CSIT (December 2016) 4(2–4):241–247
DOI 10.1007/s40012-016-0107-y

SPECIAL ISSUE REDSET 2016 OF CSIT

A review of sentiment analysis techniques for opinionated web text

Jaspreet Singh[1] · **Gurvinder Singh**[1] · **Rajinder Singh**[1]

Published online: 16 December 2016

Abstract Social Media nowadays generate huge loads of data that can be valuable in many contexts. It includes media of all formats by which groups of users interact to generate ideas in a distributed and networked process. Data scientists from Twitter have found that the main reason for attaining fame of Presidential Candidate in the upcoming elections scheduled in Nov, 2016 in US is the reach of social media. Researchers and data scientists can use data on social media to track opinions of people about products and services. Many approaches are working behind the scene to reduce errors in opinion mining and sentiment analysis and to attain a level of accuracy for meeting the growing demands of organizations to evaluate their customers. The way people express their opinions have radically changed in the past few years. This paper explores various techniques of distillation of knowledge from huge amount of unstructured information. Generic features of making use of linguistic patterns in sentiment classification are being explored in this paper. In this study investigation of all opinion extraction techniques to generate positive and negative aspects of data with appropriate feature set can help in reduction of error of misclassification.

Keywords Linear discriminant analysis (LDA) · Support vector machines (SVM) · Back propagation networks (BPN) · Homogenous ensemble neural networks (HENN) · Probabilistic neural networks (PNN) · Gaussian mixture model (GMM) · Hidden Markov model (HMM) · LEX · Scalable distance clustering (SDC)

✉ Gurvinder Singh
gurvinder4371@gmail.com

✉ Rajinder Singh
tovirk@yahoo.com

Jaspreet Singh
profjaspreetbatth@gmail.com

[1] Department of Computer Science, Guru Nanak Dev University, Amritsar 143001, Punjab, India

1 Introduction

User generated media like discussion forums, blogs, online reviews and other web contents are of great interest in opinion mining nowadays. Researchers are trying to analyze the power of social media in data science. Data from social networking media is highly unstructured text. Data scientists these days are seeking to develop numerous data services and tools to structure and analyze hidden information from social data. This need raises the thirst of current academic research to focus on language processing and sentiment analysis [7]. Ambiguity in the content and highly context dependent data of social media presents many challenges in knowledge discovery process. One of the most active forums in India, 'HP Inc. Forum' has more than 50 thousand daily users submitting their posts online. There are more than 500 million of subscribers in Sina Microblog in China [19]. Various machine learning techniques are used to extract useful knowledge for web forum content mining and web opinion clustering [21].

2 Opinion mining and sentiment analysis approaches

Opinion mining and sentiment analysis approaches can be classified into three levels of extraction namely, aspect or feature level, sentence level and document level. Further

two categories of techniques are used: (i) Machine learning based techniques (ii) Lexicon based techniques. Machine learning based techniques are basically applied at aspect and sentence levels of feature extraction. Features of these techniques include uni-grams, bi-grams, n-grams, POS tags and bag-of-words. SVM, Naïve Bayes and Maximum Entropy are three flavors of machine learning at aspect and sentence levels of feature extraction. Lexicon based or corpus based techniques use decision trees, SMO, k-NN, HMM, CRF and SDC based methodologies for sentiment classification [2, 4–6, 9]. Following are the approaches for opinion classification and sentiment extraction from opinionated web text.

2.1 Linear discriminant analysis (LDA)

Li et al. [11] found a way of dimensionality reduction of feature matrix like Principle Component Analysis (PCA). LDA searches data in all directions that have large variances and then gradually project it to diminish some. Fisher LDA considers minimization of covariance matrix J(w).

$$J(w) = \frac{\mathrm{WT}\ .\mathrm{SB}(w)}{WT.SW(w)} \tag{1}$$

where SB is 'between classes scatter matrix' and SW is 'within classes scatter matrix'. If x is overall mean of data cases and U_c is the mean of class c, then definition of scatter matrices are:

$$\mathrm{SB} = \sum (U_c - x)\ (U_c - x)^T \tag{2}$$

$$\mathrm{SW} = \sum \sum\ (x_i - U_c)\ (x_i - U_c)^T \tag{3}$$

Data from online reviews is being broken into tokens. Tokens may correspond to nouns, verbs, adverbs or adjectives. Their sentiment or opinion values are pre defined whether they are used in negative sense or positive sense. Out of all positively opinionated reviews, their xi opinion values are collected in a scatter matrix for LDA. Above two equations are used to define their matrices.

2.2 Support vector machines (SVM)

Bhadane et al. [3] presented a new approach of sentiment analysis and opinion mining in Science Direct in 2015 using SVM with 0.78 precision. SVM is a supervised learning technique based on decision planes with decision boundaries. According to the membership of different objects these decision boundaries separates classes. Greed of sentence or tweet taken from twitter is considered or classified into positive, negative or neutral depends upon SVM training algorithm. There are four SVM models. Two are of classification and other two for regression SVM. Training of these SVM techniques minimize error function:

$$E = 1/2\ W^T E\ + C\ \sum \beta_i \tag{4}$$

where C is constant and β is parameter for non separable data input. W is a vector coefficient. SVM uses a kernel function to map input feature space into new space where classes are linearly separable. It uses polynomial kernel which largely depends upon cache size, exponent, tolerance and numFolds. Vectors from nouns, verbs, adverbs and adjectives are made with their coefficient taken from defined values of their sentiments. Values less than 0.5 reflects negative influence whereas values greater than 0.5 gives positive sentiment and coefficients with 0.5 value are meant for neutral class.

2.3 Back propagation neural network

Vindhoni and Chandrasekaran [18] from Dept.of Computer Science and Engg. Anamalai University performed sentiment classification of online reviews using BPN. BPN is an adaptive learning technique with a capability of classification of sentiments from social data. For each training pattern, the inputs are applied to the network. Neurons at nodes of first layers are firstly trained with weights of defined range. Then at hidden layer various combinations of bigrams and tri-grams appearing in sentences are considered. With the applications of these varying combinations, weights get adjusted while reaching to the output layer. Use of error function signals to compute weight adjustments. According to number of nouns, verbs and adjectives present in sentences, number of neurons in three layers varies.

2.4 Probalistic neural networks (PNN)

Savchenko [15] have used PNN in recognition of discrete patterns from sets. In PNN, Gaussion Kernel function is used in the hidden layer of neural network. Its third layer was used to perform average operation of outputs for each review class. In fourth layer, final class belongingness is found by selecting largest value of class label.

2.5 Homogeneous ensemble neural network (HENN)

Su et al. [17] introduced Ensemble of learning for sentiment classification in 2013. They have used Chinese Lexical Semantics to combine predictions of multiple base models. Mixture of base models with re-sampling of training data by calling base classifiers. Interactively drawing sub samples of training data and then by

combining majority voted classes will give best possible prediction of classification.

2.6 Gaussian mixture model (GMM)

Abdel Fatteh [1] proposed "Multiple Classifiers for sentiment Analysis". in Neurocomping Elsevier Journal in 2015. GMM are used for clustering data by allocating query data points to the multivariate normal components. Assigning data points to clusters is termed as hard clustering. Power of GMM clustering can be noted because it uses soft clustering techniques. They include assignment of score to data point for each cluster.

2.7 Naïve Bayes (Bayessian networks) maximum entropy

In journal of Theoretical and Applied Information Technology, J. Jotheeswaran and Dr. Y, S Kumarswamy presented a paper on opinion using NB classifier and data set taken from Manhattan Hierarchical Cluster in 2013. NB classifier model is a directed acyclic graph whose nodes carry variables and edges contains conditional dependencies. In text classification for sentiment analysis BN is found to be very expensive.

2.8 Hidden Markov model (HMM)

L. R. Rabiner presented a tutorial on HMM and applications in speech recognition in IEEE proceedings in (1989). HMM is a classification technique which is used for putting the right label on any sequence of nodes either from biological terms or from linguistics.HMM basically associate different lexemes into chain of nodes. While processing this model of different nodes taken from online reviews, it has to go through from one state to another and path between states is noted. Depending upon overall sentiment of sentence or group of statements taken from review, AMM provides different chains, these chains are known as Markov chains and further the classification can be done through Naïve Bayes classifier.

2.9 Decision trees

Jaskarn and Shveta in 2012 presented "Analysis and identification of Human Emotions through Data Mining", published in IJCA. It is a hierarchical based classifier gives decomposition of training data space where value of some dividing attribute is used to divide data. Division of data items or phrases in case of text mining is done recursively so that last leaf nodes contain tokens for classification.

2.10 Sequential mining optimization (SMO)

Vivek et al. has given survey of various classification techniques in IJCA in Dec-2015. G. Geetika and Y. Divakar presented a paper on sentiment analysis of twitter data using SMO in IEEE International Conference 2014. SMO is used to optimize classification processes when training SVM's. It interactively breaks larger sentences into smaller phrase to tokens and then classifies these tokens according to boundary value analysis applied by SVM.

2.11 K-nearest neighbour classifier (KNN)

Pak and Paroubek [13], presented a paper titles "Twitter as a corpus for sentiment analysis and opinion mining" in IJLRC in May 2010, have talked about KNN while comparing SVM with other classifiers. It uses three types of distance functions namely Euclidean, Manhattan and Minkowski for finding gap between two terms under classification process. In this process a case is classified by using most likeliness of its neighboring values, the case being allocated to the class with amongst its K nearest neighbors identified by one of above distances.

2.12 Jaccard similarity

Mrunmayee et al., have provided a sentiment analysis tool using jaccard and cosine similarity. In this classification technique firstly sentence is tokenized and then its word root is found after removing unwanted nouns and verbs. Keywords from sentences are extracted in case of test mining. Term frequencies of each keyword from a document are found in the next step. Similarity between two terms can be found by using Jaccard's relation

$$J(A, B) = |A \cap B|/|AUB| \quad (5)$$

2.13 Lexicon based opinion classifier (LEX)

T. Christopher and KG. Nanda presented a survey in which combined classification approaches of lex and KNN with ME(maximum entropy) are discussed. In LEX based classifier polarity prediction and product features are identified with entity ranking of lexemes. Accuracy of lex alone was found to be 50.08% while mix of ME and KNN its accuracy gets improved to 80.21% observed by T. Christopher.

2.14 Conditional random fields (CRF)

Hu and Liu [8] proposed a method for aspect extraction from online reviews that can be treated in sentiment analysis. The first step of this task is to mark up words from text with" The Stanford Log-Linear Parts of Speech Tagger". Then in second step various nouns are extracted from tagged corpus. In the third step "Porter Stemming Algorithm'' is used to remove words with lesser influence in polarity i.e. positive or negative conditions are looked up manually from squeezed data.

2.15 Scalable distance clustering (SDC)

SDC is a distance based algorithm proposed by Christopher C and Tobun Dorbin in Nov, 2011 which stressed that required density of words must be accumulated in initial clusters [17]. Words from text are clustered with initial densities defined. Then modifications in their distances are making noise filtering process along with cluster iteration and thereby growing the cluster further.

3 Review of different sentiment analysis approaches

Most important feature requirement in opinion mining and sentiment analysis is correct identification of positive and negative words depicting the real greed of author of the text. The advantage of machine learning based approaches over lexicon based approaches is that, former can attain desired level of accuracy by training the network of bag of words [12]. While in latter sentiment extraction is complex and slow due to large growing size of corpus and diversity of linguistic terms. In the table below is the analysis on the basis of features, advantages and limitations of various Opinion Mining and Sentiment Analysis techniques (Tables 1 and 2).

Table 1 Review of sentiment analysis approaches

Sr. no.	Opinion mining/sentiment analysis	Features	Advantages	Limitations
1.	Linear discriminant analysis (LDA)	Between classes and within class scatter matrix	Dimensionality reduction, minimization of covariance	Complex and slow
2.	Support vector machines (SVM)	Negative, positive and neutral words and phrases	Minimise error function	Not good for non-separable words
3.	Back propagation neural network (BPN)	Auto updation of opinion words with combinations	Accuracy (0.78)	Slow and iterative
4.	Probabilistic neural network (PNN)	Term frequency and parts of speech	Ability to adopt complex texts	Low applicability to new data
5.	Homogenous ensemble neural network (HENN)	Similar text, negations, label symbols	Ability to create trained models	Availability of labelled data could be costly
6.	Gaussian mixture model (GMM)	Soft clustering, multivariate normal components	Multiple classifier for sentiment Analysis	multi-variables may cause deflection of results
7.	Naïve Bayes (NB)	Maximum entropy, frequency	Capable of aspect extraction at context level	Used for specific purposes only
8.	Hidden Markov model (HMM)	Label phrases, chain of phrases	Good for sentence level as well as paragraph level SA	Markov chains may introduce ambiguity in opinions
9.	Decision trees	Hierarchical tagging of text	Division till tokens gives better understanding	Slow and complex
10.	Sequential mining optimization (SMO)	Interactive, sentiment words as features in ML	Fast learning symbiosis	Noisy reviews
11.	K-Nearest neighbour (K-NN)	Manhattan, minkowski distance	Accuracy (0.81)	Sensitive to score of words
12.	Jaccard similarity	Tokenization, word root, term frequency	Similarity of two positive or negative influential words are considered	Sensitive to the choice of data set
13.	Lexicon based opinion classifier (LEX)	Corpus based or dictionary based	wider term coverage	Finite no. of words in lexicons and fixed sentiment orientation
14.	Conditional random fields (CRF)	Standford log-linear parts of speech tagger	Random fields arbitrarily chosen, favour requirements	Unable to find implicit aspects
15.	Scalable distance clustering (SDC)	Initial clusters required	Filter noise by removing non influential words	Lesser sensitivity to changes in topic domain

4 Recent papers exploring sentiment analysis

Most of the business organizations today believe that their success lies in the satisfaction of their customers. Also, there are the plethora of product and services reviews available on the web. So business organizations encourage young researchers and academicians for sentiment analysis and opinion extraction of their web text. Here are some recent research papers exploring new insights of web text for opinion mining and sentiment analysis.

Table 2 Review of recent papers on sentiment analysis

Title, author, publication	Dataset	Features	Tools/techniques used	Classification approach
Title: ranking product aspects through sentiment analysis of online reviews Authors: Wie Wang, Hongwei Wang and Yuan Song Publication: Taylor and Francis journal of experimental and theoretical research (2016)	Amazon dataset of 386 digital cameras for 39 months	Product aspects, review vectors, log-sales-rank, avg-rating	Seni-Strength, TF-IDF algorithm, HAC algorithm [20]	SVM, porter stemming approach
Title: Joint multi-grain topic sentiment: modeling semantic aspects for online reviews Authors: Md. Hijbul Alam, Woo-Jong Ryu and Sankeun Lee Publication: Elsevier, (2016)	Hotel-reviews from *tripadvisor.com* and Restaurant reviews from *citysearch.com*	Word-of-mouth	JMTS [2]	Gibbs sampling approach
Title: A comparative performance evaluation of neural network based approach for sentiment classification of online reviews Authors: G. Vinodhini & R.M. Chandrasekaran Publication: Elsevier, (2016)	1200 product reviews of cameras from Amazon	Unigram, Bigram, Trigram [18]	BPN, PPN, Homogeneous Ensemble NN	PCA
Title: Sentiment Analysis and visualization of customer reviews Authors: A.V. Gundla & M.S. Otari. Publication: IJESRT (Dec-2015)	Reviews of various e-commerce websites	POS tagging[6]	HAP, Proposed SA-API in Java	Graph based summary generation approach
Title: Reassessing the facebook experiment: critical thinking about the validity of big data research Authors: Galen Panger Publication: Taylor and Francis, (Oct-2015)	Facebook posts and News Feed manually collected	Positive and Negative words of News Feed	(Kramer, Guillory & Hancock) KGH Experiment	LIWC [14]
Title: Predicting the performance of online consumer's reviews: A sentiment mining approach to big data analytics Authors: Dan J. Kim, Mohammed Salehan Publication: Elsevier, (2015)	Electronic product reviews from *amazon.com* and restaurant reviews from *yelp.com*	Numeric star rating, length of reviews, word count, longevity, readership and helpfulness of reviews	Senti-strength tool	Regression analysis

Table 2 continued

Title, author, publication	Dataset	Features	Tools/techniques used	Classification approach
Title: SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection Authors: F.H. Khan, U qamar & S. Bashir. Publication: Elsevier, (2015)	Public movie review dataset of 50,000 movie reviews	SynsetScore, ObjScore, POS Tags and their ranks	Senti-MI [10]	Semi-Supervised Learning Step, Random Walk Step
Title: Twitter financial community sentiment and its predictive relationship to stock market movement Authors: Sreve Y. Yang, S.Y. Kevin Mo & Anki Liu Publication: Taylor and Francis, (2015)	1.6 Lac tweet messages corpus from 15-feb-2014 to 15-june-2014	BC Score, DC Score, Unigram, Bigram, Trigram [22]	Proposed Sentiment Analysis Algorithm	Linear Regression Analysis
Title: A supervised fine grained sentiment analysis system for online reviews Authors: Hanxio Shi, Wenping Zhan & Xiaojun Li Publication: Taylor and Francis, (2015)	Chinese hotel reviews corpus.	POS tags	CRF Model [16]	NLP Corpus based.
Title: Sentiment Analysis: Measuring Opinions Authors: Chetashri Bhadane, Hardi Dalal & Heenal Doshi Publication: Science Direct, (2015)	1940 reviews corpus manually annotated	N.grams, conjunction rules, POS tagging, Stop words	Lexicon based proposed Technique with 78% accuracy [3]	SVM
Title: A rule based approach to emotion cause detection for Chinese microblogs Authors: Kai Gao, Hua Xu & Jiushuo Wanga Publication: Elsevier, (2015)	Corpus of Chinese Microblogs	NER, POS Tags [4]	Association Rule based Technique	Bayesian Probability approach, SVM, SMO
Title: Extracting Aspects and Mining opinions in product reviews using supervised learning Algorithm Authors: A. Jeyapriya & C.S. Kanimozhi Selv. Publication: IEEE, (2015)	Twitter Corpus manually annotated	POS Tags [9]	Frequent Itemset Mining.	Naïve Bayes
Title: Twitter sentiment mining framework based learner's emotional state classification and visualization for e-learning Authors: M. Ravichandran & G. Kulanthaivel Publication: Journal of theoretical and applied information Technology, (2014)	Twitter Dataset (Stanford Gold dataset of 498 tweets)	Maximum Entropy, Bigrams	BIRT (Bigram Item Response Theory) Proposed Model	SVM, Lexicon based sentiment polarity

5 Conclusion

Several opinion mining techniques are adopted to evaluate the real greed of user generated data over social media. Dictionary based or corpus based techniques are more accurate in mining opinionated texts, while machine learning techniques are yet to improve their error rates. LDA, SVM, GMM, HMM, Jaccard's similarity and K-NN are the approaches which are quite near to real picture. These techniques are continuously working for the analysis of online data, that how much level these are able to satisfy the thrust for data over the social media. Several support vector machines are assessing the positive and negative aspects of the online data which is being posted. This assessment is basically done using certain training algorithms. Sentiments depend upon the certain range of values of features such as bi-grams and tri-grams with their polarities and also on their combinations. Their effects are slow and iterative and nature. So proceeding further to work on the hidden layer of neural network kernel function is being applied which computes the belongingness of class label. The conditional dependencies between the various nodes and edges of an acyclic graph is done with the help of Bayessian networks, which is helpful in extraction of data at the context level. For the good sentiment analysis of sentence as well as paragraphs Hidden Markov model is applied. The optimization of sentences and words leads to faster learning which generates accuracy of data on social media. Tokenization of data at word root level, helps to generate positive and negative aspects of data. All the approaches are working hard to reduce the errors in opinion mining and sentiment analysis to achieve better level of accurate data for social media. All in all, this paper focuses on the various sentiments analysis techniques for extraction of structured data from unstructured web text.

References

1. Abdel Fattah M (2015) New term weighting schemes with combination of multiple classifiers for sentiment analysis. Neurocomputing 167:434–442. doi:10.1016/j.neucom.2015.04.051
2. Alam MH, Ryu W, Lee S (2016) Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Inf Sci 339:206–223. doi:10.1016/j.ins.2016.01.013
3. Bhadane C, Dalal H, Doshi H (2015) Sentiment analysis: measuring opinions. Procedia Comput Sci 45:808–814. doi:10.1016/j.procs.2015.03.159
4. Gao K, Xu H, Wang J (2015) A rule-based approach to emotion cause detection for Chinese micro-blogs. Expert Syst Appl 42(9):4517–4528. doi:10.1016/j.eswa.2015.01.064
5. Gautam G, Yadav D (2014) Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: 2014 Seventh International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2014.6897213
6. Gundla AV, Otari PM (2015) A review on sentiment analysis and visualization of customer reviews. Int J Eng Comput Sci. doi:10.18535/ijecs/v4i10.11
7. Haenlein M, Kaplan AM (2010) An Empirical analysis of attitudinal and behavioral reactions toward the abandonment of unprofitable customer relationships. J Relat Mark 9(4):200–228. doi:10.1080/15332667.2010.522474
8. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '04. doi:10.1145/1014052.1014073
9. Jeyapriya A, Selvi CS (2015) Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In: 2015 2nd International Conference on Electronics and Communication Systems (ICECS). doi:10.1109/ecs.2015.7124967
10. Khan FH, Qamar U, Bashir S (2016) SentiMI: introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection. Appl Soft Comput 39:140–153. doi:10.1016/j.asoc.2015.11.016
11. Li Tao, Zhu Shenghuo, Ogihara Mitsunori (2008) Text categorizationvia generalized discriminant analysis. Inf Process Manage 44:1684–1697
12. Liu B (2011) Opinion mining and sentiment analysis. Web Data Mining 459–526. doi:10.1007/978-3-642-19460-3_11
13. Pak A, Paroubek P (2011) Twitter for sentiment analysis: when language resources are not available. In: 2011 22nd International Workshop on Database and Expert Systems Applications. doi:10.1109/dexa.2011.86
14. Panger G (2015) Reassessing the facebook experiment: critical thinking about the validity of Big Data research. Inf Commun Soc 19(8):1108–1126. doi:10.1080/1369118x.2015.1093525
15. Savchenko AV (2013) Probabilistic neural network with homogeneity testing in recognition of discrete patterns set. Neural Netw 46:227–241
16. Shi H, Zhan W, Li X (2015) A supervised fine-grained sentiment analysis system for online reviews. Intell Autom Soft Comput 21(4):589–605. doi:10.1080/10798587.2015.1012830
17. Su Y, Zhang Y, Ji D, Wang Y, Wu H (2013) Ensemble learning for sentiment classification, Chinese lexical semantics. Springer, Berlin, pp 84–93
18. Vinodhini G, Chandrasekaran R (2016) A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. J King Saud Univ—Comput Inf Sci 28(1):2–12. doi:10.1016/j.jksuci.2014.03.024
19. Wang H (2013) ReTweeting analysis and prediction in microblogs: an epidemic inspired approach. China Commun 10(3):13–24. doi:10.1109/cc.2013.6488827
20. Wang W, Wang H, Song Y (2016) Ranking product aspects through sentiment analysis of online reviews. J Exp Theor Artif Intell 1–20. doi:10.1080/0952813x.2015.1132270
21. Yang CC, Dorbin Ng T (2011) Analyzing and visualizing web opinion development and social interactions with density-based clustering. IEEE Trans Syst Man Cybern—Part A Syst Hum 41(6):1144–1155. doi:10.1109/tsmca.2011.2113334
22. Yang SY, Liu A, Mo SY (2014) Twitter financial community modeling using agent based simulation. In: 2014 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr). doi:10.1109/cifer.2014.6924055