International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-7,October 2015
ISSN: 2395-3470
www.ijseas.com

Performance evaluation of classifiers for sentiment labeled text Dataset

Prof. (Dr.) N. S. Chandolikar
Vishwakarma Institute of Technology, Pune, India

Abstract

Availability of large number of electronic text documents increases importance of text mining. Text Classification is the method of classifying text documents into different categories based on their content. Sentiment labeled dataset can be classified using classifiers; models trained by this classifier can be used to label unseen text. Methods for text classification such as Naïve Bayes, Support Vector Machine, K-Nearest Neighborhood, and Decision tree are used to train classifier using text dataset. This paper is indented to deal with the sentiment labeled text classification process; classifiers are evaluated on various performance measurement terms.

Keywords: *Text mining, sentiment labeled dataset, text classification*

1. Introduction

Text mining is a promising new field that attempts to gather significant and meaningful information from natural language text. Text mining methods analyzes text to extract information that is useful for particular purposes. Data mining is looking for patterns in data similarly text mining is looking for patterns in text. Unlike data mining in text mining records (or docs) are not structurally identical and are not statistically independent.

Huge amount of text is available on websites, email, postings on social media and electronic documents .Text mining can help an organization to derive potentially valuable and new information hidden in text-based content.

1.1 Text categorization

Text classification [1] is a process of automatically categorizing a set of documents into categories from a predefined set. Automated text classification is attractive because it replaces tedious and time consuming process of manually organizing text document. Usually manual process is too expensive, or simply not feasible.
Let D is set of documents and di is every individual document. These documents are categorized into C_1,C_2,.....,C_n set of categories, then text classification assigns one category C_j to a document d_i.

2. Dataset used

Experiments are performed on real world datasets available on UCI machine learning repository (http://archive.ics.uci.edu/ml)[2]. It has sentiment Labelled Sentences .The dataset contains sentences labelled with positive or negative sentiment. Dataset used for experiment is Yelp dataset, this dataset has restaurant reviews. Score is either 1 (for positive) or 0 (for negative). It has 500 positive and 500 negative sentences. The sentences come from website yelp.com.

3. Methodology

The first step is collecting the Data set and converting it to suitable file format. Next the pre-processing is used to present text collection noise free. Further tokenization, Stopwords removal, Stemming is applied .Feature Selection

is applied to select suitable words which represents sentiments. Then Classification Algorithms like Naïve Bayes, SVM, Lazy learners and Decision Tree are used to categorize text. Performance of classifiers is evaluated based on True positive Rate, Precision and Recall, F-measure)

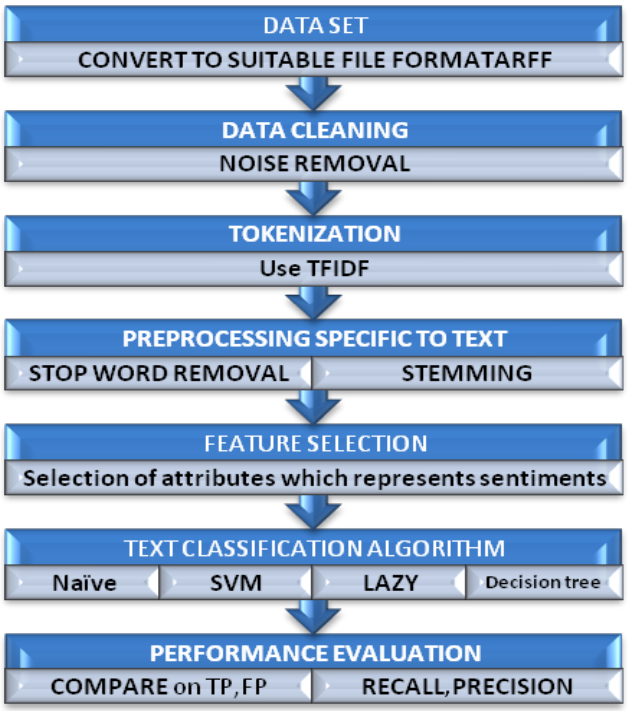

Fig1: Proposed methodology

3.1 Data set preparation

Dataset used for text classification must be converted to suitable file format.

3.2 Data preprocessing

Data cleaning is one of important task of data preprocessing, noise is removed from the data set.

3.3 Tokenization

Tokenization is the act of breaking up a text document into parts such as words, keywords, phrases, symbols and other elements called tokens.

TF–IDF(term frequency–inverse document frequency) is a term which finds importance of a word to collection of document or corpus. It is one of mostly used weighting factor in text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Term frequency : In the English text documents , with respect to query "the dell laptop" we wish to determine which document is most relevant to . For this we can eliminate documents that do not contain all three words "the", "dell", and "laptop", but this still leaves many documents. To further distinguish them, we count the number of times each term occurs in each document and sum them all together; the number of times a term occurs in a document is called its term frequency.

3.4.1 Word stemming

Word Stemming is one of important process which is applied before text mining. it is also known as lemmatization. it is a technique for the reduction of words into their stems or base word. In stemming the ‘stem’ is obtaining after applying a set of rules but without bothering about the part of speech or the context of the word occurrence. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like.

3.4.2 Stop word removal

The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. There is no standard list of stop words but determiners (like the, a, an, another),Coordinating conjunctions (like for, an, nor, but, or, yet, so) and Prepositions (like in, under, towards, before etc) are generally used as stop words.

3.5 Feature selection

With the presence of a large number of features, a Classification learning model is likely to give wrong classification or reduced performance. Feature selection is a extensively used technique for reducing number of features (attributes) among practitioners. It intend to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to better learning performance(accuracy) and less time. This gives higher model interpretability.

3.6 Text classification

Text classification methods categories labeled dataset. Labeled dataset is used to train classifiers. The classifier which gives best performance is suitable for prediction later.

3.6.1 Bayesian (Generative) Classifiers:

Bayesian classifiers attempts to build a probabilistic classifier based on modeling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

Naive bayes classifier

The Naive Bayes classifier [3] one of most popular classifier, it is usable for text classification also. It assigns a class to an unlabeled object based on maximum *likelihood* principle. The instance is assigned the class which maximizes the "*a posteriori*" probability, which is a function of the *class prior probability* and of the instance likelihood with respect to the class.

$$Cp = Amax\ P(c|d) \quad (1)$$

Where p is maximum posteriori as shown in equation 2 and 3.

$$= Amax\ \frac{P(d|c)P(c)}{P(d)} \quad (2)$$

$$= Amax\ P(d|c)P(c) \quad (3)$$

Naïve classifier assumes attribute independence hypothesis therefore it is called *naïve*. the computation of the conditional probability *P(d | c)* becomes just a calculation of a product between the probabilities of each attribute.

3.6.1 SVM

Sequential Minimal Optimization is a simple algorithm based on the concept of Support Vector Machine. Support Vector Machines (SVM)[4][5] can be used to learn with large amounts of data. SMO can solve the SVM QP problem without any extra matrix storage and without using al QP optimization steps at all.SMO algorithm solves the smallest possible optimization problem at every stage. It uses two Lagrange multipliers, because the Lagrange multipliers must obey a linear equality constraint. At every stage, SMO optimizes two Lagrange multipliers to jointly, it get the optimal values for these two multipliers, and updates the Support Vector Machine to reflect the new optimal values. Solving or two Lagrange multipliers can be done analytically, therefore SMO is advantageous.

3.6.2 K-nearest neighbors classifier

The k-nearest neighbors classifier[5] [6] assigns to an instance the class. class is assigned to instance based on the k nearest instances. k-nearest neighbors classifier uses a distance function which calculate distance between the instances and neighbors in dataset. The k parameter is chosen to be an odd number, so that a majority always exists. When k = 1, the classifier simply becomes a *nearest neighbor* classifier.

IBK is a k-nearest-neighbour classifier that uses the same distance metric. The distance function like Euclidean distance, Manhattan, and Minkowski distances are used to find out nearest

neighbour The number of nearest neighbours can be provided explicitly otherwise it automatically uses it. IBK is a nearest- neighbour classifier. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbours. A linear search is the default but further options include KD-trees, ball trees, and so-called "cover trees". IBK is a type of lazy learner.

3.6.4 Decision Trees

Decision trees[6] are designed with the use of a hierarchical division of the underlying text with the use of different text features. For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

3.7 Classifier Performance evaluation

Classifiers performance is evaluated based on precision and recall [8] [9] .Precision measures the exactness of a classifier whereas Recall measures the completeness, or sensitivity, of a classifier. Equation 4 shows formula to compute precision whereas Equation 5 shows formula to compute recall. Less false positives means higher precision, while more false positives means lower precision. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases. These two metrics can provide much greater insight into the performance characteristics of a binary classifier.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (4)$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (5)$$

Precision and recall can be combined to produce a single performance evaluation term known as F-measure, (equation 5) which is the weighted harmonic mean of precision and recall.

$$F_{measure} = \frac{2 \times presion \times recall}{precesion + recall} \quad (6)$$

4. Experiments

To assess the effectiveness of the algorithms for text classification, the series of experiments were performed in Weka. The java heap size was set to 1024 MB for weka-3-6. Weka [10] [11] is a collection of machine learning algorithms for data mining tasks. Weka provides facility for data pre-processing, classification and amny other data mining techniques. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface, Java interface.

4.1 Data pre processing

Data preprocessing [7] is very key process to make dataset suitable for further processing.

4.1.2 Converting to suitable file format

Then available data set is converted to .arff (Atribute relation file format) file format. ARFF file format is suitable to perform experiment using WEKA. Following is the created .arff file for experimentation.

@relation 'yelp'
@attribute review string
@attribute class {0,1}
@data
'Wow... Loved this place',1

4.1.1 Data Cleaning

Data set is pre processed before classification, some symbols are changed to other symbol to make it suitable for text mining.

4.1.3 Applying suitable filter

Supervised attribute selection string to word vector is used . "String to word vector" filter is applied to convert data. The

StringToWordVector filter can perform TF/IDF transformation.

Inverse document frequency: However, because the term "the" is so common, this will tend to incorrectly highlight documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "dell" and "laptop". The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "dell" and "laptop". Hence an inverse document frequency factor is incorporated which reduces the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

4.1.4 Selection of attributes

Appropriate attributes are selected to represent sentiments in text [12]. Sample of selected attributes are shown below.

Amazing, Best, Delicious, Fresh, Great , Love, Like, Love, Luckily, Never, Not, Needless, Never, No, Not, Unfortunately, Amazing, Avoid ,Awesome, Awful, Bad, Best, Dirty, Disappointed , Enjoy , Excellent , Excited Etc.

4.1.5 Classification

Various classification techniques are used to investigate text classification of Yelp dataset. The k-NN classifier is available in the *weka.classifiers.lazy.IBk* component. The best result achieved with this kind of classifiers has shown a correctness percentage of **75.1%**. In WEKA, the SVM classifier is implemented in the weka.classifiers.functions.SMO component. The result achieved with this kind of classifiers has shown a correctness percentage of 63.3**%**. The naïve bayes classifier is implemented in the weka.classifiers.bayes.naivebayes component. The result achieved with this kind of classifiers has shown a correctness percentage of 59.8%. The decision tree classifier is implemented in many methods, the weka. classifiers.trees.randomforest component is used for tree based classification. The result achieved with this kind of classifiers has shown a correctness percentage of 74.2**%**.

5. Result and discussion

Proposed model is implemented to classify sentiment labeled dataset “yelp”. To get result at first four different category of classification algorithms are compared. IBK from KNN classifier, SMO from SVM, naïve bayes from bayes based classifier and random forest from decision tree classifier are taken for performance comparison. Algorithm performance is compared based on correctly classified instance, time taken, True positive rate, precision and recall as shown in table 1 and 2 .

Table 1: Performance comparison of classifiers on precision and recall

	Naïve	K nearest neighbor IBK	Decision tree (random forest)	SMO
Precision	0.61	0.79	0.78	0.65
Recall	0.59	0.75	0.74	0.63
F-Score	0.58	0.74	0.73	0.62

Table 2: Performance comparison of classifiers

	Naïve	K nearest neighbor IBK	Decision tree (random forest)	SMO
Time Taken	0.14	0.02	0.05	0.03
Correctly Classified Instance	59.8%	75.15%	74.2%	63.3%

This study was performed to

- Investigate and pre-process the features in the text dataset.
- Define the class attributes which divide the set of instances into the appropriate classes.
- Examine various classification methods for applicability for text categorization.
- Make decision on a testing method to estimate the performance of the algorithm.
- Suggest performance improvement in text categorization.

6. Conclusion

Text mining can improve analysis of text by adding a new level of surveillance to text documents or online text review. The classifiers Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Decision Tree for Text Classification is compared with each other on their performance on text dataset yelp. From empirical results of the experiments, it is concluded that KNN Instance based classifier is best and stable classifier for organizations concerned with correct classification of restaurant review available on YELP. The experimental results on dataset evident that proposed methodology achieved high performance for sentiment labeled text. However, we would like to repeat our experiments with larger and more varied datasets.

7. References

[1] M. Ikonomakis ,S. Kotsiantis, V. Tampakas "Text Classification Using Machine Learning Techniques", WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005.

[2] Dimitrios Kotzias, Misha Denil, Nando De Freitas, Padhraic Smyth, "From Group to Individual Labels using Deep Features" KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

[3] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification", IEEE Transactions On Knowledge and Data Engineering, Vol. 18, No. 11, November 2006 .

[4] Berna Altnela.n, Murat Can Ganizb, Banu Diri, "A corpus-based semantic kernel for text classification by using meaning values of terms", Engineering Applications of Artificial Intelligence, Elsevier Ltd 2015

[5] Jagdish Raikwal Vikas Ransore, " A survey on various text mining techniques and their Issues" international Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 5, May 2015.

[6] Sadegh Bafandeh Imandoust , Mohammad Bolandraftar, " Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background" , International Journal of Engineering Research and Applications ,Volume 3,Issue 5, Page .No 605-610, October 2013.

[7]Mita K. Dalal, Mukesh A. Zaveri ,"Automatic Classification of Unstructured Blog Text", Journal of Intelligent Learning Systems and Applications, 2013.

[8] Jiawei Han And Micheline Kamber, "Data mining concepts and techniques" , Morgan Kaufmann publishers .an imprint of Elsevier .ISBN 978-1-55860-901-3,2008.

[9] S. Kabinsingha, S. Chindasorn, C. Chantrapornchai, "A Movie Rating Approach and Application" , International Journal of Engineering and Innovative Technology (IJEIT)Volume 2, Issue 1, July 2012.

[10] Abdullah H. Wahbeh and Mohammed Al-Kabi, "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text", Abhath Al-Yarmouk: "Basic Sci. & Eng." Vol. 21, No. 1, 2012.

[11] Witten IH, Frank E, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann. Second edition, 2005

[12] Horakova , M, " Sentiment Analysis Tool using Machine Learning", Global Journal on Technology, 05, pp 195-204, 2015.

Dr. Neelam S. Chandolikar received the B.Sc. (Computer Sc.), MCA and PhD degrees in 1995, 2002 and 2015, respectively. She has teaching experience of more than 15 years. Her research interests are in data mining, text mining, information retrieval and natural language processing .In these areas, she has published many papers in international journals or conference proceedings.