Semi-Supervised Non-Linear Dimensionality Reduction Technique for Sentiment Analysis Classification

M. Anandapriya, M. S. Gowtham, Kamalraj Subramaniam

Abstract*: With the quick development in data advances, client created substance, for example, reviews, ratings, recommendations can be advantageously posted on the web, which have powered enthusiasm for sentiment classification. The quantity of records accessible on both online and offline is expanding drastically. Sentiment Classification has a wide scope of utilizations in review related sites. In this paper, we present our investigations about some exploration paper in this field and exhibited our plan to distinguish the sentiment extremity of a given content as positive or negative by lessening the documents dimension, through utilizing semi-supervised non-linear dimensionality decrease technique. For Sentiment Classification, Random Subspace strategy is utilized. For exploratory assessment, openly accessible sentiment datasets can be utilized to check the adequacy of the proposed technique.

***Index Terms*: Sentiment Classification, Random Subspace, Laplacian Eigen Map, Semi-supervised Non-Linear Dimensionality Reduction**

I. INTRODUCTION

A lot of subjective web substance mirrors the people groups' sentiments on pretty much a wide range of products and services. It is the human attitude to know others supposition for the products/services they wanted to expend. What's more, it is turning into the best practice for the makers to monitor their client feelings on their items to improve their consumer loyalty. In any case, the component of those sentiments/reviews increases and bigger, it is hard for the client to foresee the dominant part supposition for the items/services. The sentiment analysis is a developing piece of literary information examination for programmed extraction of abstract substance and anticipating its subjectivity, for example, positive or negative. For dissecting the high dimensional content information, the content information representation device is utilized generally. The tool changes over the information into two dimensional or three-dimensional spaces by applying reduction procedures which compare to the first element of the informational collection [2]. There will be numerous dimensionalities reducing techniques connected to this present reality information sets [3,4,5]. The subjectivity of records can be gotten by human comment yet it should be prepared individuals to decide the real sentiment of reports. Naming the whole reports for sentiment investigation makes the tedious procedure and exorbitant as well. This makes the piece of whole archives is named by human explanation and the rest of the records are unlabeled. Hence, the methodology of semi-supervised learning is connected to content informational collections with both the marked and unlabeled data, which makes progressively practical. Sentiment Classification should be possible utilizing outfit strategies, which increment the precision rate by joining the forecasts of base students [6]. The speculation capacity of an outfit strategy is more alluring than a solitary learner. In this paper, for decreasing the dimension of content information we proposed a semi-supervised non-linear dimensionality reduction strategy, which receives label data of a piece of whole information in ascertaining low-dimensional directions for better reduction result. What's more, for classification, an ensemble technique is utilized. The classifiers utilized are random subspace, bagging and boosting. The base learners can be utilized are SVM, k-NN and Decision Tree.

II. RELATED WORKS

Sentiment Investigation used to recognize and remove the subjectivity substance from content archives and it is one of the utilization of computational linguistics, natural language processing. Regarding the primary point, Sentiment Analysis decides the passionate data of commentators and consequently identifying the extremity of archives whether it is a negative or positive mentality. Generally utilized methodologies for extremity location in Sentiment Analysis are computational linguistics methodology and AI approach [7]. The previous methodology is dependent on the semantic direction and extremity of individual words or expressions. In view of the normal semantic direction of the expressions removed from the review, the whole report is consequently characterized by computing the score dependent on the events of words [8,9]. Utilizing this methodology, we can accept that positive words exist with higher likelihood than negative words in records with positive sentiment. In any case, it takes little calculation time and cost for the extremity of words assurance. The second methodology in sentiment analysis is utilizing AI classifiers. In AI society, assurance of positive or negative extremity is seen as a paired classification issue. The classification strategies, for

Revised Manuscript Received on July 08, 2019.

M. Anandapriya, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India.

M. S. Gowtham, Department of Electronics and Communication Engineering, Karpagam Institute of Technology, Coimbatore, India.

Dr. Kamalraj Subramaniam, Department of Electronics and Communication Engineering, Faculty of Engineering, Karpagam Academy of Higher Education, Coimbatore, India

example, Naïve Bayes [10,11], support vector machine [12,13,14], and Bayesian networks can be utilized. Machine Learning Techniques [15,16] shows better expectation capacities since they utilize named information tests for processing. Ongoing investigations [17,18] recommend that ensemble learning techniques have incredible potential for sentiment classification. Wilson et al utilized Boosting for sentiment classification, which results in 96% upgrades in precision. Abbasi et al. proposed the support vector regression correlation ensemble technique for improving the performance of sentiment classification. Whitehead and Yaeger analyzed Bagging, Boosting and Random Subspace for characterizing the sentimental information and had demonstrated that ensemble learning techniques increment classification exactness in the space of sentiment classification.

III. PROPOSED SYSTEM

In the proposed framework, the sentimental information can be diminished by utilizing the methodology of the semi-supervised non-linear dimensionality reduction technique. This segment depicts the framework design. It incorporates multi-domain datasets, dimensionality reduction and sentiment classification.

A. Multi-domain Sentiment Data Set

A multi-domain sentiment dataset can be utilized for assessing the proposed framework. It incorporates item reviews from various item types taken from Amazon.com. Various items, for example, Book, DVD, Hardware and Kitchen have 1000 positive and 1000 negative reviews. These reviews are in the pseudo-XML scheme.

B. Dimensionality Reduction

For setting up the information to be handled, changing over every one of the characters to lowercase and evacuating abbreviation, stop words and after that apply to stemming. From the preprocessed content reports, the sentimental words are separated utilizing SentiWordNet 3.0. At that point, term frequency inversed document frequency matrix can be developed utilizing the separated words. At that point, apply Principal Component Analysis(PCA) for lessening the component of the informational collection, which will change them into 1000 dimensional spaces, for example, t-SNE which evacuates the correlation between features. At that point, Semi-Supervised Laplacian Eigenmap can be connected to the informational collections. Semi-supervised Laplacian Eigenmap(SS-LE) works under the premises that likeness between information tests with a similar name ought to be set as a bigger incentive than the closeness between distinctively marked examples. In the proposed SS-LE calculation, x is neighborhood set of labeled data points and is neighborhood set of points nearby a labeled data point y of same label. denotes the neighborhood set of labeled data points nearby a labeled data point y of different label. is the average pairwise distances between each data point and its neighbors and is the parameter to control the measure of the effect of mark data on similarity. The algorithm for Semi-Supervised Laplacian Eigenmap(SS-LE) is shown below:

Algorithm: Semi-Supervised Laplacian Eigenmap

1. Construct the graph $Gu(V,E)$ by utilizing input features without label information

- If $y_i \in N_u(y_j)$ or $y_j \in N_u(y_i)$, two nodes are connected by an edge.

- If two nodes, i and j, are connected, the weight of edge is put as follows:

$$wu,ij = \exp\left\{-\frac{\| yi - yj \|^2}{\sigma^2}\right\}$$

2. Construct the graph $Gl(V,E)$ by utilizing label information

-If $y_i \in N_{l+}(y_j)$ or $y_j \in N_{l+}(y_i)$ or $y_i \in N_{l0}(y_j)$ or $y_j \in N_{l0}(y_i)$,two nodes are connected by an edge

-If $y_i \in N_{l+}(y_j)$ or $y_j \in N_{l+}(y_i)$, the weight of edge is 1

-If $y_i \in N_{l0}(y_j)$ or $y_j \in N_{l0}(y_i)$, the weight of edge is 0.5

3. From two graphs, $G_u(V,E)$ and $G_l(V,E)$, graph Laplacian matrices are calculated separately

- Compute degree matrix D, where $d_i = \sum jwij$

- Compute graph Laplacian by **L= D – W**

4. Dimensionality reduction process

- Two graph Laplacian matrices, L_u and L_l,are combined with the parameter

$$L = (1-\lambda)L_u + \lambda L_l$$

- Compute eigenvalues and eigenvectors for the generalized eigenvector problem $N_u(y_i), N_{l_0}(y_i)$

$$Lz = \mu Dz$$

- Let z_0, z_1, z_d be the solutions of the generalized eigenvector problem, ordered according to increasing $Z = [z_1, ..., z_d]$ eigenvalues (z_0 has the smallest eigenvalue).

The low-dimensional information lattice relating to x is as per the following:

- This calculation develops two charts: one diagram incorporates every one of the information focuses on unlabeled and marked information however it utilizes just information highlights without name data. Henceforth this mirrors the nearby geometric properties of information. The subsequent diagram incorporates mark data and assigning weight data for the hubs to such an extent that the equivalent named hub is weighted as 1 and unlabeled hub is weighted as 0.5.

C. Sentiment Classification

The Random Subspace is an ensemble development system, in which the preparation informational index is changed as in bagging. The alteration of preparing informational collection is just in the feature space and not in the instance space. Random Subspace strategy may get advantage from utilizing random subspaces for both developing and collecting the base classifiers. At the point when the dataset has numerous repetitive or superfluous highlights, one may get better base classifiers in random subspaces than in the first component space. The consolidated choice of such base classifiers might be better than a solitary classifier built

on the first preparing dataset in the total capabilities. The base learners utilized are SVM, k-NN and Decision tree. SVM is a best in class information mining method that has demonstrated its presence in numerous applications. It requires twelve occasions for preparing. SVM can catch the inalienable attributes of the information superior to ANN can. k-NN is one of the least difficult and rather minor classification techniques. An object of KNN is ordered by a larger part vote of its neighbors. On the off chance that K = 1, at that point the article is just allocated to the class name of its closest neighbor. DT is a successive model, which consistently joins an arrangement of straightforward tests. Each test thinks about a numeric trait against limit esteem or an ostensible quality against a lot of potential qualities. The engineering of the proposed framework has appeared in fig 1.

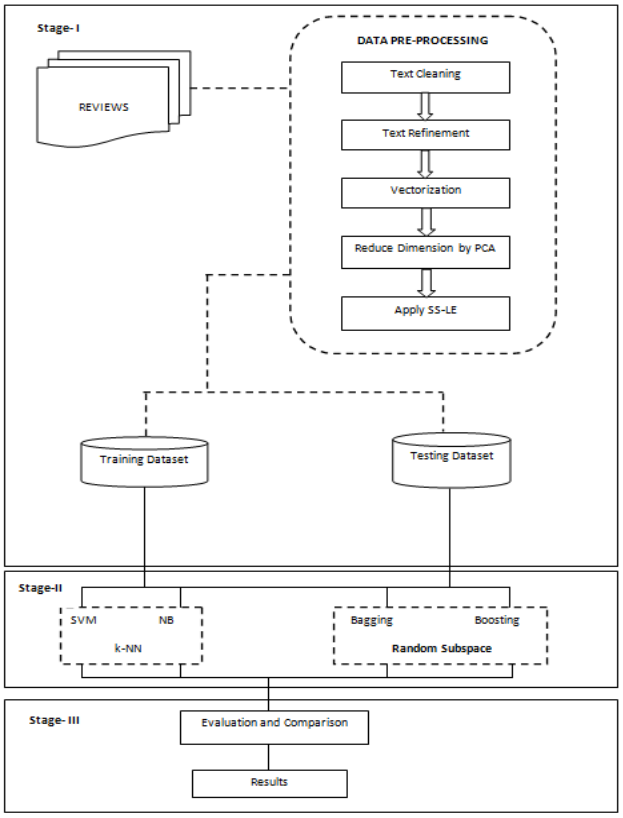

Fig. 1 System Architecture

IV. RESULTS AND DISCUSSION

This segment manages the test results. The dataset utilized is the Amazon item audit datasets. In Book dataset, there will be 1000 positive and negative reviews. In that, all out number of highlights in the negative survey is 70,394. Utilizing SentiWordNet 3.0, emphatically positive, unequivocally negative, positive, negative highlights alone are considered for further handling. With the diminished capabilities, Head Segment Examination is finished. The diminished highlights appear on table 1. The proposed technique execution can be assessed utilizing normal exactness as to the standard measure. The exactness is the extent of genuine outcomes (both genuine positives and genuine negatives) among the absolute number of cases inspected. Normal Precision can be determined as:

$$\text{Average Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

The three ensemble techniques with three base students execution can be thought about utilizing their normal precision esteems. Separating the sentimental words from the audit message in the wake of doing the stop-word expulsion and stemming. The term frequency-inverse document frequency is determined for the separated words. With the TFIDF framework, PCA is determined. The example estimations of eigenvector and eigenvalue are demonstrated as follows. The words with bigger eigenvalues are considered and the rest of the words are sifted. The graphical portrayal PCA values for each dataset Books, DVD, Gadgets and Kitchen Equipment are in the following.

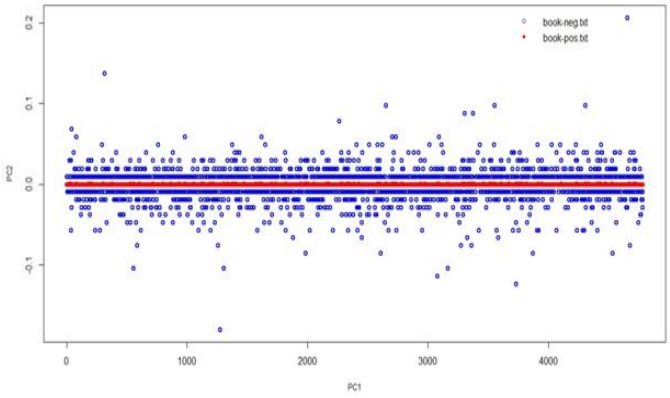

Fig. 2 PCA Values for Book Dataset

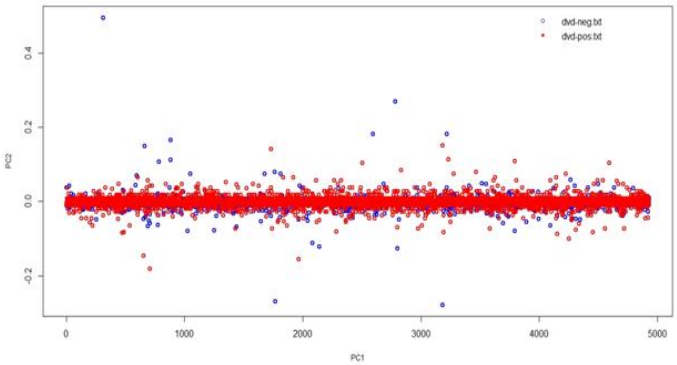

Fig. 3 PCA Values for DVD dataset

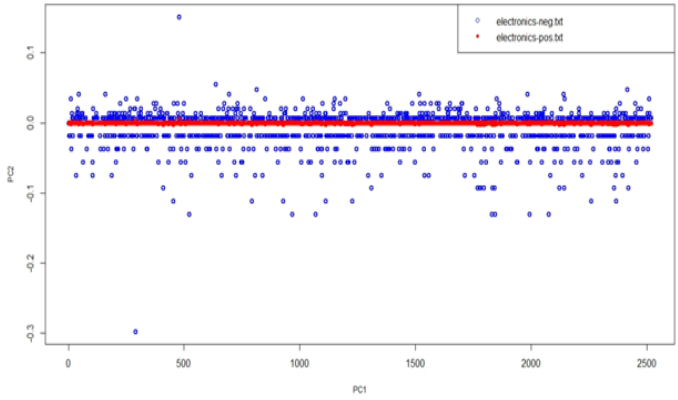

Fig. 4 PCA Values for Electronics Dataset

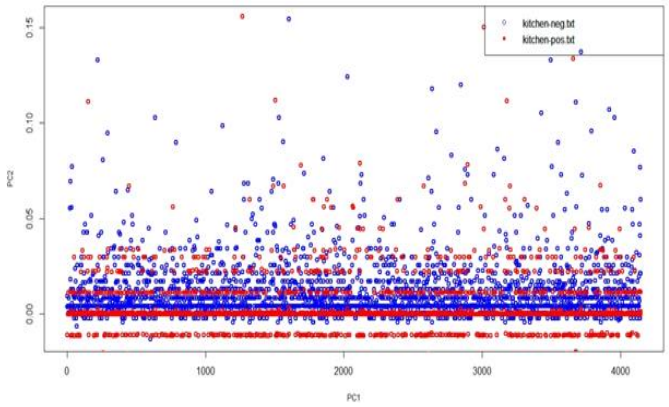

Fig. 5 PCA Values for Kitchen Dataset

Table 1. Number of Reduced Features

S. No	Amazon Review Datasets	No. of Features	Reduced Features Using SentiWordNet	After Performing PCA	After Performing LE
1	Books	1,39,571	25,330	14783	10681
2	DVD	27,44,365	4,64,934	14931	1286
3	Electronics	1,24,145	20,308	12520	9315
4	Kitchen-Hardware	4,28,570	84,976	14145	11215

Table 2. Prediction Accuracy (%) between Original set and Reduced Set

Data-set / Classifiers		Book	DVD	Electronics	Kitchen
SVM	Original set	79.26	81.34	84.42	80.36
	Reduced set	80.07	89.85	88.96	83.79
K-NN	Original set	72.14	71.87	70.65	73.86
	Reduced set	80.27	89.36	81.97	84.92
DT	Original set	79.25	80.14	78.12	80.24
	Reduced set	81.95	82.96	80.97	83.91
Bagging	Original set	80.54	79.71	81.29	80.13
	Reduced set	86.97	87.96	87.95	88.97
Boosting	Original set	82.07	81.13	80.86	79.06
	Reduced set	86.06	87.65	85.97	86.27
Random Subspace	Original set	81.27	80.38	84.26	82.19
	Reduced set	87.95	88.96	87.97	89.92

V. CONCLUSION

The issue of high dimension review data sets is settled with our given thoughts along with the ongoing examinations for characterizing that sentimental information utilizing an ensemble system. A conceivable course for future research would incorporate, the increasingly modern methodology for dimensionality reduction of sentimental information ought to be presented.

REFERENCES

1. Clarkson, P., and Rosenfeld, R. 1997. Statistical language modeling using the CMU–cambridge toolkit. In *Proc. Eurospeech'97*, 2707–2710.
2. B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (January(1–2)) (2008) 1–135.
3. S.T.Roweis, L.K.Saul, Non linear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326
4. J.B.Tenenbaum,Mapping a manifold of perceptual observations, in :Advances in Neural Information Processing Systems,vol.10,MIT Press,1998,pp.682–688.
5. B. Schölkopf, A.Smola, K.-R.Müller, Non linear component analysis as a kernel eigen value problem, Neural Computation 1299–1319.
6. Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. Decision Support Systems, 57, 77–93.
7. Abbasi, A., Chen, H., & Salem, A. (2008a). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26, 12.
8. M.Taboada, J.Brooke, M.Tofiloski, K.Voll, M.Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37(April(2))(2011) 267–307
9. A.Neviarouskaya, H.Prendinger, M.Ishizuka, Sentiful :a lexicon for sentiment analysis, IEEE Transactions on Affective Computing2(1)(2011)22–36.
10. B.Pang, L.Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the ACL,2004, pp. 271–278.
11. Salvetti, Franco, Stephen Lewis, and ChristophReichenbach. "Automatic opinion polarity classification of movie," Colorado research in linguistics 17, no. 2004
12. B.Pang, L.Lee, S.Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in :Empirical Methods in Natural Language Processing, 2002,pp.79–86.
13. B.Pang, L.Lee, Seeing stars:exploiting class relationships for sentiment categorization with respect to rating scales,in:Proceedings of the Association for Computational Linguistics.
14. R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi and T. Li, " Dual sentiment analysis : Considering two sides of one review, " in IEEE transactions on knowledge and data engineering, vol. 27, no. 8, pp. 2120 - 2133, 2015
15. Rui Xia, FengXu, Jianfei Yu, Yong Qi and Erik Cambria, " Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis, " Information Processing & Management 52, no. 1, pp. 36 - 45, 2016
16. Xia, Rui and Wang, Cheng and Dai, Xinyu and Li, Tao, "Co-training for Semi-supervised Sentiment Classification Based on Dual-view Bags-of-words Representation," Association for Computational Linguistics (ACL 1), pp. 1054-1063, 2015.
17. Yuan Wang, Zhaohui Li, Jie Liu, Zhicheng He, Yalou Huang and Dong Li, " Word Vector Modeling for Sentiment Analysis of Product Reviews, " Natural Language Processing and Chinese Computing 2014, pp. 168-180, 2014.
18. S. Das and M. Chen, "Yahoo! For Amazon: Sentiment extraction from small talk on the web, " Management science , Vol.53, issue no.9, pp. 1375-1388, 2007

AUTHORS PROFILE

M. Anandapriya completed her bachelor as well as master degree in Computer Science and Engineering from Anna University, Chennai, India in 2013 and 2015 respectively. From July 2015 to May 2016, she was working as an Assistant Professor at Kamaraj College of Engineering and Technology, Virudhunagar. From June 2016, she is working as an Assistant Professor in the Department of Computer Science and Engineering at Sri Krishna College of Engineering and Technology, Coimbatore. Her main research interests are Big Data and Analytics, Data Mining and it's applications, Software Architecture, Database Management System and so on.

M. S. Gowtham is working as an Assistant Professor at Karpagam Institute of Technology, Coimbatore since 2012. He is doing his Ph.D., (Part-Time) in the Department of Electronics and Communications Engineering at Karpagam Academy of Higher Education, Coimbatore. He completed his bachelor as well as master degree in Electronics and Communication Engineering branch from Anna University, Chennai, India in 2010 and 2012 respectively. His fundamental research interests are MANET's, Network Security, Cross Layer Optimization, and Wireless Networks.

Dr. Kamalraj Subramaniam currently works at the Department of Electronics & Communication Engineering, Karpagam Academy of Higher Education. Kamalraj does research in Biomedical Engineering, Bioengineering and Electronic Engineering. He completed the B.E. Electronics and Communication Engineering and M.E. VLSI Design degree from Anna University, Chennai in 2007 and 2009, respectively. He got the Ph.D. degree in University Malaysia Perlis, Perlis, Malaysia, in 2014. His research interests are Bio Signal Processing, Medical Image Analysis, Fractal Set Analysis, Very Large Scale Integration (VLSI), Mathematical Modeling and Algorithms, Artificial Neural Networks, Swarm Algorithms.