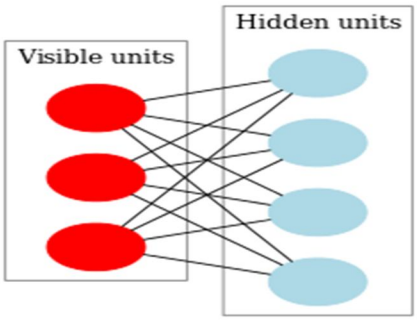


Fig.1 Deep Learning

Finally, the set of layers with learned weights could be stacked to initialize a deep supervised predictor, such as a neural network classifier, or a deep generative model, such as a Deep Boltzmann Machine.

II. LITERATURE REVIEW

A. Feature Extension for Short Text Author: Yan Tao, Wang Xi-Wei

In this paper different from the conventional word-form based automatic classification system of Chinese texts, giving further consideration on words co-occurrence relationship. The development of instant messaging technology and the popularization of information processing technology promoted the booming of short-text information processing technology, such as the mobile phone SMS, QQ chat, BBS, instant messaging software, has become an important channel for information dissemination. Such a rich resource of short texts makes people's lives easier while bringing significant information security risks. Such as waste, harassment and frequent large quantities of text messages, seriously affecting people's lives, and short text classification is the realistic tasks basis to solve the short message filtering, to promote the Chinese short-text classification has become an important research direction. Most methods of the short text is mainly traditional text classification, information filtering and retrieval methods, Specific algorithm for short text has not yet formed on its own characteristics. However, compared with the long text, because of the characteristics that short text described weak signals, noise characteristics of the data and automatic classification system of Chinese texts based on simple word-form have been unable to meet the needs of short text classification. Therefore, the comprehensive consideration of the short text data is proposed in the course of a short text classification, mining the association relationship between the short text data to assist the classification. Currently, only a small number of research at home and abroad in this area and the results are unsatisfactory. Previous studies show that the text feature rich or not is essential to the classification results. So, two methods are proposed for feature extension based on taking full account of the correlation between words: In this paper, take the training data as the background corpus, firstly, using FP-Growth algorithm to mine the co-occurrence relationship among the training set, and to construct the set of feature co-occurrence as expansion vocabulary, and then expanded the training and testing features using the set of feature co-occurrence respectively, the two methods were based on the same expansion vocabulary, but expand in different ways. Finally, experiments were carried out respectively.

Show that the method in the multi-class short-text classification is effectiveness, and has lower performance as well as inaccuracy than the traditional classification.

B. Enhancing Accessibility of Microblogging Messages using Semantic Knowledge Author: Xia Hu, Lei Tang, Huan Liu

Microblogging services such as Twitter are increasingly used for communicating breaking news, information sharing, and participating in events. This emerging medium has become a powerful communication channel in recent digital revolutions. However, the accessibility of these messages has been very limited so far. Tweets and retweets of a user's followers' appear alongside the user's own tweets in reverse chronological order. People often have only patience to skim through the first 20 - 50 messages. When the messages become over-whelming, it is impractical for a user to quickly gauge the main subjects from their followers' posts. To make a large collection of micro blogging messages accessible to users, current web systems need to provide not only accurate clusters for subtopics in micro blogging messages, but also meaningful labels for each cluster. Enhancing the accessibility of micro blogging messages entails two tasks: (1) cluster micro blogging messages into manageable categories, and (2) assign readable and meaningful labels for each cluster of messages. Unlike standard text with many sentences or paragraphs, micro blogging messages are noisy and short. In addition, micro bloggers, when composing a message, may use or coin new abbreviations or acronyms that may seldom appear in conventional text documents. Furthermore, these short messages do not provide sufficient contextual information to capture their semantic meanings. Traditional text mining methods, when applied to micro blogging messages directly, lead to unsatisfactory results. In this paper, we present a novel framework to enhance the accessibility of micro blogging messages. The proposed framework improves message representation by mapping messages from an unstructured feature space to a semantically meaningful knowledge space. First, in order to reduce the noise yet keeps the key information as expressed in each message, we propose to use natural language processing (NLP) techniques to analyze the message and extract informative words and phrases. Then, to overcome the extreme sparsity of micro blogging messages, we map the selected terms to structured concepts derived from external knowledge bases that are semantically rich. By conducting feature selection to refine the feature space, we are able to cluster all messages more accurately and generate human-comprehensible labels efficiently from related concepts. It is interesting to explore if integrating social network information can improve the quality of message clustering. The task of cluster labeling was solved without introducing much computational cost.

C. Enriching Short Text Representation in Microblog for Clustering Author: Jiliang Tang , Xufei Wang, Huiji Gao, Xia Hu, Huan

Liu

Social media websites allow users to exchange short texts such as Tweets in microblogging, user status in friendship networks. Their limited length, pervasive abbreviations, and coined acronyms and words exacerbate the problems of synonymy and polysemy, and bring about new challenges to data mining applications such as text clustering and classification, etc. To address these issues, we dissect some potential causes and devise an efficient approach that enriches data representation by employing machine translation to increase the number of features from different languages. Then we propose a novel framework which performs multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques. The proposed approach is evaluated extensively in terms of effectiveness on two social media datasets from Face book and Twitter. With its significant performance improvement, we further investigate potential factors that contribute to the improved performance.

Their limited length, pervasive abbreviations, and coined acronyms and words exacerbate the problems of synonymy and polysemy.

Investigate potential factors that contribute to the improved less performance.

Scalability and inefficiency problems occurred.

III. SYSTEM ANALYSIS

A. Problem Statement

Now a day's rapid growth of social networking services, customers, and other online information increasing day by day. Due to the short-length characteristics of microblog texts, using term frequency vectors to represent microblog texts will cause "sparse data" problem. Finding proper representations of microblog texts is a challenging issue. In the existing system the major problem is scalability and inefficiency problems when processing or analyzing such large-scale data.

B. Existing System

In the last few years, microblogging services such as Twitter and Sina Weibo have gained great popularity among the Internet users. The high volume of textual data produced by the microblogging services is very attractive to the researchers in text mining field. Transforming natural language texts into numerical vectors is an important preprocessing step for many text mining tasks, such as cluster analysis and sentiment classification. The most widely adopted model for text representation is vector space model, where each document in a corpus is represented by a vector with each dimension corresponding to a separate term and the elements denoting the frequencies of the terms. An important issue that needs to be dealt with carefully when using term frequency vectors to represent texts is the "sparse data" problem. Since the number of distinct terms occurring in one document is much smaller than the number of total terms occurring in a corpus, the term frequency vector usually contains a large proportion of zero entries, which may lead to the failures of subsequent mining operations. Considering that microblog texts are very short in length (usually less than 140 characters), it is necessary to explore more proper text representations that can overcome the sparsity problem.

Researchers have proposed many approaches to enhance the representations of short text, such as expanding the original short document by adding semantically related terms or mapping a high-dimensional term frequency vector to a low-dimensional feature vector via latent semantic analysis (LSA). In this paper we mainly focus on how to create a proper low-dimensional feature space for microblog texts, and yet we also investigate how to utilize the expansion approach to learn better features. For convenience, hereafter we use the term tweet to denote the textual message published via microblogging services.

Currently, there are few approaches specifically proposed for dimensionality reduction of tweets. Instead, low dimensional representations of tweets are usually obtained as the by-products of topic modeling. Topic models, which can discover the latent topics that occur in a collection of documents, are widely applied in microblog topic detection. In the view of a topic model, a document is a mixture of topics and each topic corresponds to a probability distribution over terms. The number of topics is usually smaller than the number of terms appearing in a given document collection, and hence the probability distribution over latent topics of one document can be viewed as a low dimensional representation of that document. However, due to the fact that both the number of topics and the content of topics change frequently in microblog environment, the topic-based representations are less applicable to tweets.

1) Disadvantages: Scalability and Inefficiency problems when processing or analyzing large scale data.

Sparse data problem will be occurred in text representation find proper text representation of data.

C. Proposed System

Instead of employing topic models to implicitly find the low-dimensional representations of tweets, in this paper we resort to deep learning approaches to explicitly learn the low-dimensional representations. Deep learning, which is an emerging area in machine

Volume 5 Issue II, March 2017
ISSN: 2321-9653

learning, refers to a class of machine learning techniques where many layers of information processing stages in hierarchical architectures are exploited. Previous studies have shown that text mining tasks, such as classification, can benefit a lot from the low-dimensional representations produced by deep networks. we can make the following two general assumptions.

If one tweet is created by retweeting another tweet, then the two tweets are semantically similar, or at least, related;

If two tweets are labeled with the same hashtag, then they are semantically similar or related.

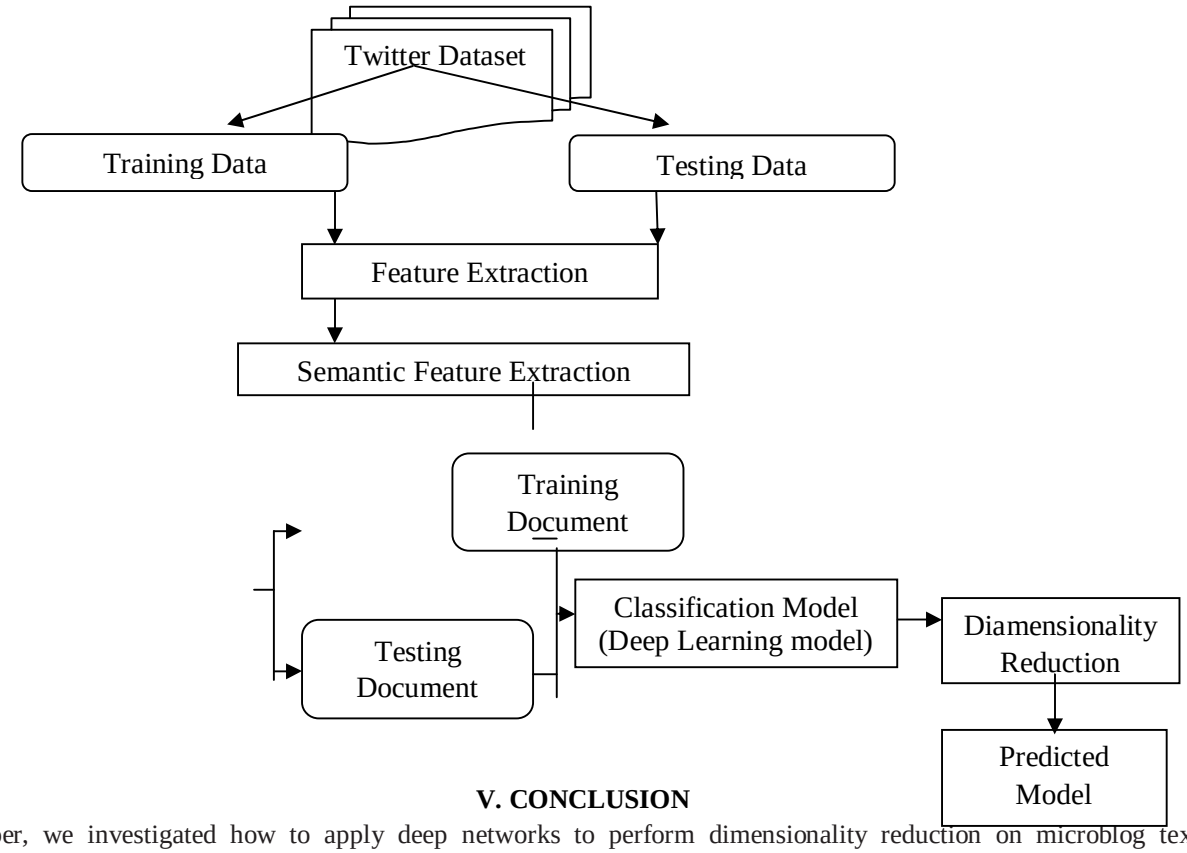
When training the deep network-based models, we make use of such semantic similarity in two ways. First, we utilize the retweet and hashtag information to modify the training set. Then, the two types of information are utilized to define a probability distribution denoting the pair wise similarity of training samples, so that we can employ t-distributed MCML (maximally collapsing metric learning) algorithm to train the model. With the help of microblog-specific information, the models are proved to be more competent in learning representations of tweets than basic deep models. We also demonstrate that the deep models are superior to LDA (latent Dirichlet allocation) and LSA, which are often adopted in current research for obtaining low-dimensional representations of tweets. Besides applying the deep architectures, we also take advantage of vector representations of words to perform dimensionality reduction. By representing each word as a continuous vector, one can explicitly measure the syntactic and semantic word similarities, which is beneficial to natural language processing tasks. In this paper, we utilize word vectors via the following way. When preparing the training set, we expand each tweet by adding words that are most similar to original words appearing in the tweet. The similarity between words are measured based on pre-learned word vectors. After the expansion, we can utilize the retweet and hashtag information to further modify the training set or directly train the deep models. Experiment results show that models trained on expanded tweets can learn better low-dimensional representations.

1) Advantages: Text representation overcame the sparsity problem.

In this approach to learn better fearure representation.

Mapping a high diamensional term frequency vector to low diamensional feature vector efficiently.

IV. PROPOSED SYSTEM ARCHITECTURE



V. CONCLUSION

In this paper, we investigated how to apply deep networks to perform dimensionality reduction on microblog texts. Priori knowledge about semantic similarity, which is derived from retweet relationships and hashtags, was explored to train the deep

networks. Two types of approaches, namely modifying the training data and modifying the objective of fine-tuning, were proposed to utilize such priori knowledge. Experiment results validated that deep models can learn better representations than LSA and LDA, and the use of microblog-specific information can further improve the performance of deep models. To improve the efficiency we implement the system in Hadoop. In addition to that to make services effective. To achieve the scalability and efficiency with help of map reduce framework in a big data environment.

REFERENCE

[1] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.
[2] Y. Xi-Wei, "Feature extension for short text," in Proc. 3rd Int. Symp. Comput. Sci. Comput. Technol., 2010, pp. 338–341.
[3] X. Hu, L. Tang, and H. Liu, "Enhancing accessibility of microblogging messages using semantic knowledge," in Proc. 20th ACM Int. Conf. Inf. Knowl. Manag.,, 2011, pp. 2465–2468.
[4] X. Yan and H. Zhao, "Chinese microblog topic detection based on the latent semantic analysis and structural property," J. Netw., vol. 8, no. 4, pp. 917–923, 2013.
[5] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in Proc. 4th Int. Conf. Weblogs Social Media, 2010, pp. 130–137.
[6] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in Proc. 20th ACM Int. Conf. Inf. Knowl. Manag., 2011, pp. 775–784.
[7] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Sci., vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[8] R. Salakhutdinov and G. Hinton, "Semantic hashing," Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, Jul. 2009.
[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," CoRR, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.378
[10] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Human Language Technol.-Volume 1., 2011, pp. 142–150.
[11] J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu, "Enriching short text representation in microblog for clustering," Frontiers Comput. Sci., vol. 6, no. 1, pp. 88–101, 2012.
[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in Proc. Int. Conf. Neural Information Processing Syst., 2004, pp. 1385– 1392
[13] C. E. Grant, C. P. George, C. Jenneisch, and J. N. Wilson, "Online topic modeling for real-time twitter search," in Proc. Text Retrieval Conf., 2011, pp. 1–9.

AUTHORS PROFILE

Name : Mr.M.Vengateshwaran

Designation : Asst.Professor & Head in CSE
Mass Group of Institutions, Kumbakonam

Qualification: B.E., M.E.,

Specialization: Bigdata, Datamining, IR, Database, SE etc.,

E-mail : mkvengatesh@gmail.com

Name : Mrs.C.RamyaPriyadarsini

Designation : Asst.Professor in CSE
Mass Group of Institutions, Kumbakonam

Qualification: B.E., M.Tech.,

Volume 5 Issue II, March 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

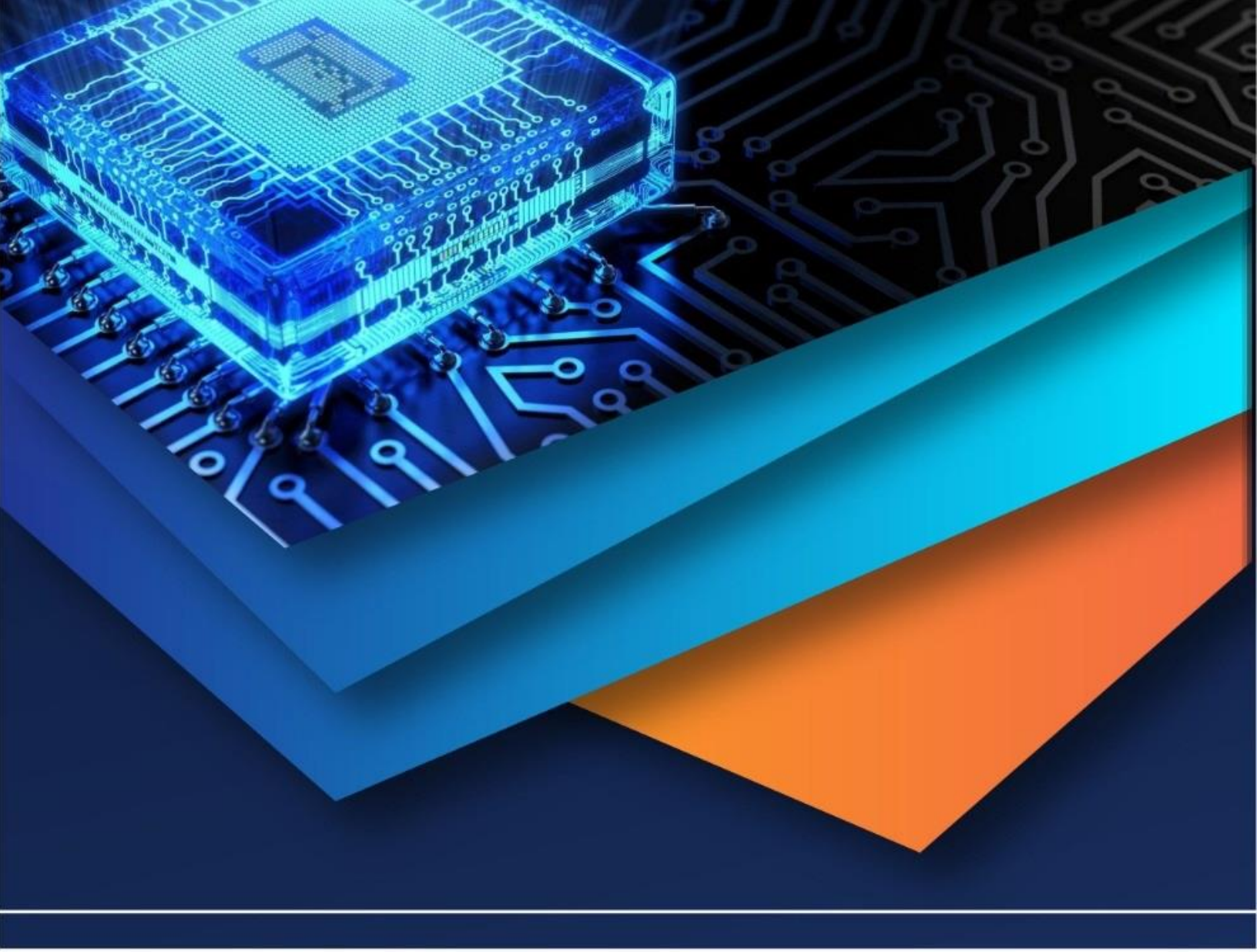
Specialization: Software Engineering, Networks, OOAD , Database

Name : Ms.N.Valarmathi

Designation : Asst.Professor in CSE
Mass Group of Institutions, Kumbakonam

Qualification: B.E., M.Tech.,

Specialization: Datastructures, Complier design, Database etc.,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY