VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2014 ISSN(e): 2309-3978;ISSN(p): 2411-6246
Volume 2, Number 1, January-December, 2014

pp. 01-07

AN EMPIRICAL STUDY ON SENTIMENT POLARITY CLASSIFICATION OF BOOK REVIEWS

FAIZA SHAHZADI[1] AND TEHSEEN ZIA[2]
[1]Department of Computer Science& Information Technology University of Sargodha
Fz_noor@yahoo.com
[2]Department of Computer Science& Information Technology University of Sargodha
Tehseen.zia@uos.edu.pk

ABSTRACT. *Sent ment polarity classification deals with automatic classification of text in sentiment polarity categories. While in most of proposed approaches for polarity classification, a dictionary containing polarity-based terms is considered. Such dictionaries are not readily available. We have adopted a machine learning based approach where classifiers are trained over a self-collected corpus of book reviews, annotated with sentimental categories. In this paper, we have presented our investigation of performance evaluation of machine learning classifiers. Five classifiers are evaluated including naïve Bayes, k-nearest neighborer, decision tree and support vector machine. Naïve Bayes has shown us best results.*

***Keywords**: :* Opinion Mining; Text Categorization; Sentiment Analysis; Sentiment Classification; Book Review.

1. Introduction. Text categorization (TC) is a popular methodology to process textual data [1]. Based on contents of text, TC methods assign text to predefine categories. With the interest of automatically indexing, organizing, summarizing and searching of enormous amount of online textual data, TC is getting popularity in information retrieval (IR). While TC deals with topic-based classification of text, opinion mining is another domain of IR that process text on non topic basis. For example, detection and extraction of opinions, feelings and emotions related to a specific subject are common tasks of opinion mining. Distinguishing expressions (such as positive, negative or neutral) within text is a subtask of opinion mining particularly known as sentiment polarity classification or identification [2]. In most of proposed polarity classification approaches, polarity related terms (such as wonderful, terrific, beautiful and bad, etc) are considered to classify sentiment polarity. These solutions though works well, the availability of dictionary of polarity-related terms is a key requirement. It has two issues with it: firstly, identifying such polarity-related terms and building a dictionary is a challenging task (such dictionaries are not readily available). Secondly, due to its reliance over specific collection of terms, it's not easy to adopt these methods. Therefore, we have analyzed the suitability of TC approach for sentiment polarity classification of text. It is more convenient since it's easy to gather opinions and annotate them with polarity categories rather building a dictionary of polarity-related terms. Secondly, it has convenience of easy adoptability.

The prototype of sentiment polarity system has broader scope for the end users who are often interested to get more information about the book before purchasing: most of the websites maintain the opinions of people who already have comments on given books. These opinions can be utilized as a valuables source of information for the purchasing the desired book. The sentiment polarity system can be integrated into any website facilitate the customer in decision making process. The processing of prototype sentiment polarity system will effectively take the customer reviews as input, classify the review as positive or negative the outcome of system in calculated form, how many percentage (positive, negative) reviews against a book. A block diagram of prototype application is shown in Figure 1.

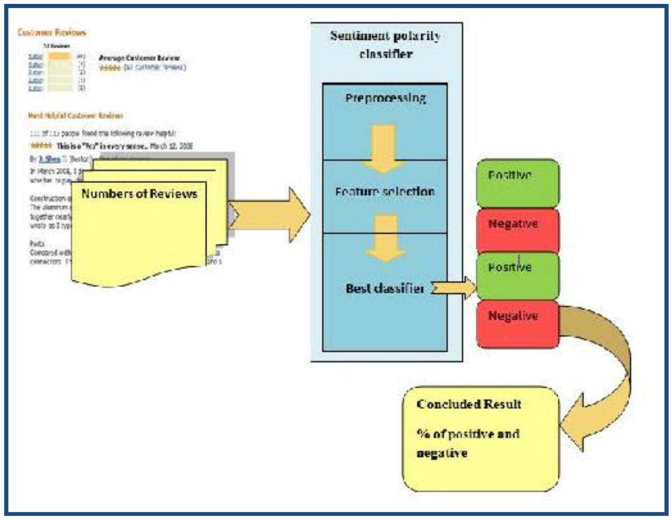

Figure 1: Prototype application.

In this paper, we have presented performance evaluation of popular machine learning algorithms over a self-collected book reviews' corpus. The algorithms considered for analysis include naive Bays, k-nearest neighbors, decision tree, support vector machines. The paper is organized as: in Section 2 we have described literature review. A short introduction of classifiers is provided in Section 3. The methodology is described in Section 4. Results are presented in Section 5. Finally, the conclusion and future work in given in Section 6.

2. Literature Review. Opinion mining and sentiment polarity identification is an active area of research among information retrieval and natural language processing communities since last few years. Different methods have been proposed to identify sentiment polarity of text. Majority of these methods can be classified into two broad approaches: rule based approach and machine learning approach. In rule based approach [16], a dictionary of polarity related terms (such as underlined in Figure 1 and 2) is employed to distinguish between for example positive and negative polarities. While rule based approach works at term –level, machine learning approach can work on document-level [17], sentence-level [18], phrase-level [19]. The methods that have followed machine learning based approach can further be sub-categorized into supervised learning and unsupervised learning approaches. For example, Bo Pang and Lillian lee, shivakumar vaithyanathan [17] have used supervised learning approach where classifiers such as Naive Bays, Maximum Entropy and Support Vector Machine are used to distinguish polarity of product reviews and movie reviews respectively. This work is at the document level. Similar kinds of techniques have been utilized to perform the domain specific sentiment analysis for financial blogs [15]. To perform classification decisions they have trained Naive Bayes, and Support Vector Machine classifiers. On the other hand, Choi Y, Kim Y, Myaeng SH (2009) have employed unsupervised learning approach where K-Mean clustering algorithm is used to cluster news articles into positive, negative and natural clusters based on frequency of polarity-related words, noun pronoun adjective [14]. A similar work was performed in the context of weblogs in order to extract the multiple topics form weblogs [16]. Hidden Markov model was used for this purpose.

In this paper, we have also adopted supervised machine learning approach. We have comparatively analyzed the performance of four popular classifiers for text categorization over a self-collected corpus of book reviews. This paper explores the detailed analysis of classifier in the methodology section.

3. Classification Algorithms. We have posed sentiment polarity classification of book reviews as a supervised machine learning problem: classification algorithms are used to classify the review among sentimental categories. To assess the performance of classification algorithms for sentiment predicting, four algorithms are evaluated: naïve Bayes (NB), k-nearest neighbors (KNN), decision tree (DT) and support vector machines (SVM). Input to these algorithms is a book review and output is a sentimental category. In NB, the task of classifying textual data is posed as estimation of posterior probabilities of categories given data [4]. KNN is an example of instance based learning method where category decision of unknown instance is made based on voting of training instances despite learning an explicit function [5]. In DT, hypothesis (or classification function) is represented in tree form where nodes represent features, edges correspond to feature values and leafs correspond to categories [6]. SVM is based on structural risk minimization principle where objective is to find a hypothesis that guarantee lowest true error [7]. By representing complexity of hypothesis as Vapnik-Chervonenkic (VC) dimension, SVM minimizes true error of hypothesis by efficiently controlling the VC dimension.

4. Data Collection and Performance Evaluation. Since we have opted machine learning methodology for detecting hidden sentiments within text, availability of data is a key resource. We have manually collected and annotated book reviews from Amazon (www.amazon.com). The corpus contains 652 book reviews: including 334 positive and 318 negative reviews. A sample of positive review is given in Figure 1.

In order to evaluate the performances of classification algorithms we have used standard macroaveragred F1 measure as define below:

$$\mathrm{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Where recall and precision are defined as:

$$\text{recall} = \frac{\text{\# of correct positive predictions}}{\text{\# of positive examples}}$$

$$\text{precision} = \frac{\text{\# of correct positive predictions}}{\text{\# of positive predictions}}$$

This book is a wealth of information for people studying Computer Science. The materials are laid-out in an easy-to-follow format and the explanations are clear and concise (for those with a mind for this type of material). The only problem with this type of book is that it refers to an ever-evolving type of science and may quickly become obsolete, but I would not allow that to deter others from using this book to gain some solid foundations for understanding the principles of CompSci.

Figure 2: A sample positive book review

The book exercises are confusing and contain mistakes.
The content is educational and interesting.
It would be helpful if next editions contain answers or similar solutions are provided, in order to avoid errors and misunderstandings.

Figure 3: A sample negative book review

The experimentation of this work is performed in WEKA software (acronym of "Waikato Environment for Knowledge Analysis"). WEKA is an open source machine learning software that provides state-of-art machine learning techniques for data analysis, pre-processing, classification, clustering, etc [8].The choice of

WEKA is due to its built-in functionalities for TC such as tokenization, stop words removal, attribution selection, feature weighting, classification and performance evaluation.

5. Model Selection for Classifiers. Model selection is the process to choose optimal classifier parameters from the existing set of data. This is often done by analyzing the performance of classifiers over different parameter values. In this section, we have described performance of classifiers over different parameter values for classifiers.

5.1. Naive Bayes. Number of studies has shown that feature selection is a critical preprocessing task for some classifiers such as naïve Bayes and KNN [9, 10]. The aspect of this criticality is to select appropriate number of informative features [11]. Naïve Bayes classifier essentially has only one parameter to be optimized. In Table 1, we have shown results of performance of naive Bayes classifier over different number of top ranked features. The feature selection of performed in a way that we have ranked features based on their effectiveness measured using information gain. The reason to use information gain is its promising performance among other feature selection methods [12]. The classifier is tested over top ranked 100, 200, 500, 1000 and 5000 features. The results have shown that top 100 features have high precision=0.784, recall=0.782 and F- Measure=0.782.

Table 1: Naive Bays classifier on different attribute

Attributes	Precision	Recall	F-Measure
100	**0.784**	**0.782**	**0.782**
200	0.766	0.764	0.764
500	0.772	0.770	0.770
1000	0.775	0.773	0.773
5000	0.726	0.725	0.725
8533	0.701	0.701	0.701

5.2. K- Nearest Neighbor (KNN). KNN essentially involves two parameters that require optimization: k (number of nearest neighborer) and number of attributes [9]. We have analyzed performance of classifier over following values of k and number of attributes.

$$K: 1,2,5,10,20,30$$

$$Number\ of\ attributes: 100, 200, 500, 1000, 5000, 8533$$

However, not all the combination of values is analysed. We have initially optimized the values of k and then by using the value of k we have optimized number of attributes as recommended in [8]. The classifier performance over values of k is shown in Table.2 where bold value shows best result. Results of classifier performance with respect to number of attributes are shown in Table 3.

Table 2: Performance of KNN with respect to K

K	Precision	Recall	F-Measure
K=1	**0.664**	**0.655**	**0.652**
K=2	0.607	0.584	0.568
K=5	0.571	0.544	0.511
K=10	0.549	0.515	0.447
K=20	0.604	0.518	0.415
K=30	0.627	0.537	0.452

Table 3: Performance of KNN with respect to number of features

Attributes	Precision	Recall	F-measure
100	**0.697**	**0.676**	**0.670**
200	0.666	0.640	0.628

500	0.654	0.638	0.631
1000	0651	.624	0.611
5000	0.627	0.607	0.596
8533	0.664	0.655	0.652

5.3. SVM with Polynomial kernel. It has essentially two parameters that require to be fixed: degree of polynomial(d) and $number\ of\ attributes$. Following values are tested for these parameters:

$$d = 1, 2, 3, 4, 5$$

$$Number\ of\ attributes: 100, 200, 500, 1000, 5000, 8533$$

In Table 5, we have shown performance of classifier given various degrees of polynomials and it can be seen that polynomial with degree 1 has outperformed the rest. By fixing the degree of polynomial, SVM is tested over various levels of features. The results are shown in Table 6.

Table 5: SVM apply on the same attributes for different values of polynomial

Polynomial degree	Precision	Recall	F-measure
1	**0.741**	**0.741**	**0.741**
2	0.719	0.718	0.718
3	0.682	0.683	0.682
4	0.660	.653	0.651
5	0.662	.635	0.623

Table 6: SVM apply on the different attributes taking values of polynomial=1

Attributes	Precision	Recall	F- Measure
100	0.800	0.796	0.796
200	0.796	0.794	0.794
500	**0.802**	**0.798**	**0.797**
1000	0.802	0.796	0.795
5000	0.774	0.771	0.771
8533	.741	0.741	0.741

5.4. SVM with RBF Kernel. The support vector machine with radial base kernel has two parameters: gamma (γ) and number of features. Following values for these parameters are tested:

$$\gamma = 0.01, 0.02, 0.03, 0.04, 0.05$$

$$Number\ of\ attributes: 100, 200, 500, 1000, 5000, 8533$$

In Table 7, we have shown performance results of classifier over γ values. Using optimal value of γ, performance of classifier over different feature levels is shown in Table 8.

Table 7: SVM Radial Base function for different values of polynomial exponent

Gamma	Precision	Recall	F- Measure
0.01	**0.770**	**0.765**	**0.765**
0.02	0.747	0.747	0.747
0.03	0.740	.738	0.737
0.04	0.714	0.709	0.706
0.05	0.690	0.676	0.668

Table 8: SVM Radial Base Function for different values of attributes

Attributes	Precision	Recall	F- Measure
100	0.763	0.748	0.746
200	0.779	0.767	0.765
500	**0.795**	**0.782**	**0.781**
1000	0.798	0.781	0.778
5000	0.786	0.765	0.762
8533	0.770	0.765	0.765

6. Result. The performance of classifiers for optimal values of parameters is shown in Table 9. It can be seen that naïve Bayes has outperformed the other. On the other hand, KNN has shown worse performance among other. Based on this analysis an application (as illustrated in Figure 1) will be developed in future for providing recommendations to users for purchasing books.

Table 9: Best results of Nave Bays, KNN, Decision Tree, SVM (polynomial, RBF)

Classifier	Precision	Recall	F-Measure
Naive Bays	**0.784**	**0.782**	**0.782**
K Nearest Neighbor	0.697	0.676	0.670
Decision Tree	0.681	0.681	0.680
SVM Polynomial	0.741	0.741	0.741
SVM RBF	0.795	0.782	0.781

REFERENCES

[1] Sebastiani, F. (2002), Machine Learning in Automated Text Categorization, *jounrnal,* ACM Computing Surveys (CSUR) , vol.34, no.1, pp: 1-47.

[2] Waltinger, U.,(2011), An Empirical Study on Machine Learning Based Sentiment Classification using Polarity Clues, L*ecture Notes in Business Information Processing*, Vol. 75, pp: 202-214,

[3] Peter D. Turney and Michael L. Littman,(2002), Unsupervised learning of semantic orientation from a hundred-billion-word corpus.*NRC/ERB-1094. CoRR, cs.LG/*0212012 .

[4] Mitchell, T. M.,(1997), Machine Learning, *McGraw-Hill, New York*, pp230-235.

[5] Dasarathy, B.V.,(1995), Nearest neighbour (nn) norms: Nn pattern classification technique, *IEEE Computer Society Press , Pattern Recognition Letters* ,Pages 809–814.

[6] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers Inc*. San Francisco, CA, USA.

[7] Burges, C., J., C., (1998), A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2, 121-167.

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, L. H. ,(2009). The Weka Data Mining Software: An Update. *SIGKKD Exploration*, 11(1), 10-18.

[9] Lewis, D, D., Yang, Y, G. Ross, T. Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5, 361-397.

[10] Joachims, T. (1998). Text Catagorization with Support Vector Machines: Learning with many Relevant Features,*Tenth European Conference on Machine Learning* (ECML-98), 137-142.

[11] Rogati, M., Yang, Y.,(2002) ,High-Performing Feature Selection for Text Classification, *Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02*, pp: 659-661.

[12] Yang, Y., Pedersen, J. (1997), A Comparative Study on Feature Selection in Text Categorization. *International Conference on Machine Learning*, 412-420.

[13] Choi Y, Kim Y, Myaeng SH (2009) ,Domain-specific sentiment analysis using contextual feature generation, In: *Proceeding of the international CIKM workshop on topic-sentiment analysis for mass opinion measurement. ACM, New York, NY, USA.*

[14] F Aisopos, G Papadakis, K Tserpes,(2012) ,Textual and contextual patterns for sentiment analysis over microblogs" *Proceedings of the 21st international conference companion on World Wide Web ACM New York, NY, USA.*

[15] N O'Hare, M Davy, A Bermingham, P Ferguson,(2009) Topic-Dependent Sentiment Analysis of Financial Blogs" *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion ACM New York, NY, USA.*

[16] Q Mei, X Ling, M Wondra, H Su, CX Zhai,(2007), Topic Sentiment Mixture:Modeling Facets and Opinions in Weblogs, *Proceedings of the 16th international conference on World Wide Web ACM New York, NY, USA.*

[17] B Pang, L Lee, S Vaithyanathan, (2002) ,Thumbs up? Sentiment Classification using Machine Learning Techniques, *proceedings of EMNLP '02 Proceedings of the ACL-02 conference on natural language processing* - Volume 10 Association for Computational Linguistics Stroudsburg, PA, USA.

[18] Khan, A, Baharudin, B. Khan, K, (2010), Sentence Based Sentiment Classification from Online Customer Reviews, *Proceedings of the 8th International Conference on Frontiers of Information Technolog*,. Article No. 25.

[19] A GO, Bhayani, R.; Huang, L;(2009), Twitter Sentiment Classification using Distant Supervision, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.