I J C T A, 9(17) 2016, pp. 8607-8613

Sentiment Analysis: A Critical Review

Neha Dwivedi[*] and Sunita Yadav[*]

ABSTRACT

In present trend, people share their opinions on the online websites available and the researchers, to collectively review these opinions, are suggesting methods to automate the task and classify the reviews to be positive or negative. This research field is termed as sentiment classification or opinion mining. This field helps in easy evaluation of the bulk of data produced by millions of reviewers. These opinions are important as they directly reflect the end user response on the particular topic or field. In this aspect, this paper surveys some of the latest techniques and approaches done by researchers in last few years and highlighting their key contributions.

Keywords: Sentiment Analysis, Opinion Mining, SVM, Naïve Bayes, LDA, Homogeneous Ensemble(HEN)

I. INTRODUCTION

The terms Opining Mining, Subjectivity Analysis and Sentiment Analysis are considered to be synonyms, though the origin of the terms is not exactly same. The term opinion can be defined as a view, emotion, attitude, positive or negative sentiment, an user's aspect or group of user's aspect about an entity or an appraisal about an entity. Entity can be a person, topic, product, event or organization. Mathematically, a 5-tuple –

$$(e_j; a_{jk}; so_{ijkl}; h_i; t_l)$$

defines the term opinion where a target entity is represented by e_j and the e_j entity's k-th feature/aspect is represented by a_{jk}, the opinion sentiment value as per the opinion holder h_i about an e_j entity's aspect a_{jk} at time t_l is represented by so_{ijkl}. The value can be positive, neutral or negative, or even can be any other granular rating form which can be used. The tuple – h_i, t_l is used for the opinion holder and the time at which the opinion was expressed [1]. The opinions are classified into various groups as comparative and regular opinions. Mostly, the opinions expressed are regular and can be further subdivided into indirect and direct opinions. The opinions expressed about an entity or about any of the entity aspect based on its effect on other entities are referred as indirect opinions whereas ideas about an entity or about an aspect of an entity are referred as direct opinions. The comparative sentences are used to express the resemblance among two entities taking into consideration their common features or aspects [2,3,4]. Opinions are also classified into implicit or explicit opinions based on the subjective or objective ideas expressed by a particular opinion [5].

Other than opinion and sentiment, emotion and subjectivity are the two concepts, close and important to these concepts. A subjective sentence expresses personal views, beliefs or feelings but not necessarily a sentiment whereas some factual information of the world is expressed by objective sentences. "They went away" is a sentence whereas "I love this chocolate" is an objective sentence expressing the speaker love for a particular chocolate.

The rest of the paper is organized as follows. Tasks involved in sentiment analysis and classifiers used are explained in section II and section III. Concluding remarks are given in section IV.

[*] Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, U.P., India, *Emails: nehadwivedi15@gmail.com; yadav.sunita104@gmail.com*

II. TASK INVOLVED

The section overviews the commonly used steps in the sentiment classification approaches as shown in the figure. The approach overviewed is the popular bag-of-words model. In this model, vector is used to represent the document and the entries in the vector represent the individual terms of the vocabulary. The steps are briefly described below:

2.1. Data pre-processing

Data pre-processing is an important step in sentiment analysis as this improves the text classification performance [6]. It involves sub-processes like tokenization, transformation and filtering of stop words as per a specific language– English, French etc. The removal of stop word likes 'the' and 'a' i.e. articles and prepositions are filtered out. The stemming process reduces the term variations into a single representation. These sub-processes are done to reduce the ambiguity. The steps outputs the bag of words or unordered collection of words.

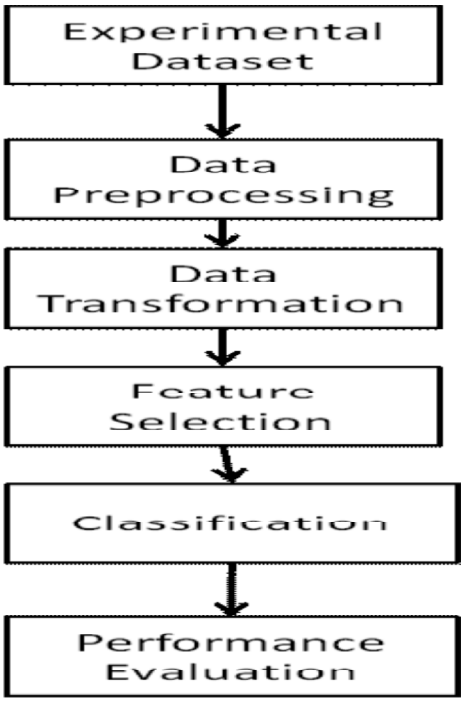

Figure 1: Steps in Sentiment Analysis

2.2. Data Transformation

By transformation, we convert the textual data to numerical representation. Commonly, the numerical representation used is binary representation in which we account for the presence or absence of the term. The other representation is maintaining the count of the number of times, a term is present in the particular document (referred as term frequency). To evaluate the importance of the each term in the document, we use TF-IDF (Term Frequency – Inverse Document Frequency), a popular numerical representation of the term.

$$TF\_IDF_{t,d} = TF_{t,d} \times IDF_t, \text{ where } IDF_t = \log \frac{N}{DF_t}$$

In the above formula, the $TF_{t,d}$ represents the occurrence number of the term t in document d, N represent the document number in the collection considered, DF_t represents the document number having the term, t. TF-IDF avoids assigning high score to the terms.

2.3. Feature Selection

The next stage in this model is feature selection. This stage is significant as by feature selection the amount of data is reduced resulting in improving the performance of classifiers. The usual methods for feature selection are document frequency, information gain, chi-square and mutual information. Among all, information gain is widely used. Information gain considers the presence or absence of the term in each class to rank each term as follows.

$$IG(t) = \sum_{k=1}^{C} P(C_k) \log \frac{1}{P(C_k)} - \sum_{t \in \{t_p, \bar{t_p}} P(t) \sum_{k=1}^{C} P(t \mid C_k) \log \frac{1}{P(t \mid C_k)}$$

where the probability that a class c_k has a particular document is given by $P(c_k)$, the probability for a term occurring in the document or nor is given by P(t) i.e. $P(t_p)$ and NOT $P(t_p)$ and the conditional probability for the term t occurring in a document of class c_k is $P(t_j|c_k)$ and the number of classes is given by C.

2.4. Classification

After refining the features in the feature selection stage, the refined features are input to the classification/ learning process. The classification of the reviews is done by classification techniques like Support Vector Machine, Naïve Bayes, Artifical Neural Network etc. Some classifiers are explained in the following section.

2.5. Performance evaluation

The evaluation of the performance of the classifiers along with the other complete approach used in the sentiment analysis can be done using the confusion matrices, overall accuracy (OA), popular indexes F1, Precision, Recall. The confusion matrix is the tabular representation of the results obtained from the classifiers. It is formed for the dataset with known inputs i.e. when the output is already known to the user. The tabular form is shown below:

Table 3
The Confusion Matrix

	Predicted positives	*Predicted negatives*
Actual positive examples	Total True Positive example number (TP)	Total False Negative example number (FN)
Actual negative examples	Total False Positive example number (FP)	Total True Negative example number (TN)

Using the confusion matrix, the accuracy, precision and recall can be also to calculated to evaluate the system performance as–

$$overall\ accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

III. CLASSIFIERS USED

The review of fundamental aspects of various supervised classifiers is discussed in this section. Some popular supervised classifiers are Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network– back propagation network and probabilistic neural network, linear discriminant analysis (LDA). The basic concepts of the popular classifiers are discussed here. Naive bayes classifier has low computational cost in comparison with SVM. The SVM and LDA are the statistical based approaches which can be implemented in rapid miner tool. The neural networks can be implemented in Matlab.

3.1. Naive Bayes

Naive bayes works on the assumption that terms occur independently in the selected context considering the probability of each term. It is a learning method working on these assumptions. It takes into account the collection of N document $\{d_j\}^n_{j=1}$, each document has a sequence of T terms, $d_j = \{t_1, t_2, \ldots\ldots, t_T\}$, the documents are classified under classes and each document probability in a class c_k is calculated by :

$$P\left(C_k \mid d_j\right) = P(C_k)\prod_{i=1}^{T} P(t_i \mid C_k),$$

where $P(t_i|c_k)$ is the conditional probability of term t_i occurring in a document of class c_k and $P(c_k)$ is the prior probability of a document occurring in class c_k. The above probability terms, $P(t_i|c_k)$ and $P(c_k)$ are estimated from the training data taken during the experiment.

3.2. Linear discriminant analysis (LDA)

LDA is a data classification method being used in many application domains. It functions by calculating a rule in order to classify the reviews, taken as data source in the experiment, as positive or negative by minimizing the probability of misclassification. It is suitable for the data set in which the within class frequencies are unequal. LDA focuses in maximizing the between-class variance and within class variance ratio in the data set thereby assuring the maximal separability in the classified reviews [7].

3.3. Support Vector Machines

Support Vector Machines is a popular learning technique under supervised approaches having many desirable and highly efficient qualities. It has a firm theoretical foundation performing classification accurately when compared with many other algorithms. It had been reported by many researchers about SVM to be the most accurate text classification technique [8] and so is used widely in opining mining [9]. SVM is a linear learning method which finds an optimal hyper plane in order to separate two classes. To obtain better classification/ generalization performance while testing with the data, it seeks to maintain maximum distance from the closest training points (known as support vectors fig. 2(a)) of each class. The SVM gives solution based on the training points which are at margin of the decision boundary. In SVM, the convex optimization problem is solved to obtain the optimal separating hyper plane parameters. In this, the global error function in gradient descent process is not minimized rather the focus is on the optimal parameters. When the input data space cannot be separated linearly, the data space is transformed into feature space of higher-dimensional so as to make the data space separable and suited for linear SVM formulation. Commonly, for such transformation the kernel function h is used [10]. This results in determining a nonlinear decision boundary in a high dimensional feature space with no focus on computing the optimal hyper plane parameters.

3.4. Artificial Neural Networks

Neural networks, centrally, derives features of the input data from the linear combinations and then models a nonlinear function of the features as output [12]. The result obtained from this classifier is one of the most

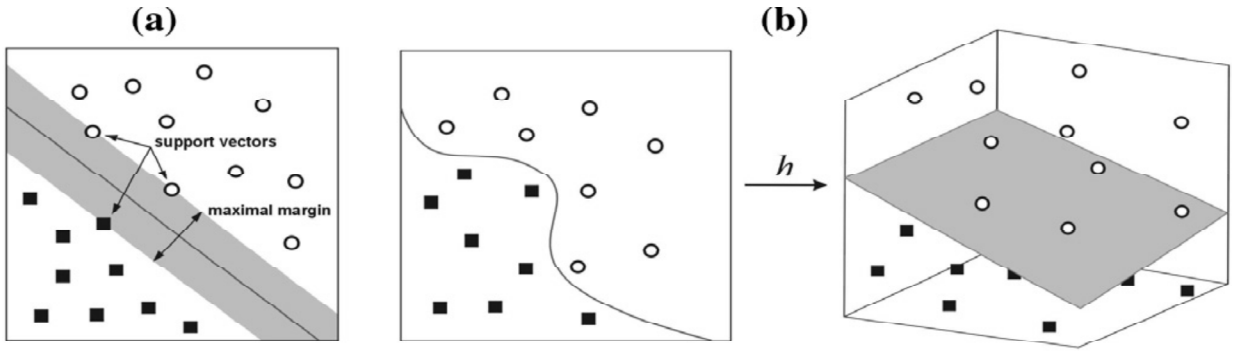

Figure 2: (a) Linear SVM (b)Nonlinear SVM [11]

effective and popular learning system forms [13]. Network diagrams are used to represent neural networks. The network diagram is made of nodes which are connected by the direct links, these nodes are arranged in layer-wise manner. Mostly, it has three layers– input, hidden and output layer. The neural network is classified as feed-forward network as the connection of the nodes is unidirectional. Each connection is weighted connection and the values of the weights are estimated using the gradient descent process minimizing the global error function. In simple terms, a neuron is a mathematical model producing results in two steps– first, it computes the weighted sum of the given inputs and then by using an activation function to these sum in order to derive the output [13]. The activation function is a nonlinear function which ensures that the entire network can be estimated using this function.

3.4.1. Back propagation neural network (BPN)

The neural networks have various advantages as parallelism, generalization, adaptive learning and fault tolerance. Generally, the neural networks are divided into feed-forward networks and feedback networks. The feed-forward network provides superior classification. The best known feed-forward network is back-propagation network which is among the most useful ones. The published papers [14-16] and the thumb rule restricts the hidden layer to be one or two. Forward pass and backward pass are included in the back propagation algorithm. The iterative gradient algorithm designed in order to reduce the mean square error between the multilayer feed-forward perceptron actual output layer and the desired output. The activation value is obtained from the forward pass and the biases and the weights are adjusted on the basis of the actual network outputs and the desired outputs in the backward pass. The two passes are iteratively continued until the convergence of the network. The back-propagation pseudo-code algorithm for training the feed-forward network is as follows: [17].

Step 1. For each training pattern (as presented in random order):

Step 1.1. the inputs are applied to the network.

Step 1.2. The output for every neuron at the input layer is calculated, through the hidden layer(s), to the output layer.

Step 1.3. The error is calculated at the outputs.

Step 1.4. Using the output error, the error signals for pre-output layers are computed.

Step 1.5. The error signals are used to compute weight adjustments.

Step 1.6. Weight adjustments are applied.

Step 2. The network performance are periodically evaluated.

3.4.2. Probabilistic neural network (PNN)–

The probabilistic neural network is a type of statistical Bayesian classification algorithm. The functioning of the PNN is separated into a feed-forward network of multiple layers. Specifically, these networks involve four layers– input layer, patter layer, summation layer and output layer. The input layer has input nodes involving the measurement set; the input layer is connected fully by the pattern layer. The pattern layer has training set having neurons for each pattern involved. The output of the pattern layer is connected selectively to the summation layer based on the pattern classes. The steps involved in the PNN model can be summarized as [18].

Step 1. The neurons in the input layer distribute the measurements involved to all the pattern layer neurons.

Step 2. Using the given data points set, Gaussian kernel function is formed and is included in the second layer.

Step 3. The third layer calculates the operation of averaging the outputs for each review class.

Step 4. Voting, selection of largest value is performed and then the class label is determined in the fourth layer.

The Matlab tool box can be used to implement the PNN.

3.5. Homogeneous ensemble (HEN)

The multiple base model predictions are combined using the ensemble methods. The base models are created by resampling the training data. The base classifiers of similar types are integrated using the homogeneous ensemble method. The base learners are trained each from a different sample of bootstrap using the base learning algorithm call in the homogeneous ensemble method. By sub sampling the training data set and with replacement, the bootstrap samples are obtained and the formed samples have the same size as that of the training data set. Once the base learners are obtained, the majority voting is used to combine the ensemble method. The majority voted class can be predicted [19]. The HEN approach design is shown below.

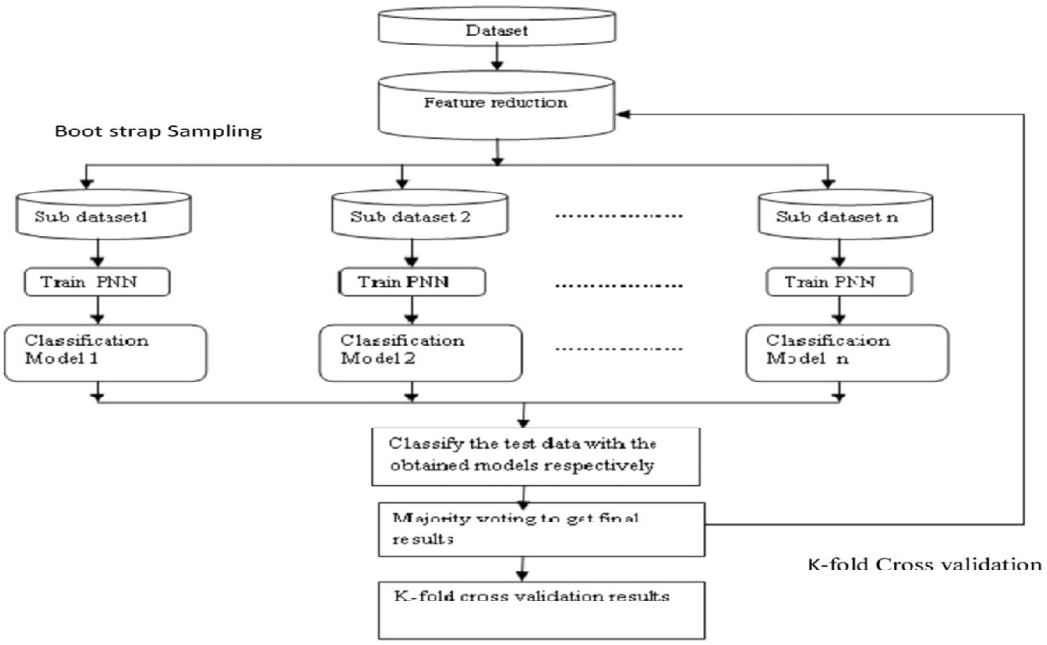

Figure 3: Design of HEN approach [20]

IV. CONCLUSION

In the above sections, the approaches and classifiers are discussed. The approach used is a general method but there exist various classifiers. Among all the classifiers used by various researchers, SVM performs the best. SVM has many benefits as it is robust in the most widely considered datasets and areas and the most of the text classification issues are independent linearly. SVM obtained efficient results in opinion mining and the response of SVM is overwhelming in comparison with other machine learning methods. SVM is a novel technique based on machine learning approaches and the statistical learning concept. It resolves major issues occurring in the ANN classifier like the over-fitting issue, low convergence ratio and local optimal solution. Conversely, for using the SVM, it is necessary for the user to have complete knowledge of the SVM as the parameters to be set for efficient results needs knowledge and the optimization results vary with the parameters.

REFERENCES

[1] B. Liu, Sentiment analysis and subjectivity, Handbook Nat. Lang. Process. 5 (1) (2010) 1–38.

[2] N. Jindal, B. Liu, Identifying comparative sentences in text documents, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '06, ACM Press, New York, New York, USA, 2006, pp. 244–251.

[3] N. Jindal, B. Liu, Mining comparative sentences and relations, in: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06), 2006b, pp. 1331–1336.

[4] S. Yang, Y. Ko, Finding relevant features for Korean comparative sentence extraction, Pattern Recogn. Lett. 32 (2) (2011) 293–296.

[5] B. Liu, Exploring Hyperlinks, Contents, and Usage Data, Springer, 2011.

[6] Salton, G., Singhal, A., Mitra, M., Buckley, C., 1997. Automatic text structuring and summarization. Inf. Process. Manage. 33 (2), 193–207.

[7] Li, Tao, Zhu, Shenghuo, Ogihara, Mitsunori, 2008. Text categorization via generalized discriminant analysis. Inf. Process.Manage. 44, 1684–1697.

[8] Li, S., Wang, Z., Zhou, G., & Lee, S.Y.M. (2011a). Semi-supervised learning for imbalanced sentiment classification. In Proceedings of international joint conference on artificial intelligence (pp. 1826–1831).

[9] Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web.Data Mining and Knowledge Discovery, 24, 478–514.

[10] Huang, T. M., Kecman, V., & Kopriva, I. (2006). Kernel based algorithms for mining huge data sets: Supervised, semi-supervised, and unsupervised learning. Studies in computational intelligence (Vol. 17). Secaucus, NJ, USA: Springer.

[11] Rodrigo Moraes, Joao Francisco Valiati, Wilson P. Gaviao Neto (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications 40 (2013) 621–633

[12] Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning.New York, NY, USA: Springer New York Inc.

[13] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: A modern approach.Prentice Hall.

[14] Su, C.-T., Hsu, J.-H., & Tsai, C.-H. (2002). Knowledge mining from trained neural network. Journal of Computer Information Systems, 61–70.

[15] Chen, L.-F., Su, C.-T., & Chen, M.-H. (2009). A neural-network approach for defect recognition in TFT-LCD photolithography process. IEEE Transactions on Electronics Packaging Manufacturing, 32(1), 1–8.

[16] Su, C.-T., Yang, T., & Ke, C.-M. (2002). A neural-network approach for semiconductor wafer post-sawing inspection. IEEE Transactions on Semiconductor Manufacturing, 15(2), 260–266.

[17] Chen, L.-S., Hsu, C.-C., & Chen, M.-C. (2009). Customer segmentation and classification from blogs by using data mining: An example of VOIP phone. Cybernetics & Systems, 40(7), 608–632.

[18] Savchenko, A.V., 2013. Probabilistic neural network with homogeneity testing in recognition of discrete patterns set. Neural Networks 46, 227–241.

[19] Su, Y., Zhang, Y., Ji, D., Wang, Y., Wu, H., 2013. Ensemble Learning for Sentiment Classification, Chinese Lexical Semantics. Springer, pp. 84–93.

[20] G. Vinodhini , R.M. Chandrasekaran. A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. Journal of King Saud University – Computer and Information Sciences (2015).