Poem Classification Using Machine Learning Approach

Vipin Kumar and Sonajharia Minz

Abstract The collection of poems is ever increasing on the Internet. Therefore, classification of poems is an important task along with their labels. The work in this paper is aimed to find the best classification algorithms among the K-nearest neighbor (KNN), Naïve Bayesian (NB) and Support Vector Machine (SVM) with reduced features. Information Gain Ratio is used for feature selection. The results show that SVM has maximum accuracy (93.25 %) using 20 % top ranked features.

Keywords Poem · Classification · Ranked feature

1 Introduction

A poem is a piece of writing in which the expression of feelings and ideas is given intensity by particular attention to diction (sometimes involving rhyme), rhythm, and imagery [5]. It is generally meant to deliver expressions such as love, happiness, success, etc. Thus, poems of many category are available. However, an effort in automatic poem classification is rare. Usually, poems are as short textual paragraphs with little discriminative value of word features for automatic classification. Therefore, poem classification is a challenging task. Some text classification algorithms have been developed to categorize news [6], patent [7], etc. Many machine learning algorithms have been attempted for automatic text classification.

A poet may use any word for the poem. In fact, poems are often structured differently from normal text documents. Therefore, there is a necessity to identify an effective machine learning algorithm for poem classification. Three machine learning

V. Kumar (✉) · S. Minz
JNU, New Delhi, India
e-mail: rt.vipink@gmail.com

S. Minz
e-mail: sona.minz@gmail.com

B. V. Babu et al. (eds.), *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012*, Advances in Intelligent Systems and Computing 236, DOI: 10.1007/978-81-322-1602-5_72, © Springer India 2014

algorithms Naïve Bayesian (NB), k-Nearest Neighbor (KNN), and Support Vector Machine (SVM) are considered. For feature selection, Gain Ratio is used. The comparison of three machine learning algorithms with respect to poem classification is considered to identify the best suited one. The paper is structured as follows: Sect. 2 is on related work. Section 3 introduces poem data set. Implementation is described in Sect. 4. Section 5 contains experiment results and analysis. Finally, we present the conclusions and future work.

2 Related Work

Many types of statistical and machine learning algorithms are available to classify text documents such as k-Nearest Neighbor, Naïve Bayesian [8], and Support Vector Machine [9]. The effort in automatic classification of literary texts such as poetry is less. Malay poetry is classified using support vector machine. In this, Radial Basic Function (RBF) and linear kernel function are implemented to classify pantun by theme, as well as poetry and none poetry [10]. Logan and Kositsky [11] have done a comparison with the acoustic similarity technique and semantic text analysis technique for collecting lyrics from the Web to analyze artistic similarity.

3 Data Set

A collection of 400 text documents from popular sites such as *http://www.poetseer.org*, *http://www.poetry.org*, and *http://www.poemhunter.com* has been considered for experiment. More than 225 numbers of labels of poem are available on the Internet. It is difficult to consider all labels for this research. Therefore, the data set includes only 8 labels as alone, childhood, god, hope, love, success, valentine, and world. Each label has 50 text documents and each text document has nearly the same length.

4 Implementation

Poem classification framework is presented in Fig. 1. RapidMiner5 [2], an open source data mining package has been used to model the NB, k-KNN, and SVM [4]-based classifiers. The preprocessing steps include two main components such as text preprocessing and feature selection. In the text preprocessing task, all poems are input and processed by transformation of upper case to lower case, tokenization, stop word removal, and stemming using WordNet [1]. The text preprocessing task transforms poems into term-document vector, where the vector represents the weight of terms in the documents. Therefore, in feature selection the weight of each term is based on the *tf-idf* weighting scheme as shown below:

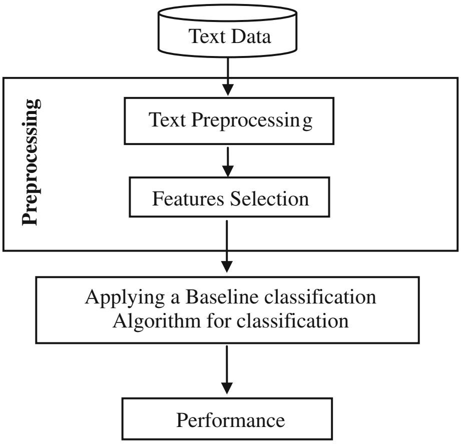

Fig. 1 Poem classification frameworks

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_j}$$

where w_{ij} is the weight of term j, N is the total number of poems, and n_j is the number of poems.

The feature selection scheme Gain Ratio (GR) is applied to the feature vector of the document to rank the features for feature reduction. It applies a kind of normalization to information gain using splitting information value defined analogously with Info(D) as

$$SplitInfo_A(D) = -\sum_{j=i}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

where D is data set into v partitions on attribute A. The Gain ratio is defined as

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

The attribute with the higher gain ratio is considered to have more relevance for classification.

After the preprocessing task, baseline classifiers KNN, NB, and SVM are applied for respective percentage of ranked features, where accuracy is a key measure to evaluate the performance of the classifiers. It can be calculated from Table 1 using the formula:

$$\text{Accuracy} = \frac{(TP + TN)}{(TO + TN + FP + FN)}$$

The true positive, true negative, false positive, and false negative are also useful in assessing the costs benefits (or risk ad gain) associated with the classification model.

Table 1 Confusion matrix for 2-class classification

Actual class	Predicted class	
	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

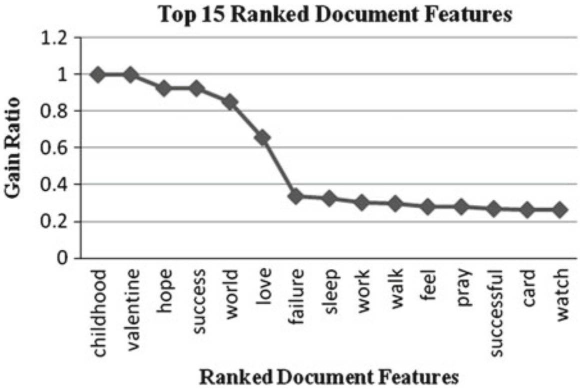

Fig. 2 Top 15 Ranked features of documents

5 Results and Analysis

The objective of the research is to find out suitable machine learning classification algorithms. NB, KNN, and SVM, three classifiers are chosen for the desired objective. Evaluation of classifiers is based on the classification accuracy. The poem data set was partitioned into training and test data. Gain Ratio was used to rank the features of the documents. Classifiers were developed using the 10, 20, 30...90 % and all features using the NB, KNN, and SVM. Ten-fold cross validation method was used to estimate the performances of all the classifiers.

5.1 Results

Gain Ratio was used to rank the document features in the experiment. To form the 3707 features, 15 top ranked features are shown in Fig. 2. It shows that childhood and Valentine features have 1 gain ratio. Hope, Success, World, Love, and Failure features are between 1.0 to 0.25 gain ratio.

Table 2 shows average accuracy of the classifiers using 10-fold cross validation with top ranked features. NB classier has maximum accuracy using 10 % top ranked features. Using 40 % top ranked features, KNN yields maximum accuracy of 87.50 %. SVM -based classifier has maximum 93.50 % accuracy by using 20 % top ranked features.

Table 2 Accuracy of NB, KNN, and SVM classifiers

% of top ranked attribute	Classification accuracy (%)		
	NB	KNN	SVM
10	**80.00**	85.50	93.00
20	79.75	86.00	**93.25**
30	76.50	84.00	93.25
40	71.25	**87.50**	92.75
50	71.25	87.00	93.25
60	68.50	83.25	92.75
70	66.25	80.00	93.00
80	66.25	76.25	92.75
90	65.75	75.50	93.00
100	79.50	85.50	92.25

Table 3 Accuracy of NB, KNN, and SVM classifiers using less than 10 % ranked features

% of top ranked attribute	Classifiers accuracy (%)		
	NB	KNN	SVM
1	**82.00**	86.00	90.00
2	81.00	85.75	90.00
3	80.00	84.00	89.75
4	80.50	86.00	91.25
5	79.50	86.26	91.50
6	79.50	85.25	92.00
7	78.50	**86.50**	91.00
8	78.00	86.00	92.50
9	78.75	85.25	92.75
10	80.00	85.50	**93.00**

From Table 2, it can be observed that SVM does not have significant difference in accuracy using 10 and 20 % top ranked features. The accuracy of KNN classifier does not have significant difference, with using 10 and 40 % top ranked features. Therefore, 10 % top ranked features are considered for the further reduced set of features.

Table 3 shows average accuracy of classifiers using 1, 2, 3... 10 % top ranked selected features. NB, KNN, and SVM have maximum accuracy of 82.00, 86.50, and 93.00 % using 1, 7 and 10 % selected features respectively.

5.2 Analysis

The objective of this research is to identify an efficient machine learning algorithm (NB, KNN, and SVM) for poem classification with reduced features. From Fig. 2, it is observed that only 6 features have Gain Ratio more than 0.6 and others have

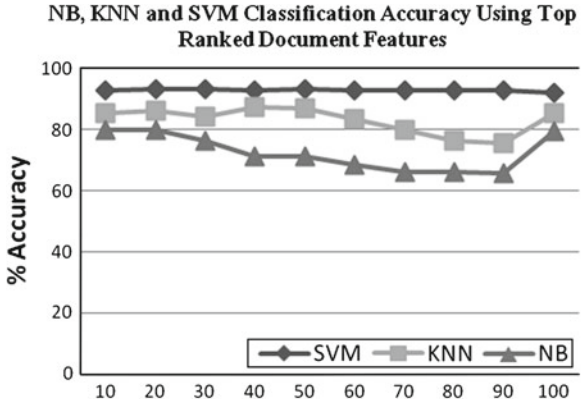

Fig. 3 NB, KNN, and SVM classification accuracy using top ranked document features

Gain Ratio less than 0.34. The top 6 features (childhood, valentine, hope, success, world, and love) are the same as labels mentioned in the data set. It means that these features (words) directly play a role in the poem classification task.

From Fig. 3 it can be easily analyzed that SVM classifier has the best performance for all percentage of ranked features and KNN has the second best performance. The performance of the classifiers displays decrease in accuracy as size of feature set increases except when 100 % features are used to model the classifiers. NB, KNN, and SVM achieved best accuracies of 80.00, 87.50, and 93.25 %, using 10, 40, and 20 % of selected features. The best performances of KNN and SVM classifiers do not indicate a significant difference using 10 % ranked features. Therefore, performance of the classifiers using 10 % ranked features has been considered for comparison of the three machine learning algorithms.

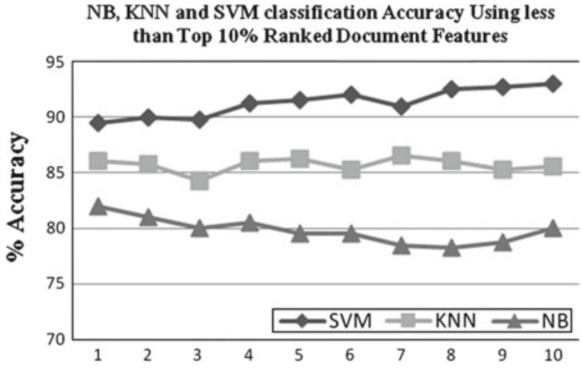

Fig. 4 NB, KNN, and SVM classification accuracy using less than 10 % top ranked document features

Figure 4 exhibits that the performance of NB-based classifiers decreases whereas the size of feature set increases. It achieves best performance 82.00 %, using 1 % of top ranked features. SVM performance is decreasing with 10 to 1 % of top ranked; therefore SVM has best performance using 10 % top ranked features. KNN has best performance 86.50 %, using 7 % of ranked features.

6 Conclusion and Future Work

A large number of poem data sets are available on Internet. Therefore, labeling poems is an important task. This study has attempted to identify an effective machine learning algorithms (NB, KNN, and SVM). The experiment results show that SVM have best performance compared to the NB and KNN. It has 93.5 % using 20 % ranked features. The top 6 features (childhood, valentine, hope, success, world, and love) have played an important role in classifying the respective poem labels.

Gain Ratio has been used for features selection of the documents. Therefore, other feature selection techniques (Information gain, Gini Index…etc) can be used in the future for features ranking. SentiWordNet [3] is a lexical resource, which gives three sentiment scores of positivity, negativity, and objectivity at each synset. Extracted document features, using SentiWordNet can play the important role for poem classification task.

References

1. Andrea, E., Sebastiani F.: SentiwordNet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, IT, 2006, pp. 417–422 (2006).
2. http://www.rapidminer.com
3. Esuli., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa (2006).
4. Chih-Chung, C., Chih-Jen, L.: LIBSVM: a library for support vector machines. In: ACM Transactions on Intelligent Systems and Technology, 2:27:127:27, (2011).
5. http://oxforddictionaries.com/definition/english/poem
6. Shih, L.K., Karger, D.R.: Learning classifiers: using URLs and table layout for web classification tasks. In: Proceedings of the 13th International Conference on World Wide Web, New York, NY, pp. 193–202 (2004).
7. Richter, G., MacFarlane, A.: The impact of metadata on the accuracy of automated patent classification. World Patent Inf. **37**(3), 13–26 (March 2005)
8. Wang, B., Zhou, S., Hu, Y.: Naive bayes-based garual Chinese documents categorization. In: Proceedings of World Multi conference on Systems, Cybernetics and Informatics, 2, July, Orlando, pp. 516–521 (2001).

9. Noraini, J., Masnizah, M.: Shahrul Azman, N.: Poetry classification using support vector machines. J. Comput. sci. **8**(9), 1441–1446 (2012)
10. Logan, B., Kositsky, A., Moreno, P.: Semantic analysis of song lyrics. In: the Proceeding of IEEE Int. Conf. on Multimedia and Expo, 2, pp. 827–830 (2004).
11. Tizhoosh, H.R., Dara, R.: On poem recognition. Pattern Anal. Appl. **9**(4), 325–338 (2006)