International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-8, June 2020

Effect of Supervised and Unsupervised Algorithm for Cross Domain Sentiment Analysis

Vaishali Arya, Rashmi Agrawal

Abstract: ***Today we are living in the "information age" where data is the capital of the new economy. With the rapidly growing data every day on online portals and social networking websites, today industries are collecting and analyzing more data than before. Though data is readily available but finding valuable insights out of it is a real task. With easy accessibility of the data, new technologies, and a cultural shift towards data-driven decision making drives the need for Sentiment Analysis (SA) and makes it relevant in most of the domains like politics, marketing, healthcare, etc. This rapidly increasing information on different domains has motivated researchers to develop a cross-domain sentiment analysis model. For the development of this model, we have analyzed the performance of supervised and unsupervised models on benchmark datasets for the cross-domain analysis. The models chosen for the supervised is the Support Vector Machine (SVM) and for the unsupervised approach we have used a combination of Vader wherein the testing results showed that the supervised algorithms performed well in comparison to the unsupervised algorithm.***

Keywords: Cross-Domain Sentiment Analysis, Supervised Algorithms for Cross-Domain, SVM for SA, Unsupervised Algorithms for Cross-Domain, Vader SA.

I. INTRODUCTION

Cross-domain sentiment classification is the classification of the target domain using the knowledge from the source domain that is independent of the target domain. This kind of model provides benefits to different organizations to provide the utmost services to their users in analyzing any domain of dataset for enhancing their information and boost their level of service they have been giving [1, 2]. These kinds of models significantly reduce the workload of data annotators in annotating any new domain dataset required for the analysis. One of the most important data platforms for cross-domain analysis is Twitter which is a popular platform for generating the user opinion online where millions of people tweets in a few seconds on viral topics of different domains in real time[3]. This has raised the issue of having the annotated data for the application of supervised algorithms. The data annotation is also a rigorous and time-consuming task and if the dataset has poorly annotated data then it always leads to poor results. Hence this leads researchers to use the lexicon-based approach. A lexicon-based approach is an unsupervised approach that does not require an already annotated dataset.

Revised Manuscript Received on June 30, 2020.
***** Correspondence Author
Vaishali Arya*, Manav Rachna International Institute of Research & Studies, Faridabad, India. arya.vaishali17@gmail.com
Rashmi Agrawal, Manav Rachna International Institute of Research & Studies, Faridabad, India. drrashmiagrawal78@gmail.com

This uses the thesaurus that has the sentiment polarity score of each word in its dictionary. Using that polarity score the cumulative score of sentence is calculated for interpreting the sentiment of the overall sentence. However, it is important to note that an unsupervised approach also suffers in giving the accuracy as per the already done researches[4]. The reason for giving less accuracy is due to the words which are not available in the lexicon dictionary and are contributing to providing the overall sentiment of the sentence. This happens because most data taken for analysis is from twitter and the brevity of Twitter is such that it uses less uniformed text but more slang language [5]. In this study to show the performance effect of supervised and unsupervised approaches, The Support Vector Machine (SVM) model is being used for supervising approach, and for an unsupervised based approach, we have used lexicon-based Vader[6]. In [8], the authors use NB, SVM, k-NN classifiers with appropriate feature selection and reduction schemes for SA of consumer feedback data which shows that linear SVM's attain good accuracy on data that is difficult even for human annotators to analyze. Hence we have adopted the Support Vector Machine (SVM) for the application of a machine learning algorithm for the proposed analysis task. We have utilized the Vader, which is a rule-based lexicon for the better handling of the context of microblogs and its benefits seen in [9–11]. For the comparison of their accuracy F1 score is calculated which is a harmonic mean of precision and recall used for an imbalanced dataset on the benchmark dataset to compare the performance of the models.which is a harmonic mean of precision and recall used for an imbalanced dataset on the benchmark dataset to compare the performance of the models.

II. LITERATURE REVIEW

In this section, we have reviewed the effect of lexicon-based unsupervised and supervised approaches used in sentiment analysis.

a. Lexicon Based:

The lexicon-based approaches are used widely for providing the domain adaptability in the model. The lexicon-based approaches utilize either the already existing lexicon or developed the lexicon to ascertain the sentiment in diverse domains based on the polarity score of each word mentioned lexicon.

In [12] the lexicon-based approach is used for feature summarization of a large number of product reviews. The author in [13] developed a general-purpose lexicon using opinion mining for the sentiment analysis.

This approach extracted the features which are carrying sentiment in the sentence and assign the polarity score for the given target domain in a domain adaptable model. The domain adaptability also uses sensitive sentiment thesaurus [14]using labeled and unlabeled data for both the source and target domains.

Sentiment sensitivity lexicons are created to find the association between words that express similar sentiments in different domains. Authors in [15] propose a lexicon approach for SA containing words as well as phrases. They have used affix patterns to generate further words like imX, disX, Xful, etc that helps in giving more accuracy by learning the lexicon-based approach on these affix rules. In [10], the authors carry out SA of movie reviews using SentiWordNet [11] and a domain-specific lexicon. They compute the polarity class and the strength per aspect.

b. **Supervised Based Approach:**

For sentiment analysis [16] was one of the earliest supervised models for SA. In this, the authors applied the three machine learning algorithms for the SA of movie reviews. They employ Naïve Bayes(NB), MaxEnt, and SVM for the analysis and analyze the factors that make the sentiment classification approach more challenging. The author in [17] describes the disadvantage of using SVM in a cross-domain approach. The author illustrates that the cross-domain models suffer from the labeled data and can not perform well with different sources and targeted datasets. Authors in [18] use the semi-supervised approaches for the sentiment classification. In the proposed approach author has utilized both supervised and unsupervised approaches for the sentiment analysis. Using the lexicon-based unsupervised approach, authors have annotated the dataset and then the SVM is applied for SA. In [19, 20] authors have shown the good results of SA for the same target and source domain datasets and showed that in cross-domain no machine learning(ML) algorithm can work well in individuals. The supervised algorithms can either work with multisource input data for any other given target data[3, 21, 22] or using the hybrid approaches[23, 24].

Contribution and Objective: In the surveyed literature, supervised and unsupervised approaches are discussed for the cross-domain SA. However, in all the researches no one has concluded about the effect of supervised only and unsupervised only approaches in cross-domain SA. They have discussed the usage of a lexicon-based approach which is utilized mostly for annotating the unlabeled datasets and their scope in providing the domain adaptability. They have also discussed the limitation of a supervised algorithm in the development of domain adaptable systems and their efficiency in non-adaptable models. This work aims to find the underlying gap of these approaches in developing a domain adaptable or a cross-domain SA model.

III. EXPERIMENTAL DATASET:

The data set is a multi-domain benchmark dataset taken from amazon reviews for four different product types: books, DVDs, electronics, and kitchen appliances [25]. Reviews contain star ratings (1 to 5 stars) that can be converted into binary labels if required. Reviews are binary classified into a positive and negative category only, and the neutral category is discarded. Reviews with rating “>3” are labeled positive reviews and those with rating “<3” are labeled as negative reviews. From the analysis standpoint, we have chosen four domains i.e. Book, DVD, Electronics, and Health. Each domain has 1000 reviews and each has 500 positive and 500 negative reviews.

IV. FRAMEWORK FOR CROSS-DOMAIN SA:

For the effective analysis of supervised and unsupervised approaches, we have selected the SVM model and Vader lexicon approach respectively. For performing analysis, all the selected multi-domain datasets are passed through the following phases:

- **Preprocessing of Dataset:**

The data set available in the repository are in tar format are extracted and cleaned using the python NLTK, the dataset is tokenized and lemmatized for getting the root of the word that will be used in possessing the sentiment in the text. All the punctuations and stopwords (excepting the negations, and word connectors like but, kind of, very, much) are also cleaned for the analysis as these are extra words only and do not contribute in sentiment. After cleaning the text, the word vector matrix is being created using the TF-IDF approach as shown in figure1.

Table I: Data Integration for Cross-Domain Sentiment Analysis

Annotation	Source Dataset 1	Source Dataset 2	Target Dataset
A1,B1	Book	Kitchen	DVD
A2,B2	Book	Health	DVD
A3,B3	Kitchen	Book	Electronics
A4,B4	Kitchen	Book	Health

- **Supervised SVM Implementation:**

To implement the supervised SVM model, we have applied the multisource source domain dataset to predict the single target source dataset (MSSTD) [25]. The reason for adopting MSSTD is to enhance the labeled dataset for the application of a supervised model on the target domain. This enhanced dataset is used to transfer the features into a model to adapt to the target domain.

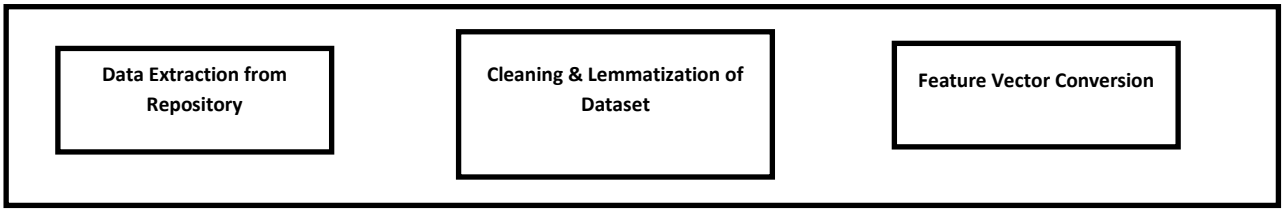

Fig. 1. Preprocessing of Dataset for SA

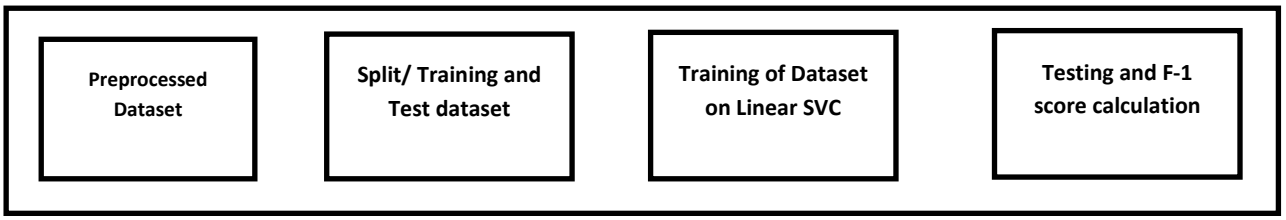

Fig. 2: Supervised SVM for SA

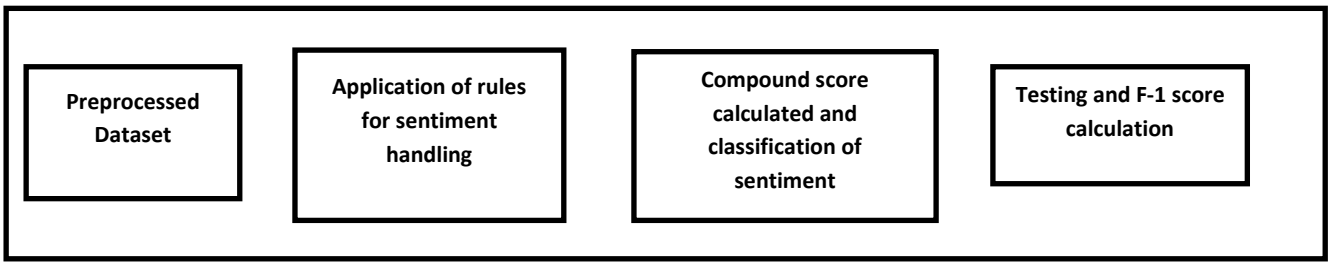

Fig. 3. Unsupervised Lexicon based Approach for SA

The domain adaptability across the different domains required the domain-independent features to be passed into the model so that it can learn on any domain [21, 26]. Hence we have applied the following combinations of domains to achieve the domain adaptability in the model. The datasets are shown in table 1. The Linear Kernel supervised SVM is applied to the four different combinations of datasets {(A1, B1),(A2, B2),(A3, B3),(A4, B4)} where A1-A4 is annotated for supervised and B1-B4 is annotated for Vader moels. The supervised SVM is passed with one gram and 2- gram features vector and the model is trained with 5 fold cross-validation on the datasets. The split ratio selected for the training and testing purpose is a 70:30% ratio. The Model is shown in figure 2.

- **Vader lexicon-based Unsupervised Approach:**

Vader is the lexicon-based model developed for the short text live streaming of the dataset for the classification of SA[6]. This is a rule-based model that extracts the aspect of the sentiment from the sentence based English grammar rules. It effectively applies the rule of negations, sentiment booster, punctuations, and the connectors in giving the sentiment about any review. This is the reason for the preprocessing phase of text data; we have not removed such words from the text. Similarly, we have tested this model on the datasets mentioned in table1. This creates the symmetry between the models to apply the test.

V. RESULT & CONCLUSION:

The cross-domain sentiment analysis model is tested on two frameworks, the Supervised and unsupervised framework. Both models are tuned with the same set of features for the application of domain adaptability in the framework. The approaches are tested on the datasets {(A1, B1),(A2, B2),(A3, B3),(A4, B4)} as shown in table1, where A1 to A4 represent the results for SVM and B1 to B4 represent the results for the unsupervised approach Vader. The results are reported in table 2. Both model's accuracy is measured using the harmonic mean of precision and recall, i.e. F1-score. The experiment carried out in this research investigated that the supervised algorithms outperform the lexicon-based approach. The supervised algorithms required a sufficient annotated dataset for achieving the good results in cross-domain that we provided through multiple sources given dataset for a single target domain and hence achieves good accuracy in the model. Whereas, Vader plays an important role when there is no annotated dataset available for the target domain then in that scenarios only they are relevant.

Table- II: Analysis Table of Supervised and Unsupervised Approach

Book+Kitchen -> DVD(A1)				Book+Health -> DVD(A2)				Kitchen+Book -> Electronics(A3)				Kitchen+Health -> Book(A2)			
SVM	Precis	Recall	F1-Score	***SVM***	Precision	Recall	F1-Score	***SVM***	Precision	Recall	F1-Score	***SVM***	Precision	Recall	F1-Score
Negative	0.79	0.72	0.75	Negative	0.73	0.81	0.77	Negative	0.77	0.8	0.79	Negative	0.76	0.81	0.79
Positive	0.74	0.81	0.77	Positive	0.79	0.69	0.74	Positive	0.79	0.77	0.78	Positive	0.8	0.75	0.77
Accuracy			76.50%	Accuracy			75.20%	Accuracy			78.20%	Accuracy			78.10%
Book+Kitchen -> DVD(B1)				Book+Health -> DVD(B2)				Kitchen+Book -> Electronics(B3)				Kitchen ->Book(B2)			
Vader	Precis	Recall	F1-Score	***Vader***	Precision	Recall	F1-Score	***Vader***	Precision	Recall	F1-Score	***Vader***	Precision	Recall	F1-Score
Negative	0.7	0.39	0.51	Negative	0.79	0.46	0.58	Negative	0.81	0.46	0.59	Negative	0.81	0.39	0.53
Positive	0.58	0.86	0.69	Positive	0.62	0.88	0.73	Positive	0.62	0.9	0.74	Positive	0.6	0.91	0.72
Accuracy			62.20%	Accuracy			67.20%	Accuracy			67.20%	Accuracy			64.90%

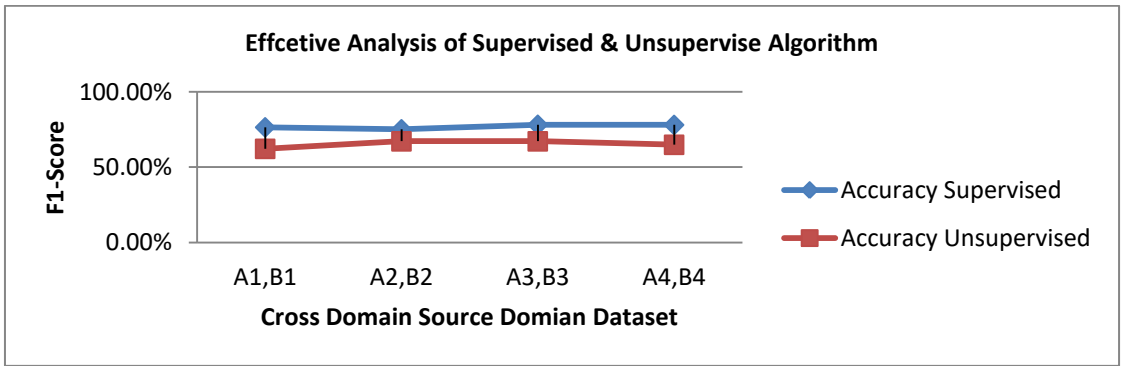

Fig. 4. Performance graph of Supervised and Unsupervised approach

REFRENCES

1. He, Y., Lin, C., Alani, H.: Automatically extracting polarity-bearing topics for cross-domain sentiment classification. ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. 1, 123–131 (2011).
2. Wang, L., Niu, J., Song, H., Atiquzzaman, M.: SentiRelated: A cross-domain sentiment classification algorithm for short texts through sentiment related index. J. Netw. Comput. Appl. 101, 111–119 (2018). https://doi.org/10.1016/j.jnca.2017.11.001.
3. Rane, A., Kumar, A.: Sentiment Classification System of Twitter Data for US Airline Service Analysis. In: Proceedings - International Computer Software and Applications Conference. pp. 769–773. IEEE Computer Society (2018). https://doi.org/10.1109/COMPSAC.2018.00114.
4. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning Word Vectors for Sentiment Analysis.
5. Ghiassi, M., Lee, S.: A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. Expert Syst. Appl. 106, 197–216 (2018). https://doi.org/10.1016/j.eswa.2018.04.006.
6. Hutto, C.J., Gilbert, E.: VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. (2014).
7. Baccianella, S., Esuli, A., Sebastiani, F.: SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010. pp. 2200–2204 (2010).
8. Gamon, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proc. 20th Int. Conf. Comput. Linguist. (2004). https://doi.org/10.3115/1220355.1220476.
9. Yiran, Y., Srivastava, S.: Aspect-based Sentiment Analysis on mobile phone reviews with LDA. ACM Int. Conf. Proceeding Ser. 101–105 (2019). https://doi.org/10.1145/3340997.3341012.
10. Islam, M.R., Zibran, M.F.: SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. J. Syst. Softw. 145, 125–146 (2018). https://doi.org/10.1016/j.jss.2018.08.030.
11. Islam, M.R., Zibran, M.F.: SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. J. Syst. Softw. 145, 125–146 (2018). https://doi.org/10.1016/j.jss.2018.08.030.
12. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004). https://doi.org/10.1145/1014052.1014073.
13. Shamshurin, I.: Extracting Domain-Specific Opinion Words for Sentiment Analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and
14. Lecture Notes in Bioinformatics). pp. 58–68 (2013). https://doi.org/10.1007/978-3-642-37798-3_6.
15. Bollegala, D., Weir, D., Carroll, J.: Cross-domain sentiment classification using a sentiment sensitive thesaurus. IEEE Trans. Knowl. Data Eng. 25, 1719–1731 (2013). https://doi.org/10.1109/TKDE.2012.103.
16. Mohammad, S., Dunne, C., Dorr, B.: Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009 (2009). https://doi.org/10.3115/1699571.1699591.
17. Pang, B., Lee, L., & Vithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. Proc. Inst. Civ. Eng. - Transp. (2019). https://doi.org/10.1680/jtran.18.00094.
18. Liu, M., Song, Y., Zou, H., Zhang, T.: Reinforced Training Data Selection for Domain Adaptation. 1957–1968 (2019). https://doi.org/10.18653/v1/p19-1189.
19. Da Silva, N.F.F., Coletta, L.F.S., Hruschka, E.R., Hruschka, E.R.: Using unsupervised information to improve semi-supervised tweet sentiment classification. Inf. Sci. (Ny). 355–356, 348–365 (2016). https://doi.org/10.1016/j.ins.2016.02.002.
20. Vadivukarassi, M., Puviarasan, N., Aruna, P.: A Comparison of Supervised Machine Learning Approaches for Categorized Tweets. In: Lecture Notes on Data Engineering and Communications Technologies. pp. 422–430. Springer (2019). https://doi.org/10.1007/978-3-030-03146-6_47.
21. M., B., B., V.: Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis. Int. J. Comput. Appl. 146, 26–30 (2016). https://doi.org/10.5120/ijca2016910921.
22. Zhao, C., Wang, S., Li, D.: Multi-source domain adaptation with joint learning for cross-domain sentiment classification. Knowledge-Based Syst. 191, 105254 (2020). https://doi.org/10.1016/j.knosys.2019.105254.
23. Multi Source domain data for sentiment classification. Zainuddin, N., Selamat, A., Ibrahim, R.: Hybrid sentiment classification on twitter aspect-based sentiment analysis. Appl. Intell. 48, 1218–1232 (2018). https://doi.org/10.1007/s10489-017-1098-6.

24. Appel, O., Chiclana, F., Carter, J., Fujita, H.: Successes and challenges in developing a hybrid approach to sentiment analysis. Appl. Intell. 48, 1176–1188 (2018). https://doi.org/10.1007/s10489-017-0966-4.
25. Hassan, F., Usman, K., Saba, Q.: Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach. Soft Comput. (2018). https://doi.org/10.1007/s00500-018-3187-9.
26. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10. pp. 751–760 (2010). https://doi.org/10.1145/1772690.1772767.
27. Thet, T.T., Na, J.C. and Khoo, C.S., 2010. Aspect-basedentiment analysis of movie reviews on discussion boards. Journal of information science, 36(6), pp.823-848.

AUTHORS PROFILE

Vaishali Kalra is currently working as an Assistant Professor in Department of CSE & IT, NorthCap University, India. She has more than 10 years of experience in teaching as an Assistant Professor. She has completed her B.Tech and M.Tech from Mahrishi Dayanand University Rohtak, Haryana. She is associated with various professional bodies in different capacity, member IEEE, Chapter Advisor of IEEE-IAS, PES. Currently, she is pursuing PhD from MRIIRS Faridabad in the field of Text Mining. She has several publications to her credit in various leading International and National Journals/Conferences in the various areas like Sentiment Analysis, Cross Domain Modelling for Sentiment Analysis, Natural Language Processing, Hybrid models for classification, Neuro-Fuzzy Models for classification, IOT based Healthcare.

Dr. Rashmi Agrawal is PhD and UGC-NET qualified with 18+ years of experience in teaching and research, working as Professor in Department of Computer Applications, Manav Rachna International Institute of Research and Studies, Faridabad, India. She is associated with various professional bodies in different capacity, life member of Computer Society of India, senior member IEEE, ACM CSTA and senior member of Science and Engineering Institute (SCIEI). She is book series editor of Innovations in Big Data and Machine Learning, CRC Press, Taylor and Francis group, USA. She has authored/ coauthored many research papers in peer reviewed national/international journals and conferences which are SCI/SCIE/ESCI/SCOPUS indexed. She has also edited/authored books with national/international publishers (Springer, Elsevier, IGI Global, Apple Academic Press, and CRC Press) and contributed chapters in books edited by Springer, IGI global, Elsevier and CRC Press. Currently she is guiding PhD scholars in Sentiment Analysis, Educational Data Mining, Internet of Things, Brain Computer Interface, Web Service Architecture and Natural language Processing. She is also an active reviewer and editorial board member in various journals of repute.