Automatic text classification of sports blog data

Mita K. Dalal
Information Technology Department
Sarvajanik College of Engineering & Technology
Surat, India
parikhmita@gmail.com

Mukesh A. Zaveri
Computer Engineering Department
S. V. National Institute of Technology
Surat, India
mazaveri@coed.svnit.ac.in

Abstract— **Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given text document to a set of pre-defined categories based on the features extracted from its textual content. This paper attempts to automatically classify the textual entries made by bloggers on various sports blogs, to the appropriate category of sport by following steps like pre-processing, feature extraction and naïve Bayesian classification. Empirical evaluation of this technique has resulted in a classification accuracy of approximately 87% over the test set. In addition to classifying the textual entries of sports blogs, it is proposed that the extracted features themselves be further classified under more meaningful heads which results in generation of a semantic resource that lends greater understanding to the classification task. This semantic resource can be used for data mining requirements that arise in the future.**

Keywords- automatic text classification; feature extraction; heuristics; intelligent data mining; machine learning; naïve Bayes classification

I. INTRODUCTION

Automatic Text Classification is a semi-supervised machine learning task that automatically assigns a given text document to a set of pre-defined categories based on the features extracted from its textual content.

Blogging is a popular means of expression over the Internet. In this paper we have attempted to automatically classify the textual entries made by bloggers (Internet users writing to blogs) on various sports blogs, to the appropriate category of sport. This task was accomplished by following steps like pre-processing, feature extraction and classification using the Naïve Bayesian machine learning technique. It is proposed that the feature extraction phase should be coupled with a semantic resource generation phase in which the extracted features (words and multi-words useful in classification) are further classified using a semi-supervised approach. The ready availability of such a semantic resource would be useful for intelligent data mining requests that arise in the future.

Several supervised machine learning techniques have been proposed in literature for the automatic classification of text documents such as Naïve Bayes [1] [2], Neural Networks [3], SVM (Support Vector Machine) [4] [5] etc. Techniques that combine Bayesian learning with other types of machine learning methods have also been proposed [6] [7] with good classification results. Naïve Bayesian learning is a well-known statistical and probabilistic approach to inference. The naïve Bayesian classifier [1] [2] [16] is a simplification of the basic Bayesian learning technique based on the assumption that attribute values in a training instance are conditionally independent.

This paper explores classification of sports blogs, however the technique used is generic and can be applied to the classification of any unstructured textual data. The text data of blog posts made by Internet users displays certain characteristics such as : 1) It is unstructured 2) It is usually short (for example, 50-150 words) 3)It contains frequent usage of abbreviations and slang terms 4) It often contains grammatical errors and, 5) Punctuations are inappropriately placed, making it more difficult to tokenize and identify phrases compared to regular written language.

Based on analysis of several sports blogs, it is found that the text posts of Internet bloggers are arranged in one of three ways:

1) *Pre-Classified* – Here separate web pages are allocated to each sport, so that entries are automatically classified. However, in this case the sports have to be pre-specified.

2) *Semi-Classified* - Here some sports which are popular among the visitors of the website are allocated separate pages, while the rest of the sports blogs appear in an ad-hoc manner.

3) *Un-Classified* - Here all blog entries appear as a jumble in the order in which they are posted by Internet users.

For visitors of the *Semi-Classified* and *Un-Classified* blogs it is sometimes difficult to identify the sport to which a blog entry is referring. In the long term, content management of such blogs becomes difficult and it is very difficult to mine any information from them. The classification technique used in this paper can solve the problem of content management for such blogs. The method used in this paper not only classifies the textual entries appearing on sports blogs but also adds value to the classification process by further classifying / labeling the features (extracted words and multi-words) into categories with semantic meaning like : “Common Terms”, “Players’ names” and “Tournaments’ names”.

978-1-4577-1719-2/12/$26.00 ©2012 IEEE

Several commercial applications with huge monetary profits are linked to sports. Although our current focus is on classification, the semantic resources generated along with the classification statistics can be further used for a several commercially useful, text mining applications such as: frequency analysis of blog posts to determine the popularity of a sport for marketing and advertisement, finding out popular sportspersons for brand endorsements etc.

The remainder of the paper is organized as follows. Section 2 of the paper explains the strategy for classification of sports blogs. Section 3 evaluates the strategy based on results of experimentation. Finally, Section 4 concludes the paper and provides pointer to future work in this field.

II. PROPOSED WORK: AUTOMATIC CLASSIFICATION OF SPORTS BLOG DATA

The automatic classification of unstructured sports blog data involves four major phases as shown in Fig. 1. These phases are: A) Pre-Processing Phase, B) Feature Extraction and Semantic Resource Generation Phase, C) Modeling Phase and D) Evaluation Phase. These phases are briefly explained next.

A. Pre-processing Phase

In this phase we pre-processed text by performing actions like sentence boundary determination [8], tokenization, stop-word elimination [1] [8] [9] and stemming by suffix stripping [10].

In text mining, the features are the words themselves. Thus, text mining applications have to deal with high dimensionality. The pre-processing phase is required to reduce the size of the training set. Stop-words are functional words which occur frequently in the language of the text (for example, 'a', 'is', 'the', 'an', 'in' etc. in English language), so that they are not useful for classification. Stemming is the action of reducing words to their root or base form. Moreover, in cases where the source documents are web pages, additional pre-processing is required to identify and remove HTML and other script tags [11].

B. Feature Extraction and Semantic Resource Generation Phase

In this phase we extracted important features, i.e. the words and multi-words useful in classification. Single word features were identified statistically using the TF-IDF (Term Frequency - Inverse Document Frequency) heuristic measure [12] [13] [14] while multi-word features [8] [12] [15] were extracted semi-automatically and by inspection.

In order to have sufficient and balanced number of features representing each class, the top 200 features of each category of sport were statistically computed using the TF-IDF measure and added to the feature set. The TF-IDF value of a feature is a statistical measure of its importance or relevance in the corpus. It is computed by taking the product of the term frequency (TF) and the inverse document frequency (IDF) [12] [13] [14].

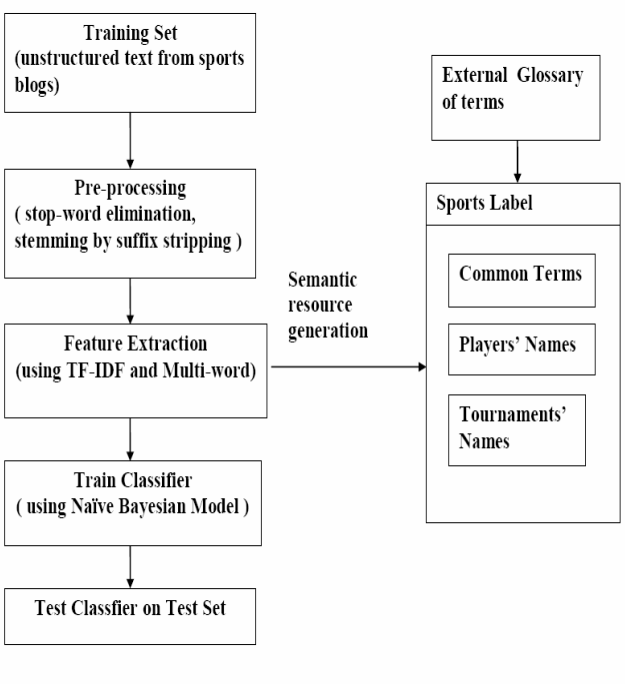

Figure 1. Classification of textual entries from sports blogs

Multi-words add some semantic sense to classification. They are useful in classification as well as disambiguation. A multi-word is a sequence of consecutive words having a semantic meaning. For example, "hat trick", "batting order", " third umpire", "clean bowled", "one day international" etc. are multi-words frequently occurring in cricket literature. Several methods can be used to extract multi-words computationally from text such as the frequency approach, sentence comparison approach [8], mutual information approach [15] etc. Since we were building classifier for unstructured text of sports blogs, the feature set was enhanced with the help of external glossaries of terms used in various sports.

The features extracted in this phase are used to train a naïve Bayesian classifier in the next phase. Instead of saving the extracted features (words and multi-words) as a BOW (bag of words), we further classify them using a semi-supervised approach. For this, the features are further classified into three groups: i) common terms, ii) Players' Names and 3) Tournaments' Names. Common terms are identified by matching them with an external glossary of terms. The heuristic used to identify Players' Names and Tournaments' names is that they are generally proper nouns or multi-words in which each separate word begins with a capital letter. Tournaments' names additionally differ from Players' names by the presence of substrings like 'Trophy', 'Challenge', 'Series', 'Tournament', 'Cup' etc. Thus, we can classify the multi-word 'M. S. Dhoni' as a Player's Name while 'World Cup 2011' is a Tournament name. Thus, each feature is classified into one of three groups. Semantic resources

generated in this way could be used for data mining requirements that may arise in the future.

C. Modeling Phase

In this phase we built the naïve Bayesian classifier [1] [2] [16]. The training data was obtained from online sports blog posts on three sports: 'cricket', 'field hockey' and 'tennis'.

For Bayesian classification, an initial knowledge of probabilities is required for good estimation. So, this method can only be used when the training data set is sufficiently large and there is sufficient training data representative of each class. Moreover, the naïve Bayesian classification method assumes conditional independence among attribute values of a training instance.

Let F indicate the feature-set obtained previously, from the Feature Extraction phase. Let C indicate the pre-defined set of distinct classes to which an unstructured text may belong. Let B indicate the set of unstructured blog texts used in the training set, while B_c is a subset of B which indicates the set of training texts belonging to some class c. The naïve Bayesian classifier calculates the prior probability of each class $c \in C$, as shown in Equation (1).

$$P(c) = \frac{|B_c|}{|B|} \tag{1}$$

We stored each instance of unstructured text (blog entry) using the Multivariate Bernoulli Model [1] [2]. Thus, each text was represented as a binary feature vector (indicating absence or presence of each feature term) of size $|F|$. In order to classify a new instance of unstructured text, its probability of belonging to each class is calculated as shown in Equation (2).

for each class $c \in C$ *do*

$$nbp(c) = P(c)\prod_{i=1}^{|F|} P(f_i / c) \tag{2}$$

The instance is finally assigned to the class with the highest probability, as shown in Equation (3).

$$finalprob = \arg\max_{c \in C}(nbp(c)) \tag{3}$$

D. Evaluation Phase

In this phase the classification accuracy of the naïve Bayesian classifier at the task of classifying unstructured text posts collected from various sports blogs is computed on the basis of experimentation. Let *b1, b2, b3, b4* represent the following values :

b1 = number of blog entries of sport *s* correctly labeled as sport *s*

b2 = number of blog entries of other sports incorrectly labeled as sport *s*.

b3 = number of blog entries of sport *s* incorrectly labeled as some other sport

Now, the classification accuracy can be computed in terms of the well-known and frequently used metrics: precision, recall and F-measure as given by the following equations.

$$precision(P) = \frac{b1}{b1 + b2} \tag{4}$$

$$recall(R) = \frac{b1}{b1 + b3} \tag{5}$$

$$F - measure = \frac{2PR}{P + R} \tag{6}$$

III. EXPERIMENTS AND RESULTS

The corpus consisted of 600 texts (about 200 texts from each of the three categories of sports – cricket, field hockey and tennis) collected from several popular sports blogs on the Internet. In order to get a good estimate of accuracy, 3-fold cross-validation was used and their results were averaged. Classification accuracy of the naïve Bayesian classifier was measured in terms of average precision, average recall and average F-measure over the test set.

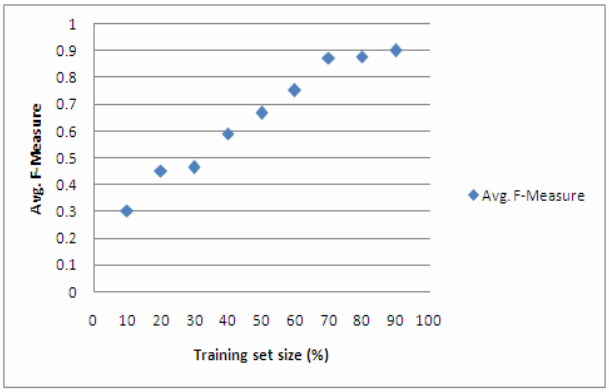

Figure 2. Plot of Training set size v/s Classification accuracy

Several experiments were conducted with varying sizes of training and test datasets. Fig. 2 shows the plot of average classification accuracy in terms of F-measure v/s training set size. It can be observed from Fig. 2 that by using the proposed method an average classification accuracy of over 87% can be achieved with approximately 70% documents in the training set and 30% in the testing set. However, more extensive testing with different types and size of training datasets is still required. Also, more empirical studies are required to observe the effect of varying the number of features extracted upon text classification accuracy.

IV. CONCLUSION

The combination of TF-IDF and multi-word for feature extraction followed by naïve Bayesian classification is an effective method for classifying unstructured texts such as

sports blog posts. However, we expect classification accuracy to improve with more extensive training using larger feature sets.

In future, we propose to use an extensive training set along with the semantic resources generated as explained in this paper, to mine information for a variety of commercially useful applications. Such applications include content management of sports blogs through automatic classification, frequency analysis of classified sports blog entries to determine the popularity of a particular sport and sportspersons and subsequently use this knowledge for determining brand endorsers, display of commercial products on classified blogs for targeted advertisement etc.

REFERENCES

[1] S. Kim, K. Han, H. Rim, and S. H. Myaeng, "Some effective techniques for naïve bayes text classification," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466, Nov 2006.

[2] M. J. Meena, and K. R. Chandran, "Naïve bayes text classification with positive features selected by statistical method," in proc. of the IEEE international conference on Advanced Computing, pp. 28 - 33, Dec. 2009.

[3] Z. Wang, Y. He, and M. Jiang, "A comparison among three neural networks for text classification," in proc. of the IEEE 8th international conference on Signal Processing, 2006.

[4] Z. Wang, X. Sun, D. Zhang, X. Li, "An optimal SVM-based text classification algorithm," in proc. of the 5th IEEE international conference on Machine Learning and Cybernetics, pp. 1378 – 1381, Aug 2006.

[5] M. Zhang, and D. Zhang, "Trained SVMs based rules extraction method for text classification," in proc. of the IEEE international symposium on IT in Medicine and Education, pp. 16 – 19, 2008.

[6] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, "Text document pre-processing with the Bayes formula for classification using support vector machine," IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. .9, pp. 1264 – 1272, Sept. 2008.

[7] R. D. Goyal, "Knowledge based neural network for text classification," in proc. of the IEEE international conference on Granular Computing, pp. 542 - 547, Nov. 2007.

[8] W. Zhang, T. Yoshida, and X. Tang, "Text classification using multi-word features," in proc. of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524, 2007.

[9] L. Hao, and Lizhu Hao, " Automatic identification of stopwords in Chinese text classification," in proc. of the IEEE international conference on Computer Science and Software Engineering, pp. 718 - 722, 2008.

[10] M. F. Porter, "An algorithm for suffix stripping," Program, 14 (3), pp. 130-137, 1980.

[11] S. Changuel, N. Labroche, and B. Bouchon-Meunier, "Automatic web page author extraction," LNAI 5822, pp. 300-311, Springer-Verlag, 2009.

[12] W. Zhang, T. Yoshida, and X. Tang, "TF-IDF, LSI and Multi-word in information retrieval and text categorization," in proc. of the IEEE international conference on Systems, Man and Cybernetics, pp. 108 - 113, 2008.

[13] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, Vol. 28, No. 1, pp. 11-21, 1972.

[14] K. S. Jones, "IDF term weighting and IR research lessons", Journal of Documentation, Vol. 60, No. 5, pp. 521-523, 2004.

[15] K. W. Church, and P. Hanks, "Word association norms, mutual information and lexicography," Computational Linguistics, Vol. 16, No. 1, pp. 22-29, 1990.

[16] T. M. Mitchell, "Machine Learning", McGraw-Hill International Editions, 1997.