IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

A Survey on Cross - Domain Opinion Mining

V. Manimekalai [1*], **S. Gomathi @ Rohini** [2]

[1] Department of Computer Technology, Dr.N.G.P Arts & Science College, Coimbatore, India
[2] Department of Computer Application, Sri Ramakrishna College of Arts & Science, Coimbatore, India

[*]*Corresponding Author: manimekalaiv@drngpasc.ac.in, Tel.: +0422-2369259*

Available online at: www.ijcseonline.org

Accepted: 16/Oct/2018, Published: 31/Oct/2018

***Abstract*—** The social network growth is increased and the interest of people in analyzing reviews and opinions for products before buy them. In this regarding research communities, academia, public and service industries are working rigorously on sentiment analysis, also known as, opinion mining, to extract and analyze public mood and views. Analyzing the sentiments in massive user-generated online data, such as product reviews and micro blogs, has become a hot research topic. Sentiment analysis is widely known as a domain dependent problem. In this paper presents a rigorous survey on cross domain sentiment analysis, challenges for social media, then identified problems in different domains usually have different sentiment expressions and a general sentiment classifier is not suitable for all domains. The main problem is the selection of sentiment from huge volume of opinionated data for different kinds of event which is available in the social networks, but there exist a huge difficulty in predicting the accurate outcome of the event at cross domain. A natural solution to this problem is to train a domain-specific sentiment classifier for each target domain. However, the labeled data in target domain is usually insufficient, and it is costly and time-consuming to annotate enough samples.

***Keywords*—**Data Mining, Opinion Mining, Machine Learning, Cross - Domain Sentiment Analysis, SentiWordNet.

I. Introduction

One of the main challenges of opinion mining is that subjective expressions vary profoundly, depending on the domain. The exact same word or phrase may or may not consider opinionated in different contexts. For example “short battery life” is clearly negative opinion, and “short article” is simply literature genre. There is a current trend to focus only on machine learning techniques as a workaround for this problem, entirely dismissing the underlying linguistic structure, but we strongly believe it is essential to take it into account as well.

It is easy to see that there are properties that subjective sentences share across domains are, Syntactic structure of subjective expressions is similar and domain-independent Some subjective words and expression are domain-independent (“fine”, “terrific”, “Happy”). Opinionated sentences usually contain more than one subjective expression and often occur next to each other in texts. Those properties allow us to find opinionated sentences in the text, using domain- independent words, as pivots and then to extract domain-specific expressions from them, based on the expected syntactic structure of opinion. Those expressions can be transformed into extraction patterns and used in mining algorithms.

Table 1. Examples for Polarity

	Electronics	**Video games**
+	**Compact**; easy to operate; very good picture quality; looks **sharp**!	A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
+	I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**	Very **realistic** shooting action and good plots. We played this and were **hooked**
-	It is also quite **blurry** in very dark settings. I will never buy HP again.	The game is so **boring**. I am extremely unhappy and will probably never buy UbiSoft again.

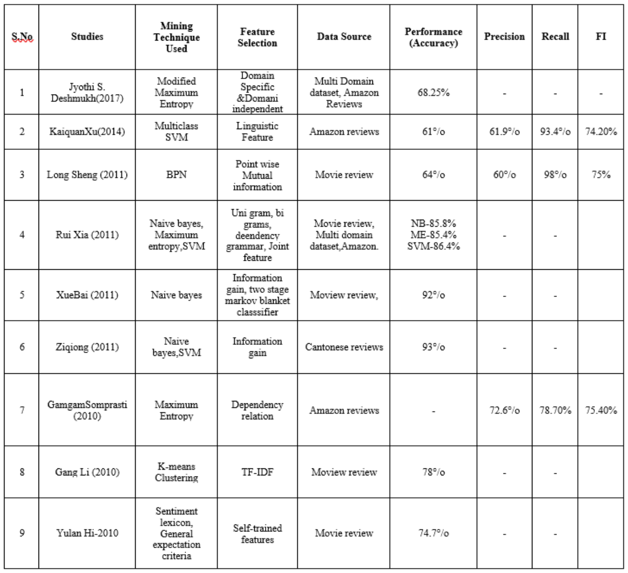
Therefore it allows us to extract a lexicon of opinionated expressions for such diverse areas as, for example, Political news, reviews on products. It is important that no manual tagging of the processed texts is required, which minimizes the need for human participation. Table 1 shows the Cross domain sentiment classification. For example, reviews of electronics and video games products. Boldfaces are domain specific words, which are much more frequent in one domain than in the other one. Italic words are some domain-independent words, which occur frequently in both domains. “+” denotes positive sentiment, and “-” denotes negative sentiment.

The paper is organized as follows, Section I contains the introduction of opinion mining, Section II contain the literature review, Section III contain cross domain sentiment classification, Section IV contain techniques for cross domain analysis, section V concludes research work with future directions.

II. LITERATURE REVIEW

This section describes literature review or the studies which give an idea that for our research done in direction of sentiment classification. Danushka Bollegala, John Yannis Goulermas,[12] proposed a method to adapt an existing sentimental extraction system to extract new features with minimum supervision for opinion identification. A supervised relation extraction system that is trained to extract a particular relation type (source relation) might not accurately extract a new type of a relation (target relation) for which it has not been trained. However, it is costly to create training data manually for every new intent expression type that one might want to extract. In this paper they mainly concentrate on source and target relation which is best to eacch other and also found the polarity. Polarity mainly focused on positive and negative and calculated using precision and performance method. Also, features from different domains were discovered by modeling their correlations with pivot features. It should be noted that the pivot feature is particularly useful in semi-supervised machine learning. Table 2 shows the details of Domain Dependent & Domain Independent feature selections.

Table 2. Survey Details of Domain dependent and Domain Independent



S.No	Studies	Mining Technique Used	Feature Selection	Data Source	Performance (Accuracy)	Precision	Recall	FI
1	Jyothi S. Deshmukh(2017)	Modified Maximum Entropy	Domain Specific &Domani independent	Multi Domain dataset, Amazon Reviews	68.25%	-	-	-
2	KaiquanXu(2014)	Multiclass SVM	Linguistic Feature	Amazon reviews	61°/o	61.9°/o	93.4°/o	74.20%
3	Long Sheng (2011)	BPN	Point wise Mutual information	Movie review	64°/o	60°/o	98°/o	75%
4	Rui Xia (2011)	Naive bayes, Maximum entropy,SVM	Uni gram, bi grams, deendency grammar, Joint feature	Movie review, Multi domain dataset,Amazon.	NB-85.8% ME-85.4% SVM-86.4%	-	-	
5	XueBai (2011)	Naive bayes	Information gain, two stage markov blanket classsifier	Moview review,	92°/o	-	-	
6	Ziqiong (2011)	Naive bayes,SVM	Information gain	Cantonese reviews	93°/o	-	-	
7	GamgamSomprasti (2010)	Maximum Entropy	Dependency relation	Amazon reviews	-	72.6°/o	78.70%	75.40%
8	Gang Li (2010)	K-means Clustering	TF-IDF	Moview review	78°/o	-	-	
9	Yulan Hi-2010	Sentiment lexicon, General expectation criteria	Self-trained features	Movie review	74.7°/o	-	-	

III. CROSS-DOMAIN SENTIMENT CLASSIFICATION

Cross-domain sentiment analysis is an interesting research problem. Due to high variation of subjectivity across domains, it is a challenging task. Cross-domain requires at least two domains: source domain on which a classifier is to be trained on, and target domain on which testing is to be performed. Work carried out in this area can be classified into two groups; the first group requires initial training set from source domain as well as from target domain [28]. The learners in the second group of study are trained on source domain and tested on target domain. These studies were carried out using lexicon based [32], machine learning based [28], and hybrid approaches [22].

Tan et al. [28] applied supervised learning approach for cross-domain sentiment classification. They proposed an effective measure i.e. Frequently Co-occurring Entropy (FCE), to select generalizable features that occur frequently in both old-domain data and the unlabelled new-domain data. To train a classification model for the new domain they employed weighted expectation–maximization based Adapted Naïve Bayes (ANB) . For the experiment, education reviews (1012 ve and 254 +ve), stock reviews (683 ve and 364 +ve) and computer reviews (390 ve and 544 +ve) were considered. FCE can be replaced with other techniques to pick out better features. Weichselbraun et al. Disambiguation and contextualization have been considered to get better result on cross-domain SA.
The Ontological lexicon method used to create in three parts

- ambiguous term detection
- calculating sentiment score based on the probabilities of co-occurring contextual terms, and finally
- SA by combining polarity values for unambiguous and ambiguous terms.

Evaluation was performed on 2500 reviews from Amazon.com, 1800 hotel reviews from TripAdvisor.com and Pang and Lee [32] movie review dataset and yielded precision of 76.5%, 82%, 79% respectively. Extension of the proposed lexicon was required by using grounded concepts from SenticNet[25], ConceptNet [120], Freebase [29], DBPedia[30], etc.

Bollegala et al. [22] proposed automatic sentiment sensitive thesaurus creation and feature expansion for cross-domain sentiment classification. Novelty of the proposed method was the exploitation of the created thesaurus to expand feature vectors at training and testing part of a binary classifier. Sentiment score was calculated based on PMI between a sentiment element and feature vectors. Experiments were performed on the Blitzer et al. [31] dataset and yielded average accuracy of 80.9%. Each lexicon was utilized to develop the labeled feature vectors from the source domains and train an L1 regularized logistic regression-based binary classifier (Classias17). These four thesauri were compared against three baseline methods and yielded better performance over all methods. Future work needs to ensure wide applicability in other domains as well. Cho et al. [17] suggested cross-domain sentiment classification by integrating multiple sentiment dictionaries viz. WordNet-Affect [16], SentiWordNet [5, 6], WordNet [133], Opinion Lexicon (OL) [18], AFINN [19], SO-CAL [20], Subjectivity Lexicon [23], General Inquirer [24], SenticNet [25], SentiSense [26], and Micro-WNOp [27] together. They proposed new concepts to remove and/or shift polarity of a sentiment word. For the experiments, 17,500, 35,000, and 90,000 reviews were collected from Amazon.com for smart phones, movies, and books to build a positive/negative review dataset.

They achieved 82.6%, 80.1%, and 81.8% accuracies for smart phone, movie, and book reviews. The proposed approach can be applied to create a custom dictionary, which may yield promising result for cross-domain sentiment classification.

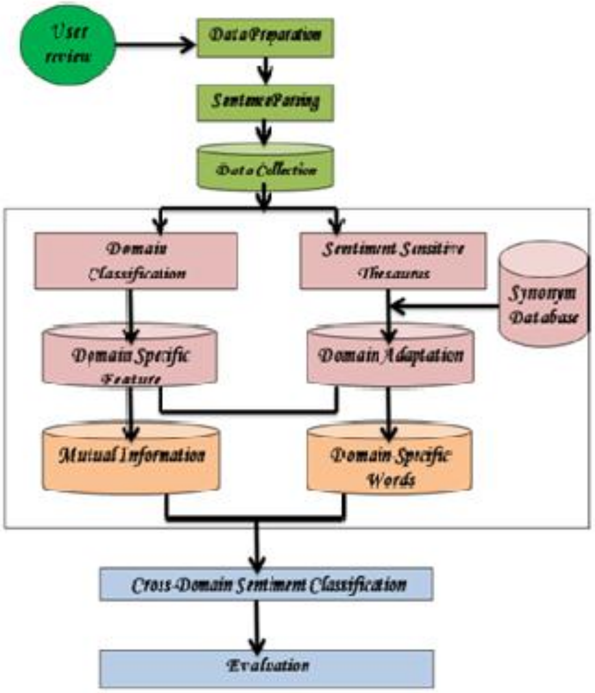


Figure 1. Basic concepts of Cross – Domain Sentiment Classification

Figure 2.

IV. Techniques for Cross – Domain Analysis

There are several techniques have been used for cross-domain sentiment analysis. In this section, we attempt to group these techniques according to a) the type learning model used, for example, topic learning and deep learning; b) feature based reasoning and case based reasoning (CBR) c the resources used by the techniques, which mainly depends on whether they are automatically or manually developed (e.g. the lexicon, such as the sensitive-sentiment thesaurus and the meta-combination of lexical-enhanced classifiers). Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Spectral Feature Alignment (SFA) Algorithm

The SFA algorithm was proposed by [1] as a way to align the words from a range of domains into a unified cluster for a specific domain by utilizing knowledge on domain-independent words. The SFA algorithm does several things: it identifies the source and target domains, recognizes the domain-independent and domain-specific features, constructs a bipartite graph, and performs co-clusteringand alignment to achieve successful sentiment classification.Pan et al. (2010) first constructed a bipartite graph to modelthe co-occurrence between domain-specific and domain-independent words into a set of clusters, with the aim ofreducing any mismatch between the domain-specific words of the source and target domains. This cluster was then utilized to train a classifier for sentiment classification. The proposed SFA algorithm collates the domain-specific words originating from the source and target domains into expressive groups, and the domain-independent words are utilized as a channel to assist in this process. In this way the distance between the domain-specific words of the two domains is reduced. The algorithm is also used to train the sentiment classifier in the target domain. Later, [2] tested an ensemble algorithm consisting of a Support Vector Machine (SVM) and the SFA algorithm on an Amazon dataset. The authors enhanced SFA by the addition of words in shorthand notation and in n-gram form. Next, a SVM-based binary classifier was trained on positive and negative examples of customer reviews. The model takes the tree information and the similarity between domains into account during sentiment classification. In addition, the closest related models in terms of target node, model weight, and model application are selected by using two strategies, the cosine function and a taxonomy-based regression model.

B. Structured Correspondence Learning (SCL) Algorithm

The SCL algorithm was proposed by [3] as a method for learning the features of a variety of domains. Essentially, the algorithm adapts the source domain to the target domain. The authors state in their work, which focuses on developing an effective binary classifier, that: ‘‘A domain is a pair consisting of a distribution D on X and a labeling function f : X [0,1]’’ [03]. To develop the model, they measured the distance between two distributions, one in the source and one in the target domain, by using hypothesized distance measures based on divergence. Also, features from different domains were discovered by modeling their correlations with pivot features. It should be noted that the pivot feature is particularly useful in semi-supervised machine learning. In the SCL algorithm, the non-pivot features are correlated with similar pivot features. Then a discriminative learner is used in training the classifier. In the SCL methodology, labeled training data is not used in testing, so it is important to be able to model and utilize correlations between features in different domains. The work in [03] was extended the following year by [04], who proposed a new model named the Structured Correspondence Learning-Mutual Information or SCI-MI model. The extension was necessary because SCL depends on the choice of pivot features, and if they are not well-chosen this can adversely affect performance.

To address this issue, in SCL-MI [04], the top pivot features are selected by using the mutual information between a feature (unigram or bigram) and a domain label. After the necessary features have been selected, the binary classifier is trained by the SCL algorithm. More recently, [05] drew on the concept of SCL to develop a method that utilizes two auxiliary tasks to help induce sentence embedding because the authors expected the embedding technique to perform well across the sentiment classification domains.

C. Joint sentiment-topic (JST) model

The JST model proposed by [46] is a probabilistic model-ing framework based on Latent Dirichlet Allocation (LDA). Several machine learning approaches for sentiment classification require labeled corpora in order to train the classifier. In contrast, the JST model is totally unsupervised. The JSTmodel is based on the authors' research on polarity-bearing topics, which they used to enhance the original feature space. Learning in the JST model is based on prior information about the domain-independent polarity words. The JST model is an extension of the LDA model proposed by [06] that wasdeveloped to detect a sentiment and a topic simultaneously from text. In the JST model, discriminative classifiers are used to search for a decision boundary that maximizes a certain measure of separation between classes. The posterior distribution is calculated by applying sequential sampling for each variable (known as Gibbs sampling). The JST method can cluster different terms that exhibit a similar sentiment. Information gain criteria are used to augment and select the features for cross-domain classification.

V. Conclusion

In this paper, we present a short survey on various techniques used for cross-domain Opinion mining. We have mainly focused on three techniques that is sentiment sensitive thesaurus, spectral feature alignment, structural correspondence learning. All these three techniques are different from one other in the way of expanding the feature vector, measuring the relatedness among the words, and finally the classier used for classification. Some methods used for performing cross-domain classification uses labeled or unlabeled data or some uses both. The solution for domain dependent problem is to train a domain-specific sentiment classifier for each target domain and then mapping the polarities with all other cross domains.

Acknowledgment

I would like to express gratitude to my supervisor Dr. S.Gomathi @ Rohini, Management of Dr.NGP Educational Institutions and Principal Dr.N.G.P. College of Arts and Science, Coimbatore for the valuable guidance and precious support.

References

[1] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, ‘‘Cross-domain sentiment classification via spectral feature alignment,’’ in Proc. 19th Int. Conf. World.Wide Web, vol. 10.2010, p. 751.

[2] C. Lin, Y. Lee, C. Yu, and H. Chen, ‘‘Exploring ensemble of models in taxonomy-based cross-domain sentiment classification,’’ in Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.-(CIKM), 2014, pp. 1279 –1288.

[3] J. Blitzer, R. McDonald, and F. Pereira, ‘‘Domain adaptation with structural correspondence learning,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2006, pp. 120–128.

[4] J. Blitzer, M. Dredze, and F. Pereira, ‘‘Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,’’ in Proc. ACL, vol. 7. 2007, pp. 440–447.

[5] J. Yu and J. Jiang, ‘‘Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 236–246.

[6] D. M. Blei and M. I. Jordan, ‘‘Modeling annotated data,’’ in Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2003, pp. 127–134.

[7] Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.Association for Computational Linguistics, 2002.

[8] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics.Association for Computational Linguistics, 2002.

[9] Yan Dang; Yulei Zhang; Hsinchun Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews, Intelligent Systems, IEEE , vol.25, no.4, pp.46,53, July-Aug.2010 doi: 10.1109/MIS.2009.105

[10] Hung, Chihli, and Hao-Kai Lin. "Using objective words in SentiWordNet to improve word-of-mouth sentiment classification." IEEE Intelligent Systems 28.2 (2013): 0047-54.

[11] Bhaskar, J.; Sruthi, K.; Nedungadi, P., "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers," Recent Advances and Innovations in Engineering (ICRAIE), 2014 , vol., no., pp.1,6, 9-11 May 2014 doi: 10.1109/ICRAIE.2014.6909220.

[12] Muhammad faheem Khan, Aurangzeb and khairullah khan efficient word sense disambigutionteqnique for sentence level sentiment classification of online review‖ Sci.Int(Lahore).25(4),2013.

[13] Bollegala, D.; Weir, D.; Carroll, J., "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," Knowledge and Data Engineering, IEEE Transactions on , vol.25, no.8, pp.1719,1731, Aug. 2013 doi: 10.1109/TKDE.2012.103

[14] A. Esuli, F. Sebastiani, SENTIWORDNET: a publicly available lexical resource for opinion mining, in: Proceedings of the 5th Conference on Language Resources and Evaluation LREC-06, Genoa, Italy, 2006, pp. 417–422.

[15] S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of LREC10, Malta, 2010, pp. 2200–2204.

[16] A. Moreo, M. Romero, J.L. Castro, J.M. Zurita, Lexicon-based commentsoriented news sentiment analyzer system, Expert Syst. Appl. 39 (2012) 9166– 9180.

[17] H. Cho et al., Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews, Knowl.-Based Syst. 71 (2014) 61–71.

[18] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 168–177.

[19] F.Å. Nielsen, A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs, 2011.

[20] M. Taboada, J. Brooke, M. Tofiloski, Lexicon-based methods for sentiment analysis, Comput.Linguist. 37 (2) (2011) 267–307.

[21] A. Weichselbraun, S. Gindl, A. Scharl, Extracting and grounding contextualized sentiment lexicons, IEEE Intell. Syst. 28 (2) (2013) 39–46.

[22] D. Bollegala, D. Weir, J. Carroll, Cross-domain sentiment classification using a sentiment sensitive thesaurus, IEEE Trans. Knowl. Data Eng. 25 (8) (2013).

[23] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), 2003, pp. 105–112.

[24] P.J. Stone, E.B. Hunt, A computer approach to content analysis: studies using the general inquirer system, in: Proceedings of the Spring Joint Computer Conference (AFIPS 1963), 1963, pp. 241–256.

[25] E. Cambria, R. Speer, C. Havasi, A. Hussain, SenticNet: a publicly available semantic resource for opinion mining, in: AAAI Fall Symposium: Commonsense Knowledge, vol. 10, p. 02, 2010.

[26] J.C. de Albornoz, L. Plaza, P. Gervas, Sentisense: an easily scalable conceptbased affective lexicon for sentiment analysis, in: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 2012, pp. 3562–3567.

[27] S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, C. Gandini, Micro- 862 WNOp: a gold standard for the evaluation of automatically compiled lexical resources for opinion mining, in: A. Sanso (Ed.), Language Resources and Linguistic Theory, Franco Angeli, 2007, pp. 200–210.

[28] S. Tan, X. Cheng, Y. Wang, H. Xu, Adapting naive bayes to domain adaptation for sentiment analysis, in: M. Boughanem et al. (Eds.), ECIR 2009, LNCS 5478, 2009, pp. 337–349.

[29] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, FreeBase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, 2008, pp. 1247–1250.

[30] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia-A crystallization point for the web of data, Web Semant.: Sci., Serv. Agents World Wide Web 7 (3) (2009) 154–165.

[31] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL'07, vol. 7, 2007, pp. 187–205 (13, 29).

[32] B. Pang, L. Lee, A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, July 2004, p. 271.

[33] Z.Yan et al, EXPRS: an extended page rank method for product feature extraction from online consumer reviews, Inform Manage (2015), http://dx.doi.org/10.1016/j.imm.2015.02.002

[34] A.Hendari, Mhd. A.Tavakoli, N.Salim, Z.Hendari, Detection of Review Spam:aSurvvey, Expert Syst. Appl. 42 (7) (2015) 3634 – 3642.

[35] JyotiS.Deshmukh a, Amiya Kumar Tripathy, Entropy based Classifier for Cross – Domain Opinion Mining, Applied Computing and Informatics 144 (2017) 55-64.

Authors Profile

Mrs.V.Manimekalai pursued Bachelor of Science from Bharathidhasan University in 2008 and Master of Computer Applications from Anna university in 2011 and Master of philosophy from Bharathidhasan University in 2003. She is currently pursuing Ph.D and working as an Assistant professor in Department of Computer Technology in Dr.N.G.P Arts and Science College, Coimbatore. She has published more than 5 research papers in reputed international journals. Her main research work focuses on Data mining, Sentiment analysis, Opinion mining, Big Data Analytics. she has 8 years of teaching experience.

Dr. S. Gomathi @ Rohini owns 13 years of academic experience and 3 years of industrial experience. Image processing and analytics are her fields of interest. She has more than 20 publications in her credit. She is currently serving as Associate Professor Computer Science in Sri Ramakrishna College of Arts & Science, Coimbatore. She is supervising a few Ph.D. / M.Phil. projects. She is a member in IAPA and IEDRC. She has organised a few funded workshops.