COMPARATIVE PERFORMANCE OF RANDOM FOREST AND SUPPORT VECTOR MACHINE ON SENTIMENT ANALYSIS OF REVIEWS OF INDIAN TOURISM

Smita Selot
Professor, Dept of CA
SSTC, Bhilaismitaselot504@gmail.com

Sreejit Panicker
Associate Professot, Dept of CSE
SSTC, Bhilai
sreejit.bhilai@gmail.com

Abstract

Automatic sentiment extractions from social media reviews is new trend in business analysis. It visualizes and summarizes the sentiments extracted from millions of reviews in set of predefined three classes: positive, negative and neutral. Foundations of automated sentiment analysis lies in Natural language Processing (NLP) and Machine Learning (ML) algorithms. Through this paper we are presenting results of applying two robust supervised machine learningalgorithms on Indian tourism reviews: Random Forest(RF) and Support Vector Machine(SVM) and compare the performance of both on the 11K dataset collected through Tripadvisor.com. It is found that using a limited feature, RF outperforms SVM in terms of accuracy and execution time.

Keywords: Sentiment analysis, natural language processing, Random Forest, Support Vector machine.

1. Introduction

Sentiment analysis (SA) is state-of-art technique for analyzing people's opinions, thought, expressionand attitude towards an entity through written text[5].With exponential increase of internet users; data in form of reviews has grownrapidly; as a result, more than 94% of the customer read online reviews before making a decision of acquiring a product.[2][3].These reviews are bulky and unstructured in nature; but, with proper visualization and analysis it can offer meaningful information to business community. SA of reviews aims at automatic detection of sentiment polarity of the sentence with respect to target [12][13]. Hence, SA is gradually becoming trend in business analysis since 2015 as shown by comparing trends of the terms' sentiment analysis (in blue) with feedback (in red) in Fig1. A significant growth in the field of SA is observed from 2015 onwardsthere by improving scope of research avenue in the field.

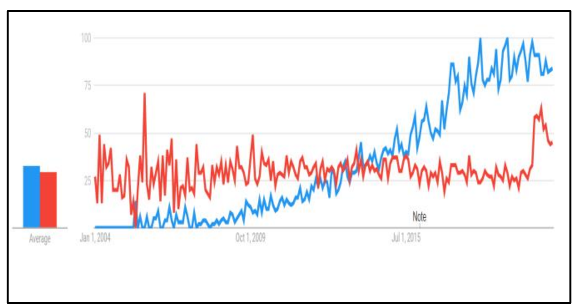


Fig 1: Increase trend in Sentiment Analysis

Sentiment analysis is(also called opinion mining, review mining or appraisal extraction, attitude analysis) is the task of detecting, extracting and classifying opinions, sentiments and attitudes as expressed in textual input [Balahar].

Indian tourist places are source of attraction not only for local travelers butalso international toursits.With rich heritage and diversified culture, Indian tourist places are

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

source of interest and draws attraction from all segments of society. It plays a vital role in economic and socio growth of the country. The World Travel and Tourism Council reported that in 2018; tourism in India contributed 9.2% of India's GDP growth and predicted to grow at rate of 6.9% annually.

Sentiment analysis of reviews on Indian tourism will bring an insight in detecting public opinion about tourists' places. Automatic categorization of reviews will further aid in further enhancing placed by resolving specific issues observed through reviews. Substantial work on SA of product reviews is available; but application of SA on Indian tourism review is seldom observed. Most of the websites accept reviews in English and India is country with diverse language. Hence, sentiment expressed by an individual is limited by the vocabulary used by authors. With objective of building classifier on Indian tourism reviews;online reviews of tourists spot of Chhattisgarh and Maharashtra are gathered from Tripadvisor.com using web scratcher tool and datasetwas generated. Nearly 11K reviews were collected and tagged manually as positive and negative for training supervised learning models. Two robust supervised machine learning algorithms; RF and SVM are implemented and evaluated on test data. Paper is divided into five section; next section presents work done in the area of SA. Data collection and methodologies are discussed in section 4. Results and comparative analysis are done in section 5.

2. Literature Review

Sentiment Analysis has evolved as an active research domain since last one decade[6]. It is subset of NLP; a very challenging field in computer science research. Intelligent system with natural language understanding has to deal with ambiguities within that language; like, word sense disambiguation (WSD); where a word has different meaning in different context; coreference resolution (CR); where pronoun is substituted with correct proper noun. All of these subproblems themselves are individual research areas [23]. In past few years; availability of NLP tools for part-of-speech(POS) tagging; tokenization, lemmatization, vectorization, syntactic and semantic representation has channelized research work in more promising direction.

Different types of sentence possess different types of sentimentvalues. Not all the sentences in the review areopiniated; some are factual with no sentiments. They are called as objective sentences; whereas sentences with opinion are termed as subjective sentences[7][11][13].Subjective sentences areclassified as positive and negative based on overall sentiment value of sentence [17].Supervised machine learning models are designed for understanding the sentiments and classifying them as positive and negative [30]. Sentiment or opinions are associated with target[4]. For example, sentence "*I like the phone but battery life is less.*"; has two target *phone* and*battery life* and two sentiments *like* and*less*. Single or multiple targets and opinions are present in the sentence Identification of accurate target and its sentiment from complex sentence is also a challenging task handled by ML based algorithm [10] . Not only adjectives orientations are being analyzed for SC task; but emoticons and thumps up are also used as features for sentiment classification task.

Features from set of reviews or documents play a significant role in training a model. Some of the important features are word frequency, term frequency inverse document frequency (tfidf), part of speech tags (POS), type-token ratio (TTR). one hot encoding(OHE) of words in vocabulary, vector representation of words and sentences[15][16].In real scenarios, training samples may not contain target values for all possible combination of features. Some combination occurs more frequently than others; This leads to data sparsity in high dimensional feature set. Training a model with sparse data leads to overfitting situation; as model will learn frequent occurring patterns and will not perform well with new or less frequent pattern in test data. Model is not well generalized in such scenarios. This problem of generalization is also a challenge in predicting correct sentiment from a new type of sentence. Vectors at word level and sentence level give a better presentation of feature set as it tends to reduce the dimensionality problem by projecting similar words or sentence in same axis in multidimensional representation of text [9][20].

Indian tourism sector is an important segment affecting social and economic parameters of the country. Implementing SA task in reviews of this sector will assist in visualizing potential problems in the specific areas by projection of emotions of tourists through their reviews and its analysis through ML approaches.

Most robust algorithms in SA task are support vector machine (SVM) and Random Forest (RF) [33][12]as reflected in literature survey of NLP task. Hence these two algorithms are implemented for the sentiment classification reviews and their performance is reported in this paper. Results are further enhanced by tuning parameters of the model.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

3. Methodology

Process of building a robust model for sentiment analysis is divided into six steps as shown in Fig1. *Preprocessing* steps removes noise from data and make it ready for next phase.

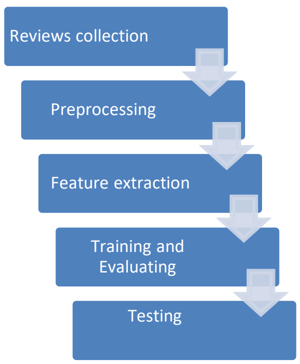
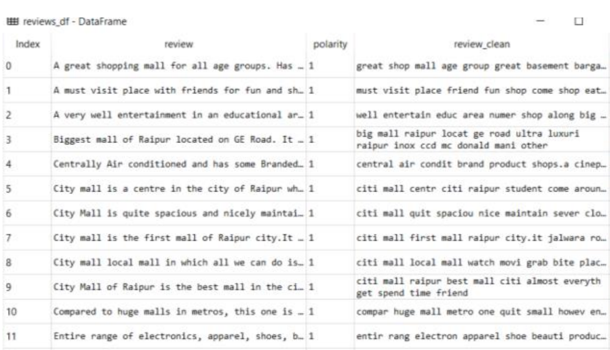


Fig1: Steps for Sentiment classification

Intelligent system cannot be trained on textual data. Hence, numerical representation of words and sentences is obtained in vectorsform through word2vev and doc2vec tools.*Features extraction* in text is obtained through vectorization process.Vectorization process assigns vector to words and sentences in the documents. Each word or vector represent a dimension in n-dimension space [7] Words with similar meaning fall on same axis. Model is built and is *trained* on 80% of the data. Parameters of the model are tuned for optimum performance. Trained model is tested on 20% of the data. Performance of the model is *evaluated* on test data and accuracy scores are determined.



reviews_df - DataFrame

Index	review	polarity	review_clean
0	A great shopping mall for all age groups. Has …	1	great shop mall age group great basement barga…
1	A must visit place with friends for fun and sh…	1	must visit place friend fun shop come shop eat…
2	A very well entertainment in an educational ar…	1	well entertain educ area numer shop along big …
3	Biggest mall of Raipur located on GE Road. It …	1	big mall raipur locat ge road ultra luxuri raipur inox ccd mc donald mani other
4	Centrally Air conditioned and has some Branded…	1	central air condit brand product shops.a cinep…
5	City mall is a centre in the city of Raipur wh…	1	citi mall centr citi raipur student come aroun…
6	City Mall is quite spacious and nicely maintai…	1	citi mall quit spaciou nice maintain sever clo…
7	City mall is the first mall of Raipur city.It …	1	citi mall first mall raipur city.it jalwara ro…
8	City mall local mall in which all we can do is…	1	citi mall local mall watch movi grab bite plac…
9	City Mall of Raipur is the best mall in the ci…	1	citi mall raipur best mall citi almost everyth get spend time friend
10	Compared to huge malls in metros, this one is …	1	compar huge mall metro one quit small howev en…
11	Entire range of electronics, apparel, shoes, b…	1	entir rang electron apparel shoe beauti produc…

Fig 2: Sample data after preprocessing

Approximately, 11K reviews of Indian tourist places; collected from Tripadvisor.com are used for experimental purpose. Reviews from 11 tourist places of Chhattisgarh and 18 tourists places from Maharashtra are manually tagged as positive and negative. Summary of state wise data is depicted in Fig 3.

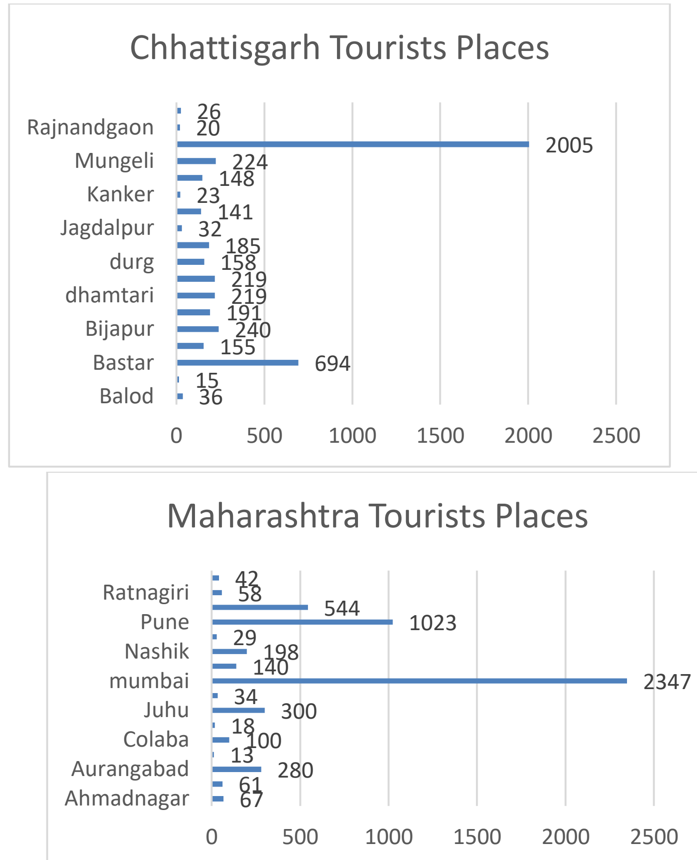
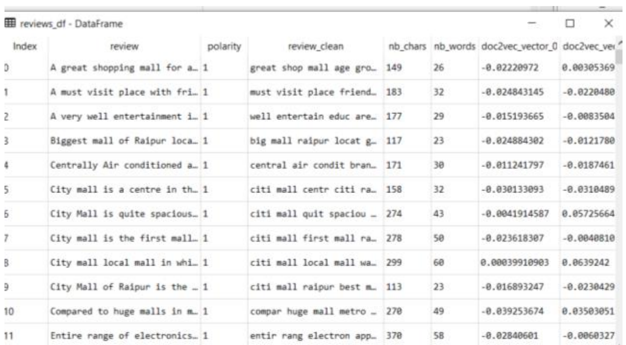


Fig3: Summary of data collected from two states

With availability of open-source tools for NLP and ML, experimenting with ML models for solving NLP based complex problem has increased manifolds. These techniques take input as feature vectors which represent properties or characteristics feature of text and train a system to learn these characteristics. Some of features commonly used are: POS tags, number of words, number of characters, TTR ratio, tfidf, one-hot-encoding (OHE),word and sentence vectors. Total number of unique words in corpus form vocabulary and type token ratio (TTR) of a corpus defines the diversity of the corpus. TTR is ratio of vocabulary size by number of tokens in corpus. Apart from these conventional feature set; word embeddings play an important role in text classifications task [3]. Word embeddings not only give vector representation to words but also capture semantic and syntactic relevance [8]. Basic vector representation is given byOne-Hot-Encoding(OHE). Every word in vocabulary is identified by a unique vector and size of this vector depends on number of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

words in vocabulary. OHE gives sparse representation of words as similar words are represented by different vectors. More dense representation is obtained by word2vec and doc2vec where similar meaningwords or sentences are projected onto same axis in a multidimension representation [9]. This reduces the vector size and gives better numerical representation of sentences. Word2vec converts words to vector and doc2vec converts sentence to a vector. Doc2vec is unsupervised model build on word2vec where paragraph id is added as additional parameter. Here we are using doc2vec as one of key feature in building a model; apart from number of characters, number of words, POS tag, tfidf and TTR values.Doc2vec generates vector of for reviews Sample feature set is shown in Fig 4.


reviews_df - DataFrame

Index	review	polarity	review_clean	nb_chars	nb_words	doc2vec_vector_0	doc2vec_ve
0	A great shopping mall for a…	1	great shop mall age gro…	149	26	-0.02220972	0.00305369
1	A must visit place with fri…	1	must visit place friend…	183	32	-0.024843145	-0.0220480
2	A very well entertainment i…	1	well entertain educ are…	177	29	-0.015193665	-0.0083504
3	Biggest mall of Raipur loca…	1	big mall raipur locat g…	117	23	-0.024884302	-0.0121780
4	Centrally Air conditioned a…	1	central air condit bran…	171	30	-0.011241797	-0.0187461
5	City mall is a centre in th…	1	citi mall centr citi ra…	158	32	-0.030133093	-0.0310489
6	City Mall is quite spacious…	1	citi mall quit spaciou …	274	43	-0.0041914587	0.05725664
7	City mall is the first mall…	1	citi mall first mall ra…	278	50	-0.023618307	-0.0040810
8	City mall local mall in whi…	1	citi mall local mall wa…	299	60	0.00039910903	0.0639242
9	City Mall of Raipur is the …	1	citi mall raipur best m…	113	23	-0.016893247	-0.0230429
10	Compared to huge malls in m…	1	compar huge mall metro …	270	49	-0.039253674	0.03503051
11	Entire range of electronics…	1	entir rang electron app…	370	58	-0.02840601	-0.0060327

Fig 4: Sample reviews with feature set

A word gets its numeric representation through word2vec using CBOW or skip-gram variations of neural network model. CBOW works on the principle of predicting a word; given the context and skip gram is flipped model of CBOW; which predicts the context, given the word. To build vector representation of varying length sentences or reviews; paragraph identification is added as a new input to base model of word2vec. System learns document specific representation and returns vector using two implementations:

- Paragraph Vector - Distributed Memory (PV-DM)
- Paragraph Vector - Distributed Bag of Words (PV-DBOW)

PV-DM is an extension of CBOW which predicts the word;given context along with paragraph id. It not only learns the semantically related words in a sentence; but also captures topic specific representation. Likewise; PV-DBOW is enhanced model of skip-gram implementation.

Support Vector Machine (SVM)

Support vector machine (SVM) is a ML algorithm used for classification as well as regression task. In classification problem; its objective is to find decision boundary, a hyperplane, that divides data points into two regions. It implements**kernel**, a mathematical function of linear and non-linear nature for the finding the decision boundary; that separates data point with maximum margin from support vectors. Support vectors are data points nearest to hyperplane amongst all data points of a class. A plane that separates datapoints with maximum distance from both the sets is an optimum hyperplane. Training data set is collection of :

$$D=\{(\mathbf{x}^{(1)},y^{(1)}),(\mathbf{x}^{(2)},y^{(2)}),(\mathbf{x}^{(3)},y^{(3)}),\ldots\ldots,(\mathbf{x}^{(n)},y^{(n)})\} \quad \text{----(7)}$$

where $\mathbf{x} = [x_1, x_2, x_3 \ldots x_n]^T$ is n dimensional input vector for i^{th} example in the real-valued space; y is class label. SVM finds a linear function $g(\mathbf{x})=\mathbf{w}^T\mathbf{x}+w_0$;such that the input vector $x^{(i)}$ is allocateda class based on the value of $g(\mathbf{x}^{(i)})$.and to other class if it is less than zero.

$$y^i = \begin{cases} +1 \ldots \mathbf{w}^T x^{(i)} + w_0 > 0 \\ -1 \ldots \mathbf{w}^T x^{(i)} + w_0 < 0 \end{cases} \quad \text{----(8)}$$

w is the weight vector and w_0 is the bias. There may exists many separators; dividing data sets, but the one which maximizes the margin between vectors of two class is optimum separator. It is called principle of maximum marginality. Hyperplane, selected is also expected to minimize outliers in separating two classes. C is regularization parameter for generalizing our classifier. A generalized classifier will have low error rate not only in training set but also on unseen test data.

.

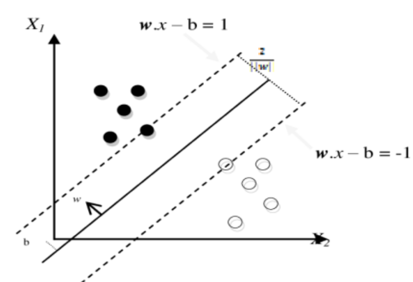


Fig 2: Decision boundary of SVM

SVM searches for two things; a hyperplane with largest minimum margin and a hyperplane that accurately separates maximum data points. It is difficult to achieve both objectives; hence value of parameter C of SVM tries to balance and tune the model. Lower value of C generates maximum margin, but outliers may occur as presented in

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Fig3. However, higher values of C will minimize outliers with minimum margin.

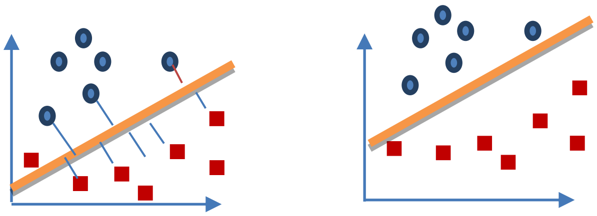


Fig 3(a)Maximum Margin with outliers low C values (b) Reduced misclassification rate with higher C values.

Random Forest(RF)

It is a robust ML technique used for regression and classification.It is an ensemble model, made of large number of small decision tree called estimators. Each estimator produces their own results; which are combined to obtainoverall accurate predictions [2].Forest is an ensemble of decision trees; trained with *bagging* method[1]. Idea of bagging method is that, combination of learning models increases the overall result. Each tree is built on random set of features and it uses subset of these feature for splitting a node. Importance of a feature is measured by reduction in impurity of a node using that feature. It removes the overfitting problem of decision tree and generalizes well in case of unseen data. It works well with large, heterogenous dataset with high dimensionality. Let us take a training set with N examples and pick-up sample repeatedly from subsets of training set of size k; k<N. Sampling is random with replacement and is called bootstrapping or bagging. If number of features is M, a subset m of M is used for training each estimator. Each estimator is trained with only m feature of k training data set. Estimators are number of decision trees used for creating a forest and each estimator is trained on sampled feature and training data. Final outcome is obtained by voting method where output of each estimator is compared with others; predictions given by maximum trees is considered as final outcome.

Hyper Parameters of Random forest: Set of parameters are tuned to obtain the best result. Some of the important hyperparameters are:

- *Number of estimators*: Number of trees algorithm uses for building forest before recording maximum voting or average prediction score. Higher number of trees; better is the performance of algorithm at higher computation cost.
- *Maximum_features:* Maximum number of features used by algorithm to split a node
- *Minimum number of leaves* required to split a node
- *No of jobs*: Number of processors used by algorithm. -1 value indicates any number of processors can be used.
- *Random state:* For a definite value of random state and same training data with same other hyperparameters; algorithm will output same result.
- *Out of Bag samples(oob):* About one-third of training sample is used for validation; which helps to generalize performance of the model. It is cross validation for evaluating the model

RF model is basically updated version of decision tree; which is more robust and handles the problem of overfitting well. 11K data collected from Tripadvisor.com is used for building model after preprocessing and feature extraction. Build model is tested and evaluated and results are discussed in next section.

4. Result and Discussion

Data set is divided into 80-20 ratio for training and testing. Experiment is conducted with following features:

- Number of words
- Number of characters
- tfidf
- Word2vec
- Doc2vec (with vector size =100)

System performance is evaluated by building confusion matrix and finding accuracy, F measure, precision and recall. Precision measures fraction of relevant instances among total retrieve instances. Recall, also known as sensitivity, represents the fraction of relevant instances retrieved among all relevant set. F-Score is calculated as given in Eq 3 and accuracy as per Eq 4

F_Score=2*(Recall*Precision)/(Recall+ Precision)------(3)
Accuracy=(TP+TN)/(TP+FP+FN+TN)------------------(4)_

Where TP,FP,TN and FN is defined as follows:
True positive: +ve review correctly identified as +ve
False positive: +ve review incorrectly identified as -ve
True negative: -ve review correctly identified as -ve
False negative: -ve review incorrectly identified as +ve

Table2: Comparative performance of SVM and RF

S. No	Algorithm	Accuracy	F Score	Precision	Recall

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

1	SVM (rbf)	73.1	0.785	0.798	0.788
1	SVM (linear)	82.3	0.8959	0.8965	0.8956
2	RF	84.48	0.8736	0.8830	0.8727

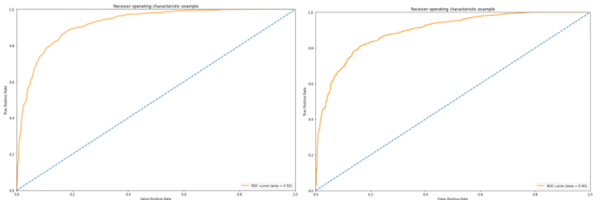


Fig 3: ROC curve with SVM and Random forest (a) AUC=0.92 (b)AUC-0.90

False negative: -ve review incorrectly identified as +ve

Computational complexity of SVM is much higher than random forest. SVM takes longer time to train as compared to RF as optimization of hyperplane gets expensive. Between two flavors of SVM; linear SVM performs better than SVM with rbf kernel due to linear nature of data. Regularization parameter, C value is tuned to improve result. Lower value of parameter(C=1) exhibits better results as it generates optimum hyperplane with higher margin. Decision boundary of RF is crisper and handles outliers in a better way; thereby improving classification rate. When experiment was conducted with basic feature set and low vector size; linear SVM was slightly better than RF. Enhancing vector size increased feature set and Rf outperformed SVM.

Receiver operating Characteristic (ROC) curve is useful tool in understanding accuracy of a predictive modelTrue positive rate (TPR)is plotted against false positive rate (FPR).TPR is fraction of samples that were correctly predicted to be positive out of all positive observation. FPR is fraction of samples that are incorrectly predicted to be positive out of all negative samples. Farther the curve is from the diagonal line; better is the model in prediction.As it is shown in Fig 3; SVM has the best ROC curve as it is most deflected away from diagonal as it generates optimum model parameters.

5. Conclusion and Future Scope

We have experimented on tourism data set extracted from Tripadvisor.com and reported the results of three supervised leaning algorithms. Doc2vec and word2vec was used to obtain vector representation of each review which improved the results of SVM and RF classifiers. Dataset can be enhanced to include reviews with double negation, sarcasm and phrases. Accuracy can be further improved by using word embeddings generated from larger corpus; so that system can be trained with higher number features.

Acknowledgements

This work is funded under the project " Opinion mining of tourism sector of India using ML algorithmfrom reviews collected through social media" funded by TEQIP III , CSVTU, Bhilai, under the collaborative research project.

Reference

1 Alex Davies ZoubinGhahramaniThe Random Forest Kernel and creating other kernels for big data from random partitions https://arxiv.org/abs/1402.4293 Feb 2014

2 Alexander Statnikov1, Lily Wang2 and Constantin F Aliferis* "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification" BMC Bioinformatics 2008, 9:319 doi:10.1186/1471-2105-9-319

3 Barman D, Chowdhury N,(2020)"A novel semi supervised approach for text classification ",International Journal of Information Technology vol 12 pp 1147-1157

4 Boiy, E., Hens, P., Deschacht, K., and Moens, M. F.(2007)Automatic sentiment analysis of on-line text. In Proceedings of the 11th International Conference on Electronic Publishing (Vienna, Austria).

5 Cambria E, Poria S, Gelbukh A, and Thelwall M,(2017)"Sentiment Analysis Is a Big Suitcase," IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, 2017.

6 Dave, K., Lawrence, S., and Pennock, D. M.(2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the 12th internationalWWW conference (Budapest, Hungary, May 20–24). 2003, 519–528.

7 Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining WSDM '08 (pp. 231–240). New York, NY, USA: ACM. doi: 10.1145/ 1341531.1341561 .

8 Edgar A, Sidarta R, Mariano S, and. Diego F. S, (2016)"Comparative study of LSA vsWord2vec embeddings in small corpora: a case study in dreams database," in ASAI SimposioArgentino de Inteligencia Artificial.

9 F. Enr´ıquez, J. A. Troyano, and T. L´opez-Solaz,(2016) "An approach to the use of word embeddings in an opinion classification task,"Expert Systems with Applications, vol. 66, pp. 1–6, 2016.DOI: 10.1016/j.cities.2019.03.019

10 Hailong Z, Wenyan G, and Bo J, (2014)"Machine learning and lexicon basedmethods for sentiment classification:Asurvey," in Proceedings of the 11th Web Information System and Application Conference, WISA 2014, pp. 262–265, China, September 2014.DOI: 10.1109/WISA.2014.55

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

11 Hu M and Liu B(2004), "Mining and summarizing customer reviews,"in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), pp.168–177, August 2004.https://doi.org/10.1145/1014052.1014073

12 Jadav S, Tanawal B, and Guadani H (2017) "Sentiment analysis:a review," International Journal of Advance Engineering and Research Development, vol. 4, pp. 957–962, 2017.2017. 10.21090/IJAERD.75190

13 Jo,Yohan.,Oh,Alice.(2011).Aspect and sentiment unification model for online review analysis. Conference: Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China.DOI: 10.1145/1935826.1935932

14 Kamps J, Marx M, Moken R.J, and M. De Rijke(2004), "Using WordNet to measure semantic orientations of adjectives," in Proceedings of the 4th International Conference on LanguageResources and Evaluation, LREC 2004, pp. 1115–1118, Portugal,May 2004.

15 Li X, Xie H, Chen L, Wang J, and Deng X, (2014)"News impact on stock price return via sentiment analysis," Knowledge-Based Systems, vol. 69, no. 1, pp. 14–23, 2014.https://doi.org/10.1016/j.ins.2014.03.096

16 Li X, Cao J, and Pan Z, (2018)"Market impact analysis via deep learned architectures," Neural Computing and Applications, pp.1–12, 2018.31

17 Liu B(2012) "Sentiment analysis and opinion mining," Morgan & Claypool, 2012.(https://doi.org/10.2200/S00416ED1V01Y201204HLT016)

18 Marzijarani SB, Sajedi H. (2020)Opinion mining with reviews summarization based on clustering. International journal of Information Technology Sep 10 2020

19 Maynard D.and Funk A, (2011)"Automatic detection of political opinions in tweets," in Proceedings of the 1st Workshop on Making Sense ofMicroposts 2011: BigThings Come in Small Packages, MSM 2011 - Co-located with the 8th Extended Semantic Web Conference, ESWC 2011, pp. 81–92, Greece, May .

20 Medhat W, Hassan A, and Korashy H,(2014) "Sentiment analysis algorithms and applications: a survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.https://doi.org/10.1016/j.asej.2014.04.011

21 Mika V.Mäntylä DanielGraziotin MiikkaKuutila Computer Science Review, Elsevier,Volume 27, February 2018, Pages 16-32 https://doi.org/10.1016/j.cosrev.2017.10.002

22 MudinasAndrius,ZhangDell,Levene Mark.(2012)Combining lexicon and learning based approaches for concept level sentiment analysis,WISDOM ,China.https://doi.org/10.1145/2346676.2346681

23 Nasukawa T and Yi, J(2003) "Sentiment analysis:Capturingfavorability using natural language processing," in Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003,pp. 70–77,USA, October 2003.https://doi.org/10.1145/945645.945658

24 Ortigosa A, Mart´ın J.M., and Carro R.M.(2014)"Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, vol. 31, no. 1, pp. 527–541, 2014.DOI: 10.1016/j.chb.2013.05.024

25 Pang,B.,&Lee,L.(2004).A sentimental education on:Sentimentanalysisusing Subjectivity summarization based on minimum cuts. Proceedings of the 42nd annual meeting on Association for Computational Linguistics (pp.271-279). ACL doi:10.3115/1218955.1218990

26 Pang,B.,Lee,L.,&Vaithyanathan,S.(2002).Thumps up?:Sentiment classification using machine learning techniques.proceedings of ACL-02 conference on Empirical methods in Natural Language Processing.(Vol-10,pp79-86) doi: doi:10.3115/1118693.1118704

27 Qiu G, He X, Zhang F, Shi Y, Bu J, and Chen C(2010), "DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis,"Expert Systems with Applications, vol. 37, no. 9, pp. 6182–6191, 2010https://doi.org/10.1155/2018/9839432

28 Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification.Proceedings of the ACL Student Research Workshop (pp. 43-48). Association for Computational Linguistics. doi:10.3115/1628960.1628969.

28 Sindhwani V and Melville P (2008)"Document-word co-regularization for semi-supervised sentiment analysis," in Proceedings of the 8th IEEE International Conference on Data Mining, ICDM 2008, pp. 1025–1030, Italy, December 2008.DOI: 10.1109/ICDM.2008.113

30 Tripathi,A.,Rath,sk,. (2013)Classification of Sentiment of Reviews using Supervised Machine Learning Techniques International Journal of Rough Sets and Data Analysis Volume 4 • Issue 1 • January-March 2017DOI: 10.4018/IJRSDA.2017010104

31 Turney P.D(2002), "Thumbs up or thumbs down?" in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424, Philadelphia, Pennsylvania, July 2002.https://doi.org/10.3115/1073083.1073153

32 Yang A, Lin J, Zhou Y, and Chen J,(2013) "Research on building a Chinese sentiment lexicon based on SO-PMI," AppliedMechanicsandMaterials, vol. 263-266, no. 1, pp. 1688–1693 https://doi.org/10.4028/www.scientific.net/AMM.263-266.1688

33 YassineAlAmraniMohamedLazaarKamal EddineElKadiri" Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis"Procedia Computer Science. Elsevier
Volume 127, 2018, Pages 511-520 https://doi.org/10.1016/j.procs.2018.01.150

34 Zhang Z, Ye Q, Zhang Z, and Li Y,(2010) "Sentiment classification of internet restaurant reviews written in cantonese," Expert Systems with Applicaions, vol. 38, no. 6, pp. 7674–7682, DOI: 10.1016/j.eswa.2010.12.147

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731