IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques

Vishal Shirsat[1*], Rajkumar Jagdale[2], Kanchan Shende[3], Sachin N. Deshmukh[4], Sunil Kawale[5]

[1]Dept. of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad and 431004, India
[2]Dept. of Statistics, Babasaheb Ambedkar Marathwada University, Aurangabad and 431004, India

[*]*Corresponding Author: vss.csit@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i5.16 | Available online at: www.ijcseonline.org

Accepted: 07/May/2019, Published: 31/May/2019

Abstract— Now a day's sentiment analysis performs a very vital role in text mining. In essence web mining is a very broad area in a data mining field for extracts the sentiment of the text. To identify the sentiment of the textual data is a very challenging task. The present work focuses on sentence level negation identification and calculation from the News articles and Blogs. Two step approaches generally used for analysis namely preprocessing and post processing. Preprocessing consists of the tasks like stop word removing, punctuation mark removal, number removal, white space removal etc. Post processing comprises identification of sentiments from the text and calculation of score. The work analyses the performance of support vector machine, Naïve Bayes for the dataset collected online.

Keywords— Sentiment Analysis, Support Vector Machine, Naïve Bayes, Machine Learning Algorithm

I. INTRODUCTION

Sentiment analysis is one of the applications of Natural Language processing and text mining to extracts the subjective information from the text, it is also useful for extracting the contextual polarity, emotions of the textual data. Data extraction can be mined from various data sources like tweets, blogs, social media and online News articles. Sentiment analysis has been categories in three levels: document level, sentence level and entity and aspect level. Document level methods analyze polarity of whole document. The document generally contains reviews of one item; therefore system will calculate or express the overall polarity about item. Sentence level processes and analyzes each statement to determine polarity and gives polarity of each sentence. Third level i.e. entity and aspect level which is the most important level in sentiment analysis. As compared to the previous levels, this level is feature based sentiment analysis and useful to find out the sentiment of entities and their aspects. The paper proposes the role of negation in the News articles and Blog with the help of machine learning algorithms. Now a day's news articles and blogs are most important platforms that allow to users to express their personal opinion about several issues, it may be related to the politics, social responsibilities and national or international issues etc. There is vast amount of data on web in the form of text, the aim of sentiment analysis is to find out the polarity of opinion of the user. Sentiment analysis where we can easily predict the opinion of single person or group of people about particular issues as mentioned above.

II. RELATED WORK

Researchers working in the area of Sentiment analysis have developed and applied various algorithms to predict the sentiment of text from News articles and Blogs. These may be based on Natural Language Processing (NLP), Pattern-based techniques and machine learning algorithms such as Naïve Bays (NB), Support Vector Machine and Random Forest. Some researchers have used unsupervised and semi-supervised learning techniques.

Researchers working in the area of Sentiment analysis have developed and applied various algorithms to predict the sentiment of text from News articles and Blogs. These may be based on Natural Language Processing (NLP), Pattern-based techniques and machine learning algorithms such as Naïve Bays (NB), Support Vector Machine and Random Forest. Some researchers have used unsupervised and semi-supervised learning techniques.

M. Thelwall and K. Buckley [1] has shown, in Sentiment analysis, there are three approaches like machine learning based methods, lexicon based methods and Linguistic analysis.Jagdale, R. S [2] distinct the Different events on twitter has been analyzed and calculated the sentiment

polarity of each event. L. Tan, J. Na, Y. Theng[3] is to analyze text direction, linguistic approach uses syntactic features of the words or phrases, the negation, and the structure of the text. Sentiment Analysis of Natural Language texts is a broad and expanding field. A text may contain both Subjective and Objective sentiments. Wiebe [4] defines Subjective text as the linguistic expression of somebody‘s opinions, sentiments, emotions, evaluations, beliefs and speculations. In her definition, the author was inspired by the work of the linguist Ann Ban field [5], who defines subjective as a sentence that takes a character‘s point of view and that present private states (that are not open to objective observation or verification), defined by Quirk [6], of an experience, holding an attitude, optionally towards an object. Bing Liu [7] defines Objective text as the facts that are expressed about entities, events and their properties. Esuli and Sebastiani [8] define Sentiment Analysis as a recent discipline at the crossroads of Information Retrieval and Computational Linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. Shoukry [9 shows an application for Arabic language sentiment analysis and per-formed a sentiment classification for Arabic tweets. The collected tweets are examined to provide their polarity. Their study proposed hybrid system that used all the identified features from the ML approach, and the sentiment lexicon from the SO approach, resulting in an accuracy and recall of 80.9%, while it’s precision and F-measure is 80.6%. Alexandra Balahur [10] this paper examined it is essential to define the importance of the tasks in three levels. They performed the targeted text and separated good and bad news content.

III. METHODOLOGY

Preprocessing plays a very major role in a sentiment analysis, the role of preprocessing is to remove the unwanted data from the text. This type of data does not contain any important information. In this work we used BBC News article data set and Multisensory dataset for experimentation. Every online text contains information like html tags, scripts and advertisements. This step helps cleaning and preparing the data for the post processing. The present work uses Multisensor News articles and BBC News articles dataset [12, 13]. The Multisensor News articles dataset consist of 12,073 documents in categories that are economy, health, lifestyle, nature environment, politics and science and technology and BBC dataset consists of 2225 documents in categories like business, politics, entertainment, sport, technology.

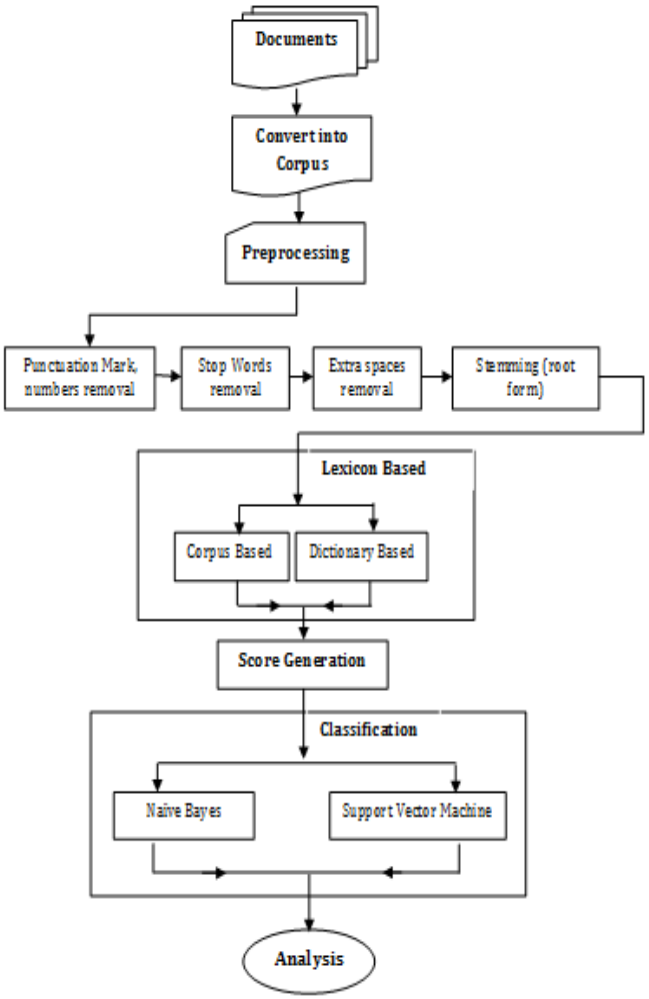

Figure. 1 Proposed Methodology

IV. MACHINE LEARNING ALGORITHMS

This work uses machine learning algorithms for the classification purpose. Classifiers used are Naïve Bayes (NB) and Support Vector Machine (SVM). Naïve Bayes classifier finds the probability of occurrence given the probability of another occurrence that has already occurred. NB classifier does particularly well for problems which are linearly separable and even problems which are nonlinearly separable and it performs reasonably well [14].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

In the above formula (1), A is the sentiment of the text and B is the text, whereas P (A|B) is the posterior probability of class and P (A) probability of class. P (B|A) is the likelihood and P (B) is the prior probability of predictor.SVM is non probabilistic algorithm which is used to separate data

sequentially and Non-sequentially [15]. It is basically used for text classification and yields good performance in high-dimensional feature space. SVM denotes the instance points Space, maps so that the instances of the different classes are separated by a clear margin as extensive as possible [16].

V. RESULTS AND DISCUSSION

This section presents the post processing work. As mentioned above, two dataset were used i.e. BBC and Multisensor. The proposed methodology uses five steps. First step performs data cleaning which removes URL, Stop words, Punctuation, Strip white spaces and Numbers from the data. Further, number removal is an important step as number very rarely represents the sentiment and hence not significant. The next step is to convert the document in to lower case to have uniformity. Stemming is used to change to root form of the word, for example "Connection", "Connecting", "Connected" into a single word i. e. "Connect". Term document frequency describes the frequency of terms that occur in document, where rows in the output are assumed as a collection and column assumes as a related terms. Identification of the sentiment is a major task, to achieve this we used sentimentr and syuzhet packages with dictionary based approach and lexicon based approach. Sentimentr package uses ten lexicon dictionaries for the sentiment identification with eleven arguments and calculates the polarity of the sentences and utilizes sentiment dictionary to tag polarized words. This package calculates sentiment by using following formula.

$$c'_{i,j} = \sum((1 + w_{amp} + w_{deamp}).w^{p}_{i,j,k}(-1)^{2+w_{neg}})$$

$$w_{amp} = (wb > 1) + \sum(w_{neg} .(z.w^{a}_{i,j,k}))$$

$$w_{deamp} = \max(w_{deamp'}, -1)$$

$$w_{deamp'} = (wb < 1) + \sum(z(-w_{neg}))$$

$$w_b = 1 + z_2 * w_{b'}$$

$$w_{b'} = \sum(|w_{adversative\ conjuncion}|, \ldots, w^{p}_{i,j,k}, w^{p}_{i,j,k}, \ldots, |w_{adversative\ conjunction}| * -1)$$

$$w_{neg} = \left(\sum w^{n}_{i,j,k}\right) mod\ 2$$

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \qquad (2)$$

Sentimentr package uses above formula to calculate sentiment of the text [6]. Syuzhet package uses three methods to calculate the sentiment that are NRC, Bing and Afinn, These three methods gives different results [17]. However, when experimented, we find Sentimentr package is best for sentence level as compare to other packages because Sentimentr package calculates the polarity of each sentence with the help of more parameters like, Positive word, Negative word, Downtowners, amplifiers, deamplifiers, adversative conjunction etc.

Table 1. Shows the comparison of Sentimentr and Syuzhet Packages with its parameters and methods for calculating the sentiments. Whereas Table no.2 shows the categeorywise polarity of BBC dataset. In the given table there are five category business, entertainment, politics, sport and technology. To calculate the polarity of sentence here we used sentimentr package. In Table No. 3. We used two machine learning techniques Naïve Bayes and support Vector Machine there we calculated accuracy, precision and f- score. Here entertainment category got more accuracy by using Naïve Bayes and Support vector machine give more accuracy for the business category. These both machine learning algorithm are the probabilistic and non-probabilistic algorithms which most popular for the sentiment analysis. After that Table no 4 shows the confusion matrix of BBC dataset which classified the results in diagonally. In confusion matrix results should be in diagonal. In second experiment we used Multisensor blog dataset whereas Table no 5 shows the results with positive, negative and neutral. In this experiment we used sentimentr package to calculate the polarity of blogs. For the Statistical measurements calculated Mean Error, Root Mean Square Error, Mean Absolute error and Mean Absolute Square error as well. Finally we come to know that here those results we got that are proved by machine leaning algorithms and statistical measurements.

Table. 1 Comparison of Sentimentr and Syuzhet Packages

Sr. No	Name of Package	Parameters and Methods
1	Sentimentr	Positive word, Negative word, Downtowners, Amplifiers, Deamplifiers, Adversative Conjunction
2	Syuzhet	NRC, Bing and Afinn

Table 2. Category wise Polarity for BBC Dataset

Sr. No	Name of Category	Articles			
		Total	*Positive*	*Negative*	*Neutral*
1	Business	510	262	214	34
2	Entertainment	401	136	244	21
3	Politics	417	210	190	17
4	Sport	511	151	327	33
5	Tech	401	136	244	21
	Total	2,240	895	1219	126

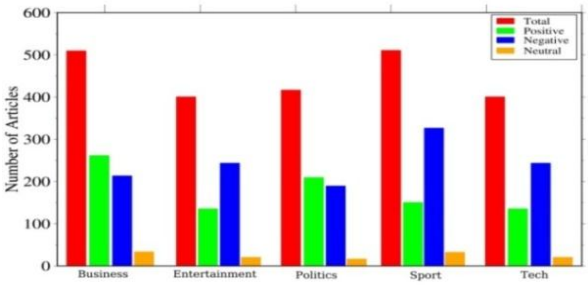

Figure 2: Category wise Polarity for BBC Dataset

Table 3 Comparative Analysis of Naïve Bays

Dataset	Naïve Bayes		
	Accuracy	*Precision*	*F-Score*
Business	92.63	89.76	91.32
Entertainment	96.46	94.80	97.33
Politics	93.33	88.88	93.33
Sport	93.00	90.74	95.14
Tech	96.46	94.80	97.33

Table 4: Comparative Analysis of SVM

Dataset	SVM		
	Accuracy	*Precision*	*F-Score*
Business	82.60	79.67	89.34
Entertainment	69.91	68.22	84.39
Politics	94.16	89.06	94.21
Sport	69.23	69.01	81.66
Tech	69.91	68.22	81.11

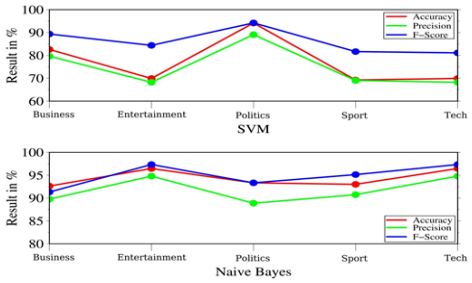

Figure 3: Comparative Analysis of Naïve Bays and SVM

Table 5. Confusion Matrix

a	b	c	d		Classified as
1	0	0	0	a	Entertainment
0	1	0	0	b	Politics
0	0	1	0	c	Sport
0	0	0	1	d	Tech

Table 6. Category wise Polarity for MULTISENSOR Dataset

Sr. No	Name of Category	Blogs			
		Total	*Positive*	*Negative*	*Neutral*
1	Economy Business Finance	3689	2488	371	830
2	Health	326	198	47	81
3	Lifestyle Leisure	3353	2471	387	495
4	Nature Environment	990	673	162	155
5	Politics	561	315	87	159
Total			2386	2047	97

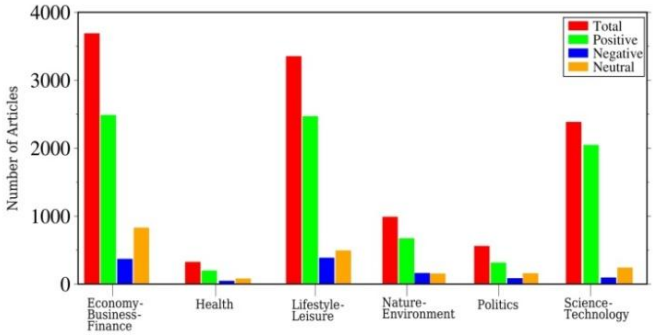

Figure 4: Category wise Polarity for MULTISENSOR Dataset

Table 7: Training Statistical measurements of Multisensor Dataset

Dataset	Training			
	ME	RMSE	MAE	MASE
Economy_Business_Finance	-4.752	0.122	0.090	1.000
Health	1.369	0.273	0.195	1.000
Lifestyle_Leisure	-1.781	0.178	0.128	1.000
Nature_Environment	4.158	0.168	0.128	1.000
Politics	2.232	0.095	0.074	1.000
Science_Technology	-1.110	0.157	0.110	1.000

Table 7: Testing Statistical measurements of Multisensor Dataset

Dataset	Testing			
	ME	RMSE	MAE	MASE
Economy_Business_Finance	-8.429	0.115	0.076	0.837
Health	-4.581	0.121	0.098	0.503
Lifestyle_Leisure	-4.891	0.203	0.142	1.111
Nature_Environment	3.219	0.205	0.142	1.106
Politics	1.757	0.102	0.075	1.014
Science_Technology	-4.571	0.130	0.097	0.883

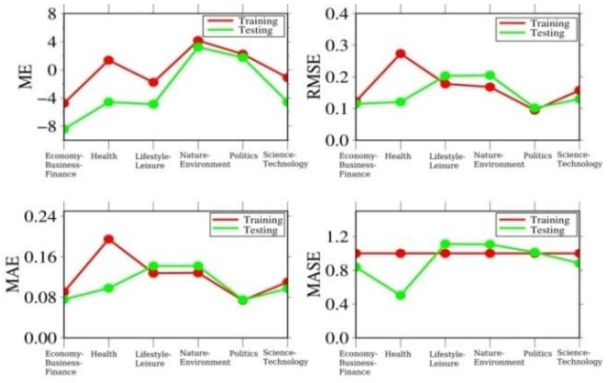

Figure 5: Comparative Analysis of Statistical measurements of Multisensor Dataset

VI. CONCLUSIONS AND FUTURE WORK

In this work we performed sentence level sentiment analysis of news articles and blogs. Experimental work performed to calculate polarity of the news articles and blogs. The results shows category wise sentence level polarity. The classification proposed work used Naive Bayes, support vector machine and random forest. Whereas Naive Bayes achieves 96.46 % accuracy and Support Vector Machine algorithms achieves 94.16 %. With these results we come to know that naive Bayes achieves more good results as compare to Support Vector Machine because Naive Bayes work on the probability of each word and its features in the given sentences as compare to Support Vector Machine. Future work will focus on other Classification machine learning techniques on other data from the websites of News articles and blogs.

ACKNOWLEDGMENT

I am indebted to the Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad and University Grant Commission, Delhi for providing all facilities to me related to my research work in the form of Rajiv Gandhi National Fellowship (SRF).

REFERENCES

[1] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment in twitter events", Journal of the American Society for Information Science and Technology 62(2), (2011) 406 418.
[2] Jagdale, R. S., Shirsat, V. S., Deshmukh, S. N, "Sentiment Analysis of Events from Twitter Using Open Source Tool." International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.5, Issue.4, April- 2016, pg. 475-485.
[3] L. Tan, J. Na, Y. Theng, K. Chang, "Sentence-level sentiment polarity classification using a linguistic approach, Digital Libraries: For Cultural Heritage", Knowledge Dissemination, and Future Creation (2011) 77 87.
[4] Wiebe, J., "Tracking point of view in narrative.‖ Computational Linguistics", 20, 1994.
[5] Banfield, A., "Unspeakable sentences: Narration and Representation in the Language of Fiction", Routledge and Kegan Paul, 1982.
[6] Quirk, R., "A Comprehensive Grammar of the English Language", Longman Publishing House, 1985.
[7] Bing Liu, "Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing, Second Edition, 2010.
[8] Esuli, A. and F. Sebastiani, "SentiWordNet: A Publicly Available Resource for Opinion Mining", In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2006, Italy, 2006.
[9] Shoukry, Amira, Collaboration Technologies and Systems (CTS), 2012 International Con-ference technologies and systems ,21-25 May 2012, Page(s):546 – 550
[10] Alexandra Balahur, Ralf Steinberger, 'Rethinking Sentiment Analysis in the News", Theory to Practice and back‖, European Commission, Joint Research Centre, University of Alicante, Department of Software and Computing Systems.WOMSA, pp.1-12, 2009.
[11] Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. 36, 6527–6535 (2009)
[12] D. Liparas, Y. Hacohen-Kerner, A. Moumtzidou, S. Vrochidis and I. Kompatsiaris, "News articles classification using Random Forests and weighted multimodal features", 3rd Open Interdisciplinary MUMIA Conference and 7th Information Retrieval Facility Conference (IRFC2014), Copenhagen, Denmark, November 10-12, 2014.
[13] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.
[14] Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. 36, 6527–6535 (2009)
[15] Bhumika, M., Jadav, V., Vaghela, B.: Sentiment analysis using support vector machine based on feature selection and semantic analysis. Int. J. Comput. Appl. 146(13) (2016).
[16] BholaneSavita,D., Deipali, G.: Sentiment analysis on twitter data using support vector machine. Int. J. Comput. Sci. Trends Technol. 4(3) (2016)
[17] Rinker, T. W. (2018). sentimentr: Calculate Text Polarity Sentiment version 2.6.1. http://github.com/trinker/sentimentr
[18] Jockers ML (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text. https://github.com/mjockers/syuzhet.

Authors Profile

Sachin N. Deshmukh is currently working as Professor in Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and having experience of around twenty four in teaching for post graduate (M. Tech, M.Sc. and MCA) and graduate B.E., B. Tech courses. He has published more than 80 Research papers in National and International reputed journals and conferences. University Authorities also have given the responsibility as Director (University Network InformationCenter), Director (Center for Vocational Education and Training), Chief Coordinator of Spoken Tutorial Project ofIIT, Mumbai. He also worked on research projects of UGC and AICTE. Apart from University, worked in AICTE New Delhi on deputation as Deputy Director (e-Governance) and as Associate Professor at COEP Pune on Lien. He is member of EQASA Member, Manber NAAC Peer Team, Member UGC Committee, Member AICTE-SCSC, & AICTE-SCAC, PEIN-Fellow San-Diago Spain, Fellow IETE.His area of research is Text mining, Social Media Data Analytics, Sentiment Analysis and Opinion Mining, Intension Mining.

Sunil Kawale is currently working as Professor in Department of Statistics, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and experience of around twenty years in teaching for post graduate (M.Sc. Statistics). His research area is Operations Research, Stochastic Processes, Computer Programming and Applications, Data Mining.

Vishal S. Shirsat has received M. Phil. (Computer Science) From Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and now he is pursuing Ph.D. in same department and He got National fellowship for his Research Work. His research area is Sentiment Analysis and Opinion Mining.

Rajkumar S. Jagdale has received M.Sc. (Computer Science) From Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and now he is pursuing Ph.D. in same department and He got DST Inspire fellowship for his Research Work. His research area is Sentiment Analysis Opinion Mining.

Kanchan Shende has received M. Phil. (Computer Science) From Department of Computer Science and IT, Dr Babasaheb Ambedkar Marathwada University, Aurangabad and now she is pursuing Ph.D. in same department and He got National fellowship for her Research Work. Her research area is Oceanography, Metrology , Remote Sensing and GIS.

