Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Automatic Sentiment Analysis for Unstructured Data

Jalaj S. Modha* *Computer Engineering, LJIET Gujarat Technological University Gujarat, India*

Prof & Head Gayatri S. Pandi *Computer Engineering, LJIET Gujarat Technological University Gujarat, India*

Sandip J. Modha *Computer Engineering DA-IICT Gujarat, India*

Abstract— Now-a-days Big Data have been created lot of buzz in technology world. Sentiment Analysis or opinion mining is very important application of 'Big Data'. Sentiment analysis is used for knowing voice or response of crowd for products, services, organizations, individuals, movie reviews, issues, events, news etc... In this paper we are going to discuss about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s). So, we propose new approach classify and handle subjective as well as objective statements for sentimental analysis.

Keywords— Sentiment Analysis, Text Mining, Machine learning, Natural Language Processing, Big Data

I. INTRODUCTION

Big Data is trending research area in computer Science and sentiment analysis is one of the most important part of this research area. Big data is considered as very large amount of data which can be found easily on web, Social media, remote sensing data and medical records etc. in form of structured, semi-structured or unstructured data and we can use these data for sentiment analysis.

Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes[1]. Sentiment Analysis includes branches of computer science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding. By using approaches, methods, techniques and models of defined branches, we can categorized our data which is unstructured data may be in form of news articles, blogs, tweets, movie reviews, product reviews etc. into positive, negative or neutral sentiment according to the sentiment is expressed in them.

Sentiment analysis is done on three levels [1]

- Document Level
- Sentence Level
- Entity or Aspect Level.

Document Level Sentiment analysis is performed for the whole document and then decide whether the document express positive or negative sentiment. [1]

Entity or Aspect Level sentiment analysis performs finer-grained analysis. The goal of entity or aspect level sentiment analysis is to find sentiment on entities and/or aspect of those entities. For example consider a statement "My HTC Wildfire S phone has good picture quality but it has low phone memory storage." so sentiment on HTC's camera and display quality is positive but the sentiment on its phone memory storage is negative. We can generate summery of opinions about entities. Comparative statements are also part of the entity or aspect level sentiment analysis but deal with techniques of comparative sentiment analysis.

Sentence level sentiment analysis is related to find sentiment form sentences whether each sentence expressed a positive, negative or neutral sentiment. Sentence level sentiment analysis is closely related to subjectivity classification. Many of the statements about entities are factual in nature and yet they still carry sentiment. Current sentiment analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment [1]. For Example, "I bought a Motorola phone two weeks ago. Everything was good initially. The voice was clear and the battery life was long, although it is a bit bulky. Then, it stopped working yesterday. [1]" The first sentence expresses no opinion as it simply states a fact. All other sentences express either explicit or implicit sentiments. The last sentence "Then, it stopped working yesterday" is objective sentences but current techniques can't express sentiment for the above specified sentence even though it carry negative sentiment or undesirable sentiment. So we try to solve out the above problematic situation using our approach. [1]

Our Proposed classification approach handles the subjective as well as objective sentences and generate sentiment form them.

II. RELATED WORK

A. *Sentiment Analysis for Subjective Sentences*

Sentiment Analysis is basically express sentiment of the individual person. These kinds of text or sentences are categorized as subjective sentences. It is possible that may be subjective sentences also don't carry any sentiment and

they are categorized as neutral sentences. Till now sentimental analysis techniques are performed on subjective sentences efficiently and research work done by so many researchers.

Bo Pang and Lillian Lee, research paper explains degree of positivity polarity, subjectivity detection and Opinion identification using SVM and N-gram algorithms [8]. Pang and Lee, a mincut-based algorithm was proposed to classify each sentence as being subjective or objective [9]. The algorithm works on a sentence graph of an opinion document. They also express supervised, unsupervised approaches for classification for sentiment analysis. Ana C.E.S Lima and Leandro N.de Castro presents hybrid approach of emotional-based and word-based for automatic sentimental analysis of twitter messages (i.e Tweets) and they also use basic text mining techniques and naive-Bayes classification algorithm which provide good efficiency.[6] Generally sentimental word dictionaries will be used for labelling of Small piece of data called "crunches". These kinds of dictionaries contain certain threshold value for sentiment word and the defined value is used to decide sentiment of word is positive or negative for subjective sentences. SentiWordNet V3.0 or WordNet are the online available sentiment word dictionaries [21]. For Example,

1.) Positive Sentiment in subjective sentence: "I like my new Dell Laptop" Defined sentence is expressed positive sentiment about the laptop brand Dell and we can decide that from the sentiment threshold value of word "like". Threshold value of word "like" has positive numerical threshold value. Use this threshold value in the classification algorithm like naive-Bayes.

2.) Negative sentiment in subjective sentences: "Phata poster nikala hero is the flop movie" defined sentence is expressed negative sentiment about the movie named "Phata poster nikla hero"and we can decide that from the sentiment threshold value of word "flop". Threshold value of word "flop" has negative numerical threshold value. Use this threshold value in the classification algorithm like naive-Bayes.

3.) Neutral sentiment in subjective sentences: "I'm going for a long drive" defined sentence is expressed fact. It doesn't carry any sentiment so we put this kind of statement in the neutral category. We can decide that the defined sentence is neutral because there is absence of words that express sentiment.

Polarity, subjective detection and opinion identification all are very important things in this kind of sentiment analysis.

.

B. *Sentiment Analysis for Objective Sentences*

Sentiment Analysis for objective sentences is very trending research topic now-a-days because there are so many data sources which have objective sentences that carry sentiment but because of lake of proper algorithms and contexts we can't get the fruitful result from the objective sentences. According to recent article published by Ronen Feldman express that objective sentences that carry sentiment should be analysed for getting efficient sentiment analysis and this is one of the challenging task in sentiment analysis. [1], [5]

Source of objective sentences are including news articles, blogs, social media etc. where we get good amount of objective sentences. [5]

We consider following examples which are objective sentences but still carry sentiment. [1], [5], [12]

- "Firefox keeps crashing." defined sentences carry negative sentiment about Firefox web browser.
- "The earphone broke in two days." defined sentence carry negative sentiment about the earphones.
- "I get relaxed time after today's session." define positive sentiment about person's routine.

In this particular area just challenges are proposed but still researchers are trying to find out efficient solution to get analysed these kinds of implicit opinions in the objective sentences.

Available sentiment dictionaries don't have enough vocabulary to get analysed objective sentences and categorised them efficiently into positive, negative or neutral.

Provide proper context or semantic orientation is also very important part of sentiment analysis of objective sentences.

III. EXISTING TECHNIQUES AND APPROACHES

A. *Machine Learning Techniques*

Machine learning techniques are most useful techniques for the sentiment analysis for categorized document or sentences into positive, negative or neutral categories.

Machine learning techniques classified into two basic techniques as defined below. [1]- [4]

1.) Supervised Machine Learning Techniques: Supervised machine learning techniques are used for classified document or sentences into finite set of class i.e into positive, negative and neutral. Training data set is available for all kind of classes. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.[23]

We are using Support Vector Machine (SVM), Naive-Bayes, K-nearest neighbour (KNN), Logistic regression for classification purpose.SVM efficiently classifies news articles, Blogs into positive, negative or neutral category. Naive-Bayes efficiently classifies tweets or small piece of sentences called "Crunches". KNN also give good result for sentence level sentiment analysis.

2.) Unsupervised Machine Learning Techniques: Unsupervised machine learning techniques don't use training data set for classification. Clustering algorithms like K-means clustering, Hierarchical clustering used to

classify data into categories. Semantic Orientation also provides to generate accurate result for classification. Neural network can be also used for defining threshold values to the words and classify them based on the defined values. Point wise mutual information (PMI) is also one of the unsupervised classification methods for sentiment analysis.

B. Natural Language Processing

Natural language processing techniques play important role to get accurate sentiment analysis. NLP techniques like Beg of words, Hidden markov model, part of speech (POS), N-gram algorithms, large sentiment lexicon acquisition and parsing techniques are used to express opinion for document level, sentences level and aspect level.[1,2,12]

Large sentiment lexicon acquisition is used sentiment word dictionary which contains lot of sentiment words with their numeric threshold value for particular domain [1, 5]. Now-a-days SentiWordNet dictionary is used for subjective sentiment analysis. The method defines distance $d(t1, t2)$ between terms *t1* and *t2* as the length of the shortest path between *t1* and *t2* in WordNet. The orientation of t is defined as $SO(t) = (d(t, Like) - d(t, Hate))/d(\text{Like,Hate})$. $|SO(t)|$ is the strength of the sentiment of *t*, $SO(t) > 0$ entails *t* is positive, and *t* is negative otherwise[1],[5],[12]. For objective sentiment classification we have to expand the vocabulary of SentiWordNet or WordNet by adding more words with proper threshold value.

Noun phrase (NP), verb oriented, adjective oriented sentimental analysis concentrate on NP, verb and adjective respectively to classify the sentence or entity into positive, negative or neutral.[2], [5], [13]

Word based techniques, Emotional based techniques are part of the NLP domain for sentiment analysis classification particularly for twitter message analysis. [6], [7]

C. Text Mining Techniques

Text mining techniques are also useful for efficient automatic sentiment analysis for twitter messages. Text mining process divides into four stages. In this approach supervised machine learning algorithms are used for classification purpose. [3].Text Mining Process is shown in the figure 1.

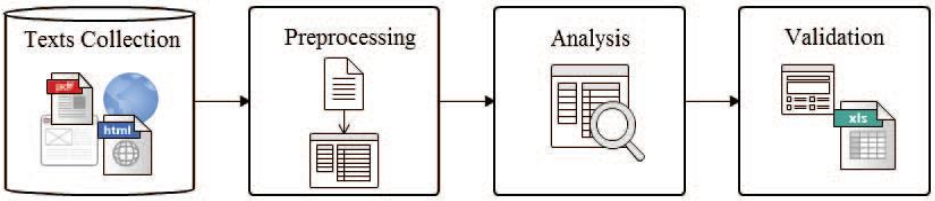

Fig 1: Text mining process [6]

Text mining classifier architecture stages are depicted in figure 2 for classification of twitter messages.

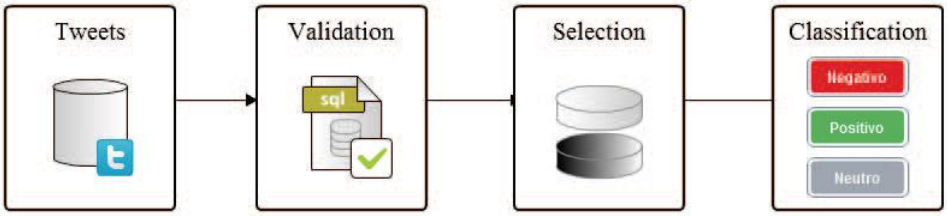

Fig 2: Text mining classification architecture stage [6]

D. Techniques of Information Theory and Coding[14], [18]

The concept of mutual information (MI), TF-IDF and random process are also used for sentiment analysis and its classification.

E. Semantic Approach[22]

For sentence level and entity or aspect level sentiment analysis the semantic approach is really useful and gives efficient result. In this approach we can us ontology learning techniques or description logic (DL) for defining semantic rules and put them together in the knowledge base. Using the rules of ontology and/or DL we can attach semantic orientation to the sentences or to the entities for proper sentiment analysis. Contexts which are provided to the sentences or to the entities are more clear and accurate.

F. Hybrid Approaches [5]-[19], [22]

We can combine any of the above approaches, techniques or methods as and when needed for efficient sentiment analysis. Some hybrid approaches are defined as below.

- We can use combination of supervised, semi-supervised and unsupervised machine learning algorithms for sentiment analysis.
- We can use naive-Bayes , word based and emotional based techniques together as hybrid model in automatic sentiment analysis.
- We can combine any of the machine learning techniques with NLP techniques like HMM, N-gram, POS, Bag of word and large sentiment lexicon acquisition for better and accurate result for implicit and explicit sentiment analysis.SVM and N-gram algorithms are used together for emotion identification of twitter messages.
- We can also combine any of the machine learning techniques, NLP techniques with semantic approach to generate proper semantic orientation for implicit opinion.
- We can also combine any of the machine learning techniques, NLP techniques with/without semantic approach to generate proper semantic orientation as and when needed for analysing objective sentences that carry sentiment.

TABLE I
ENLIST OF EXISTING APPROACHES, METHODS AND TECHNIQUES [5-19], [22]

Sr. No	Title of Research paper	Authors	Approaches, Techniques and Methods
1	Automatic Sentiment Analysis of Twitter Messages	Ana C.E.S Lima and Leandro N.de Castro	Text mining approach, Naive Bayes algorithm, word based approach, emotional based approach
2	Harnessing Twitter 'Big Data' for Automatic Emotion Identification	Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P.Sheth	POS, SentiWordNet,N-gram, Naive Bayes algorithm
3	Opinion mining and sentiment analysis	Bo Pang and Lillian Lee	Machine Learning Techniques, NLP Techniques like BOW, SentiwordNet
4	A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts	Bo Pang and Lillian Lee	Graph theory ,Cut based Classification
5	Sentiment Text classification of customers Reviews on the Web Based on SVM	Huising Xia, Min Tao and Yi Wang	Support Vector Machine Algorithm
6	Domain Independent Sentiment Classification with Many Lexicons	Bruno Ohana,Brendan Tierney and Sarah-Jane Delany	Large sentiment lexicon acquisition
7	Using Objective Word in SentiWordNet to Improve Word-of-Mouth Sentiment Classification	Chihil Hung and Hao-kai Lin	Add sentiment threshold values to the objective words
8	Verb Oriented Sentiment Classification	Mostafa Karamibekr and Ali A.Ghorbani	Verb oriented
9	Sentiment Classification based on Random Process	Jintao Mao and Jian Zhu	Random Processes
10	Sentiment Analysis of Social Issues	Mostafa Karamibekr and Ali A.Ghorbani	Verb oriented and opinion dictionary
11	Twitter Part of Speech Tagging Using Pre-Classification Hidden markov Model	Shichang Sun, Hongbo Liu,Hongfei Lin, Ajith Abraham	Hidden Markov Model
12	Sentiment Analysis of Stock Market News with Semi-supervised Learning	Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka	Polarity word dictionary
13	Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary	Sang-Hyun Cho and Hang-Bong Kang	BOW, Opinion or SentiWordNet Dictionary
14	Sentiment Classification	Aurangzeb Khan and Baharum	SentiWordNet Dictionary

Sr. No	Title of Research paper	Authors	Approaches, Techniques and Methods
	Using Sentence-level Semantic Orientation of Opinion Terms form Blogs	Baharudin	
15	Social Business Intelligence Using Big Data	Gautam Shroff, Lipika Dey and Puneet Agrawal	Ontology Based Approach

IV. OUR PROPOSED APPROACH

A. *Traditional Approach:*

Here we are not following the traditional concept of sentiment analysis for classification which considers subjective sentences and ignores objective sentences even if they carry sentiment(s). Afterwards subjective sentences are categorised into positive, negative or neutral. This approach doesn't consider objective sentences.

Many of the statements about entities are factual in nature and yet they still carry sentiment(s). Current sentiment analysis approaches determine the sentiment of subjective statements and overlook such objective statements. There is a need for algorithms that use context to attach sentiment scores to objective (factual) statements.We try to address this research problem in our proposed approach. [5].

B. *Our Proposed Approach:*

In Sentiment Analysis, we have numbers of sentences or sentences of documents. All these documents or sentences may convey opinion or maybe not. Formally, there is document set D= $\{d_1, d_2, ..., d_N\}$, sentence set S= $\{S_1, S_2, ..., S_n\}$ and all these documents and sentences belong to some specific entity e where e is a product, service, topic, issue, person, organization, or event. It is described with a pair, e: (T, W), where T is a hierarchy of parts, sub-parts, and so on, and W is a set of attributes of e. So an opinion is a quintuple,$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or neutral. We can use sentiment word dictionary with c terms C= $\{c_1, c_2, ..., c_N\}$ which contain threshold value V= $\{v_1, v_2, ..., v_N\}$ for the particular term ci we have threshold value v_i. So, by using these values we can express our s_{ijkl}. [1]

In our approach we have followed four steps of classification.

1.) *First step*: First classify sentences or sentences of documents into two categories Opinionated and No-Opinionated, regardless whether it is subjective or objective.

2.) *Second Step*: In this step we have opinionated sentences so now they are classified as subjective sentences and Objective sentences.

3.) *Third Step*: The third step is classifying subjective sentences into positive, negative or neutral category. For complex type of sentences we may need to attach context or semantic orientation

4.) *Fourth Step:* The fourth step is classifying objective sentences into positive, negative or neutral category. Here also we have to provide context or sentiment orientation as and when needed.

Figure 3 and 4 describes four stage sentiment analysis classifications.

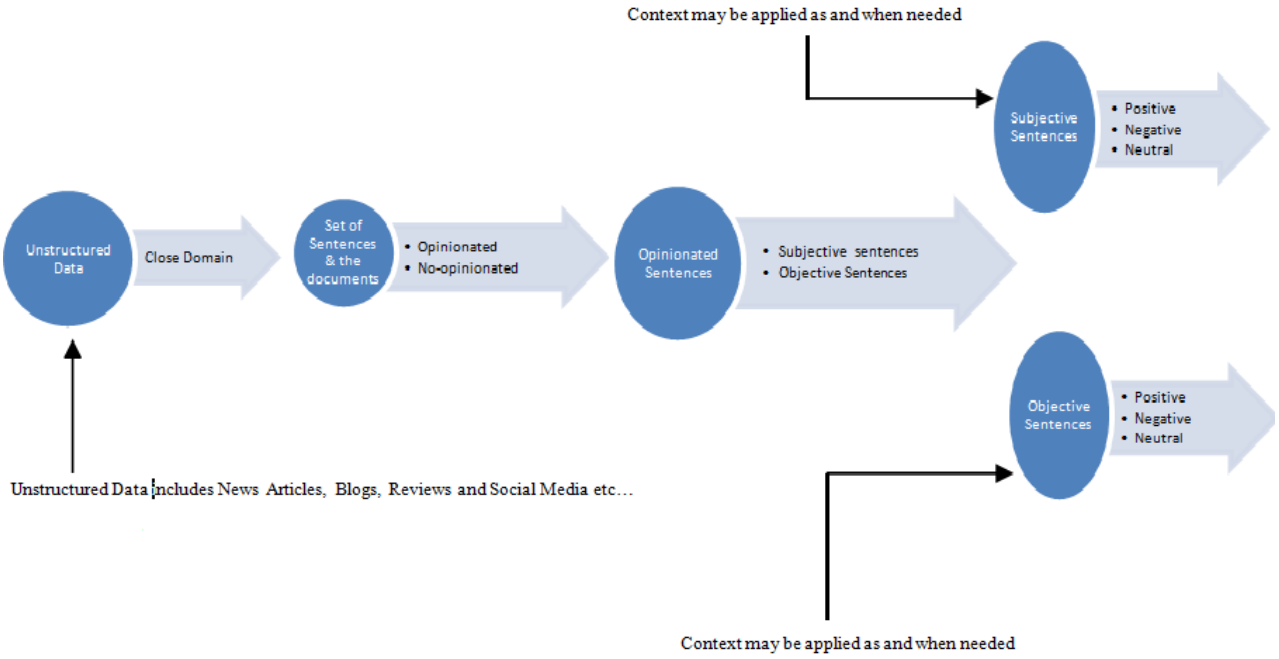

Fig 3: Our Approach of classification for sentiment analysis.

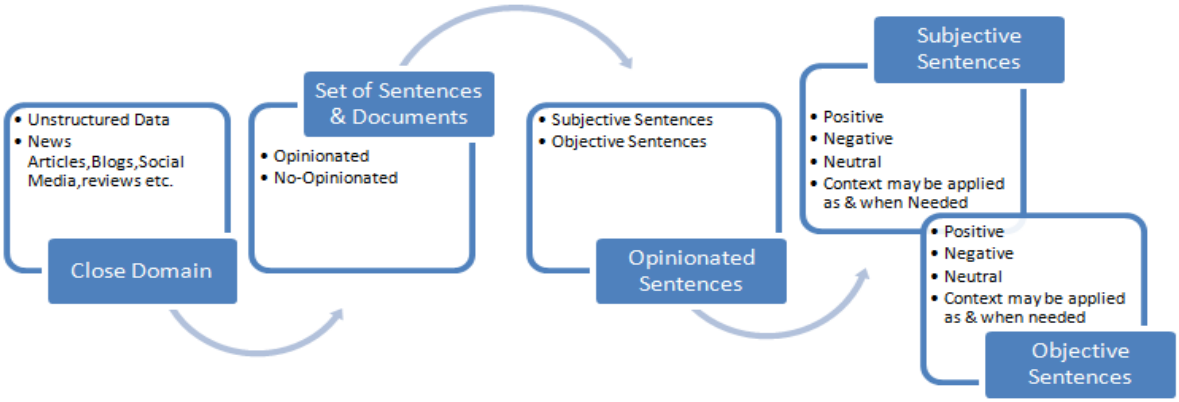

Fig 4: Flow of Our Approach of classification for sentiment analysis.

For our approach, we are going to use SVM, Naive-Bayes, BOW, POS, Large sentiment lexicon acquisition, SentiWordNet or WordNet, N-gram, grammar rules and Text mining techniques for the classification and sentiment analysis.

V. ON GOING WORK

A. *Experimental Work*

We are going to apply this approach on close domain .Our close domain is the Indian political news articles. We find good amount of objective sentences in political news articles so we take the news articles documents as well as subjective comments of people on those articles as well. Now we are actually creating our data set for the experiments. We are going to prepare the SentiWord dictionary for our domain which is Indian Political News Article. This dictionary will contain the terms and threshold values for opinionated words.

We are going to apply context by defining grammar rules and semantic orientation to the sentences for accurate sentiment analysis.

We will evaluate our experiment results by using following Information Retrieval matrices. [20]

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F-measure = 2*Precision*recall/(Precision + recall)
- Accuracy = TP + TN /(TP + TN + FP + FN)

VI. FUTURE WORK

We proposed an approach which considers subjective sentences and objective sentences both but here we are applying this approach on close domain named "Indian Political News Articles". We are expecting good efficiency for our proposed approach. In future we will apply this approach for the other close domain. We need a global SentiWord dictionary which we can use for either for open or close domain and contain proper threshold values. In future we can use the pure semantic, ontology and Description logic approach to classify subjective sentences and objective sentences and then after classify into positive, negative or neutral category.

VII. CONCLUSION

As we know today's world is becoming a narrower. We get reaction of people for particular products, events, issues very fast on web. Automatic sentiment analysis is very useful to identify and predict current and future trends, product reviews, people opinion for social issues, effect of some specific event on people; ROI and Business Intelligence applications use the sentiment analysis at big organizations like SAP, SAS and TCS. We are concentrating on both objective sentences and subjective sentences so we are going to improve the efficiency and effectiveness of sentiment analysis.

ACKNOWLEDGMENT

The authors would like to thank fellows of LJIET and DA-IICT for their reviews on this paper. I'm grateful to my guide Prof & Head Gayatri S. Pandi for her valuable suggestions and encouragement. Special thanks to Mr. Sandip J. Modha for giving me constant guidance and suggestions.

REFERENCES

[1] Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan and Claypool Publishers, May 2012.p.18-19,27-28,44-45,47,90-101.

[2] Nitin Indurkhya, Fred J. Damerau , *Handbook of Natural Language Processing, Second Edition*, CRC Press, 2010.
[3] Ronen Feldman, James Sanger,The Text Mining Handbook-Advance Approaches in Analyzing Unstructured Data, Cambridge University Press,2007.
[4] Jiawei Han, Micheline Kamber, Data *Mining Concepts and Techniques*, Morgan Kaufmann Publications, 2006.
[5] Ronen Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, April 2013, vol. 56.No.4.
[6] Ana C.E.S Lima and Leandro N.de Castro, "Automatic Sentiment Analysis of Twitter Messages", *IEEE Fourth International Conference on Computational Aspect .of Social Networks (CASoN)*, p.52-57, 2012.
[7] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P.Sheth, "Harnessing Twitter 'Big Data' for Automatic Emotion Identification " , *ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on privacy, Security, Risk and Trust*,p.589-592, 2012.
[8] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol.2, No1-2(2008)1-135.
[9] Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of ACL, 2004.
[10] Huising Xia, Min Tao and Yi Wang, "Sentiment Text classification of customers Reviews on the Web Based on SVM", *IEEE Circuts and System Society, Sixth International Conference on Natural Computation (ICNC)*, p.3633-3937, *2010.*
[11] Bruno Ohana,Brendan Tierney and Sarah-Jane Delany, "Domain Independent Sentiment Classification with Many Lexicons ", *IEEE Computer Society, Workshops of International conference on Advanced Information Networking and Application*, p.632-637, 2011.
[12] Chihil Hung and Hao-kai Lin, "Using Objective Word in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", IEEE Computer Society, P.47- 54, March-April 2013.
[13] Mostafa Karamibekr and Ali A.Ghorbani, "Verb Oriented Sentiment Classification", IEEE Computer Society, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent technology,p.327-331, 2012.
[14] Jintao Mao and Jian Zhu, "Sentiment Classification based on Random Process", IEEE Computer Society, International Conference on Computer Science and Electronics Engineering, p.473-476, 2012.
[15] Mostafa Karamibekr and Ali A.Ghorbani, "Sentiment Analysis of Social Issues", IEEE Computer Society, International Conference on Social Informatics, p. 215-221, 2012.
[16] Shichang Sun, Hongbo Liu,Hongfei Lin, Ajith Abraham, "Twitter Part of Speech Tagging Using Pre-Classification Hidden markov Model", IEEE International Conference on Systems, Man and Cybernetics, October 14-17,p.1118-1123, 2012.
[17] Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka, "Sentiment Analysis of Stock Market News with Semi-supervised Learning", IEEE Computer Society,IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328,2012.
[18] Sang-Hyun Cho and Hang-Bong Kang, "Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary", IEEE International Conference on Conference on consumer Electronics (ICCE), p.717-718, 2012.
[19] Aurangzeb Khan and Baharum Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms form Blogs", 2011.
[20] Ms.K.Mouthami, Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran, "Sentiment Analysis and Classification Based on Textual Review".
[21] Online SentiWordNet dictionary source http://sentiwordnet.isti.cnr.it/.
[22] Gautam Shroff, Lipika Dey and Puneet Agrawal, "Social Business Intelligence Using Big Data",CSI Communications, April 2013,p.11-16.
[23] Wikipedia article on supervised machine learning http://en.m.wikipedia.org/wiki/Supevised_learning.

About Authors:

Jalaj S. Modha is pursuing her Master of Engineering in computer engineering branch form LJIET, Gujarat Technological University, Ahmedabad, Gujarat. Her research interest lies in BigData and BigData analytics applications, Cloud Computing, Relational and Non-relational database System.

Prof. & Head Gayatri S. Pandi holds M.tech Degree. She is Heading the department of M.E. Computer Engineering at LJIET, Ahmedabad, Gujarat. Her research interest lies in cloud computing and Big Data.

Sandip J. Modha holds M.Tech Degree in Computer Science and Engineering form Nirma Institute of Technology. Currently he is pursuing his Ph.D from DA-IICT, Gandhinagar, Gujarat. His research interest lies in Big Data, Link Open Data, Semantic web, Cloud Computing and Service oriented architecture.

