Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Using Machine Learning for Automatic Text Classification of Unstructured Blog Data

Dr. Nitin Rajvanshi

Lecturer (Sel. Grade), Govt. Women Polytechnic College, Jodhpur

Abstract- *Opportunities for integrating applications of machine intelligence into the daily lives of people are growing with the increasing popularity of computing systems, the widening diversity of web services, the growing popularity of portable devices that contain general-purpose operating systems, and ongoing inventions in human-computer interaction— including the cases of speech recognition, handwriting, and sketch-understanding interfaces. Much of machine learning (ML) research is inspired by problems and its solutions from biology, medicine, finance, astronomy, etc.This paper presents automatic classification of unstructured blog entries by following pre-processing steps like tokenization, stop-word elimination and stemming. It uses Machine Learning techniques for feature set extraction, and feature set enhancement by semantic resources followed by modeling using a alternative machine learning model—the naïve Bayesian model.Empirical evaluations and calculations done in this paper indicate that this multi-step classification approach has resulted in good overall classification accuracy over unstructured blog text datasets with machine learning model alternative.Automatic classification of blog entries is generally treated as a semi-supervised machine learning task, in which the blog entries are automatically assigned to one of a set of pre-defined classes based on the features extracted from their textual content. The naïve Bayesian classification model clearly out-performs the other classification model when a smaller feature-set is available which is usually the case when a blog topic is recent and the number of training datasets available is restricted.*

Keywords- *Automatic Blog Text Classification; Feature Extraction; Machine Learning Models; Semi-Supervised Learning; Polysemy ; Prior Probability.*

1.Introduction

Blogging is a popular way of communicating, information sharing and opining on the Internet. There are blogs devoted to sports, politics, technology, education, movies, finance etc. Popular blogs have millions of visitors annually, so they are also important platforms for mining consumer preferences and targeted advertisement. Most of the content posted on blogs is textual and un-structured. Classifying blog text is a challenging task because blog posts and readers' comments on them are usually short, frequently contain grammatical errors and make use of domain-specific abbreviations and slang terms which do not match dictionary words. They are also punctuated inappropriately making tokenization and parsing using automated tools more difficult.

The blog posts of Internet users are organized in one of three ways [1]—1) Pre-classified; 2) Semi-classified or Partially Classified; or 3) Unclassified. These three categories are briefly explained next.

1) Pre-classified—Pre-classified blogs have separate web-pages allocated to each sub-class, so that the content posted is automatically sorted. For example, a blog that posts on computer technology could have previously allocated pages for categories like "hardware", "software", etc.

2) Semi-classified—Semi-classified blogs are those which have some web-pages pre-classified exclusively for popular categories. For example, a sports blog might contain separate web-pages for popular sports, while posts on less popular sports often simply referred to as the category "Others".

3)Un-classified— Un-classified blogs contain no fine classification and allow all blog postings to appear in antemporary manner. For example, an movie blog could contain posts on movies, music, actors, viewers' opinions and replies to other bloggers etc. all on one page.
Most blogs fall into the semi-classified or un-classified category.Machine learning techniques like naïve Bayesian [1-3], Artificial Neural Networks [4], Support Vector Machines [5] as well techniques that combine various machine learning methods [6] have been used by researchers for automatic text classification. In addition to content management, classification and summarization of blog text data has several important applications.

2. **Related Work**

The major steps are : pre-processing raw text for dimensionality reduction, extracting the word feature set useful for classification and handling difficulties arose because of synonyms and polysemes in text classification.

2.1 Pre-Processing

Unstructured text needs to be pre-processed before features can be extracted from it. Here, features are the words of text. As natural languages have a vast vocabulary , this pre-processing reduces the amount of term matching involved. This involves Stop-word removal [1-3] and stemming.

Stop-words are the words which occur frequently in a language and does not indicate a particular class of documents, so not useful for classification.Words like “the”, “is”, “in”, “or”, “it”, “for” etc. are stop-words in English. Removing stop- words reduces the size of the text to be processed .

Stemming converts the word to its root or base form and thus reduces and results in reduction of word features to be processed.

2.2 **IdentifyingFeatures for Text Classification**

The most common feature extraction techniques include TF-IDF . By this the most significant words or words with high discrimination are extracted and identified as features.

TF-IDF is an acronym for term frequency-inverse document frequency. It is based on the heuristic that a term is a good discriminator if it occurs frequently in a document but does not occur in many distinct documents of the corpus.

2.3 “Synonymy” and “Polysemy”

“Synonymy” and “Polysemy” are most frequently occurring problems in text classification. “Synonymy” is the capacity for different words to have the same meaning. For example, “wealthy” and “rich” are synonyms. A relevant document could be left during key basedinformation retrieval or wrongly classifiedif it uses a synonym of the feature term instead of the exact same feature term. “Polysemy” is the capacity for the same word to have different senses. For example, the word “output” could mean a “finished product of any factory” or an “result from computer”. A polyseme of a feature term could cause anless or irrelevant document to be retrieved during search. During the task of performing text classification using word features, the problem of “synonymy” can be solved by using a thesaurus or an online lexical database like WordNet. “Polysemy” is more cumbersome to handle, however techniques for word sense disambiguation have been successful in handling this issue to some extent.

3. Automatic Classification of Blog Entries

Using Machine Learning Techniques

The dataset used for experimentation consisted of a variety of blog entries of type Sports, Computer Technology and Environment as shown in table 1. The automatic classification of blog text has four major phases—1) Pre-Processing Phase; 2) Feature Extraction and Enrichment Phase; 3) Classifier Modeling and Training Phase; and 4) Evaluation Phase

3.1. Pre-Processing

The pre-processing steps performed on the blog text weretokenization, stop-word elimination, stemming and spell error correction.

3.2 Feature Extraction

After stop-word elimination and stemming performed, the vocabulary of the blog posts is still very large. All these words are not useful in classification. In order to extract most significant and relevant word features, tf-idf“term frequency-inverse document frequency” is used. In this case a document means a single blog post

Table 1. Dataset description.

Sr. No.	Category	Sub-Categories
1	Sports	Cricket Tennis
2	Computer Technology	Hardware Software

Table 2. Partial list of acronyms for "cricket".

Acronym	Expanded Form
BCCI	Board of Control for Cricket in India
ICC	International Cricket Council
LBW	Leg before Wicket
ODI	One Day International
T20	Twenty 20

Features are based on their tf-idf values (Term frequency and Inverse document frequency) and extracted the top ranking 40% words to be used as feature extractor for unstructured/unorganized blog text classification .

Feature Vector Set for sub categoryCricket: [Ball, Bat, Batsman, Bowler, ODI, Spinner ,Batting Average, Bold out, Caught behind]

3.3. Classifier Selection and Training

This model simply showswhether each feature term is present or absent. For ex-ample, if there are n features extracted in the feature extraction process, then, a blog post entry "e" would be internally represented as an ordered sequence, e = (i1, i2, i3, ⋯, in) where each "ik" is a binary variable indicating "1" for presence of feature term and "0" for absence. The well-known machine learning model Naïve Bayesian Model is used hereto classify unstructured blog text data.

3.3.1. Naïve Bayesian Model

The naïve Bayesian model is a probabilistic approach. It is based on the assumption of conditional independence among attributes. Given a training set containing attribute values and corresponding target values (classes), the naïve Bayesian classifier predicts the class of an unseen (new) instance, based on previously observed probabilities of the feature terms occurring in that instance.

Let C indicate the set of pre-defined classes to which the blog post may belong. Let B indicate the training set of pre-processed blog posts, while Bc is a pre-labeled subset of B that contains blog posts of some class c Є C. Let F be the final feature set generated during the Feature Extraction and Enrichment Phase. The probabilities of occurrence each of the features in the feature set F for each class, was computed by making one pass over the blog training set. First, the naïve Bayesian classifier [1] computes the prior probabilities of each class c Є C as indicated by Equation (1).

for each class c Є C do

$$P(C) = \frac{|B_C|}{|B|} \qquad (1)$$

So applying this model :
Here, Class **C**= [Sports, Computer]

Subclass of class sports = Cricket
B is Training set of predefined blog, so in this case Training set for **cricket** :
[Ball, Bat , Batsman, Bowler, Spinner, Batting Average, Caught Behind, Bold out, BCCI, ICC, IBC,ODI, T20]

Similarly Training set for Sub class **Tennis**:
[Deuce, service, Ace, Match Point, Game Point]

B_c :Some post Є C
So, foreach class c Є C (like sports, Computer), prior probabilities are calculated by equation 1.
P(ball/Cricket) = 3/7
P(ball/Tennis) = 2/9
P(Deuce/Cricket) = 3/9
P(Deuce/Tennis) = 3/7

Likewise all the prior probabilities are calculated.
Now, in order to classify a new blog entry e, theprobability of it belonging to each class is predicted as shown in Equation (2)

In Equation (2) the P(fi/c) terms indicate the statistical probability of occurrence of the ith feature term in a blog entry of category c.

$$Pe(C) = P(C)\prod_{i=1}^{|F|} P(^{fi}/_{fc}) \qquad (2)$$

The blog post is then assigned to the class with the highest probability as given by the following equation 3:

$$maxpr = \arg max\left(pe(c)\right) \qquad (3)$$

c Є C

3.4**. Evaluation Phase**
In this case pre-processing and feature set are generated over 2 categories of blogs with 2 sub-categories under each type. These categories have been depicted in Table 1. Classification Accuracy of the model has been evaluated in terms of precision, recall and f-measure over the data sets of blog. Experiments with varying sizes of feature-set was repeated.

4. **Results**

The results shown in Table 3 clearly depicts that the experiment done for automatic categorization of unorganized/unstructured blog text using Naïve Bayesian Model gives relevant retrieval as shown by the parameters calculated.

Table 3

FS Size	**Category**	**Avg. Precision**	**Avg. Recall**	**Avg. F-measure**
10%	Sports	0.7473	0.7134	0.7304
10%	Computer	0.7444	0.7396	0.7419
20%	Sports	0.7630	0.7598	0.7613
20%	Computer	0.7626	0.7630	0.7628

5. **Conclusion**

In this paper an attempt is made for automaticclassification of unstructured/unorganized blog posts using a semi-supervised machine learning approach. Evaluation indicate that the multi-step classification strategy can classify blog text with good accuracy. Tf-Idf is an effective statistical feature-set extractor for blog entries. Moreover, results indicate that the naïve Bayesian classification model clearly gives better results especially when a restricted featureset is available.
More experiments can be performed with larger feature set and other models of machine learning like Artificial Neural Networks can be applied to it, and a comparative performance analysis can be done.

REFERENCES

[1] M. K. Dalal and M. A. Zaveri, "Automatic Text Classification of Sports Blog Data," *Proceedings of the IEEE International Conference on Computing, Communications and Applications* (*ComComAp*2012), Hong Kong, 11-13 January 2012, pp. 219-222.

[2] 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, 2006, pp. 1457-1466. doi:10.1109/TKDE.2006.180

[3] M. J. Meena and K. R. Chandran, "Naïve Bayes Text Classification with Positive Features Selected by Statisti- cal Method," *Proceedings of the IEEE International Con- ference on Advanced Computing*, Chennai, 13-15 De- cember 2009, pp. 28-33. doi:10.1109/ICADVC.2009.5378273

[4] Z. Wang, Y. He and M. Jiang, "A Comparison among Three Neural Networks for Text Classification," *Pro- ceedings of the IEEE* 8*th International Conference on Signal Processing*, Beijing, 16-20 November 2006, pp. 1883-1886. doi:10.1109/ICOSP.2006.345923

[5]Z. Wang, X. Sun, D. Zhang and X. Li, "An Optimal SVM-Based Text Classification Algorithm," *Proceedings of the IEEE* 5*th International Conference on Machine Learning and Cybernetics*, Dalian, 13-16 August 2006

[6] M. Zhang and D. Zhang, "Trained SVMs Based Rules Extraction Method for Text Classification," *Proceedings of the IEEE International Symposium on IT in Medicine and Education*, Xiamen, 12-14 December 2008, pp. 16-19. doi:10.1109/ITME.2008.4743814

[7] R. D. Goyal, "Knowledge Based Neural Network for Text Classification," *Proceedings of the IEEE International Conference on Granular Computing*, Fremont, 2-4 De- cember 2007, pp. 542-547. doi:10.1109/GrC.2007.108

