APPLICATION OF FEATURE EXTRACTION TECHNIQUE TO UNSTRUCTURED TEXTS

Isabella J.[1], Suresh R.M.[2]

[1] Research Scholar, Sathyabama University, Chennai, India,
[2] Principal, Jerusalem Engineering College, Chennai, India
Email: [1]isabellajones71@gmail.com

Abstract

Inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The 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 control for the fact that some words are generally more common than others. Variations of the IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification. Text Classification is a semi-supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features. Text Classification has important applications in content management, contextual search, opinion mining, product review analysis, spam filtering and text sentiment mining. This paper explains the usage of Inverse document frequency for dealing with unstructured text, handling large number of attributes and the application of K-Nearest neighbour classifier to classify the documents.

Keywords: Inverse document frequency, IMDb, Mining, KNN Classifier and Naïve-Bayes Classifier

I. INTRODUCTION

Inverse document frequency (IDF) is a popular measure of word's importance. The IDF invariably appears in a host of heuristic measures used in information retrieval. However, so far the IDF has itself been a heuristic .It is a popular measure of a word's importance. It is defines as the logarithm of the ratio of number of documents containing the given word. This means rare words have high IDF and common function words like 'the' will have low IDF.IDF is believed to measure a word's ability to discriminate between documents. Text Classification involves assigning a text document to a set of pre-defined classes automatically, using a machine learning technique . The classification is usually done on the basis of significant words or key-features of the text document. Since the classes are pre-defined it is a supervised machine learning task. This paper explains the classification of unstructured data which includes steps such as pre 'processing (eliminating stop-words [1] [2] [3] , stemming [2] etc.), feature selection using various statistical or semantic approaches using appropriate machine learning for training a text classifier. Text classification has several useful applications opinion detection [4] and opinion mining from online reviews of products [5], movies or political situations and sentiment mining .

Blogging has become a popular means of communication over the Internet. On applying computational methods for the detection and measurement of opinion, sentiment and subjectivity in text. Sentiment classification is used in a number of areas such as recommender systems, online product review to retrieve information based on sentiment orientation. Evaluating sentiment orientation for the purposes of classification has received considerable research attention, and several approaches are surveyed in literature [6,7,8,9,10]. Thus the need to develop sophisticated text classification method arises.

II. LITERATURE SURVEY

Bo Pang et al., 2004 [6] investigated the effectiveness of classification of documents by overall sentiment using machine learning techniques. Experiments showed that the machine learning techniques give better result than human produced baseline for sentiment analysis on movie review data. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews.

Peter Turney 2002 [7] proposes an unsupervised learning algorithm, using semantic orientation of the phrases containing adjectives and adverbs, to classify

reviews. The approach initially extracts phrases containing adjectives and adverbs; the semantic orientation of the phrase is estimated using PMI-IR; based on the average semantic orientation the phrases the review is classified as recommended (Thumbs up) or not recommended (Thumbs down). Experiment was conducted using 410 reviews on various topics; an average accuracy of 74% was achieved.

Xiaowen Ding et al., 2008 [8] proposes a holistic lexicon-based approach which uses external indications and linguistic conventions of natural language expressions to determine the semantic orientations of opinions. Advantage of this approach is that opinion words which are context dependent are easily handled. The algorithm used uses linguistic patterns to deal with special words, phrases.

Wiebe J et al., 2004 [9] proposed a learning method for creation of subjective classifiers, which can be used on unannotated text. The method developed is superior to other previously used supervised learning approaches. In an attempt to build classifiers which can distinguish subjective and objective sentences, a new objective classifier was created using new objective clues which achieved higher recall than previous works. Their approach began with seeding process which uses known subjective words to automatically create training data.

Pang et al., 2004 [10] proposed a machine-learning method to find subjective portions in a document. Extracting of the subjective portions can be done using techniques for finding the minimum cuts in graphs. This makes it easy to incorporate the cross sentence related constraints. Pang et al., studied the relationship between polarity classification and subjectivity discovery, showing that shorter extracts got from compressed reviews retain polarity information as that of the full review.

In this paper it is proposed to compute the inverse document frequency of the extracted features by considering the individual predictive ability words and applying the classifiers to calculate the classification accuracy for various number of text documents.

III. METHODOLOGY

Data pre-processing reduces the size of the input text documents significantly. It involves stop-word elimination[1]and Stemming. Stop words are functional words which occur frequently in the language of the text.(For example,'a','the','an','of' etc) so that they are not useful for classification. Stemming is the action of words to their root or base form. Feature extraction/Selection helps identify important words in a text document. This is done by Inverse Document Frequency. The output after feature selection is a document vector in which the document vector simply indicates the absence or presence of feature on which an appropriate machine learning algorithm is used to train the text classifier. The classification accuracy is then calculated,

In this paper it is proposed to use online movie reviews as data due to the availability of a large number of reviews online. Internet Movie Database (IMDb) is an online database of information related to movies, television shows, actors, production crew personnel, video games and fictional characters featured in visual entertainment media. Bo Pang and Lillian Lee [6] provide collections of movie-review documents collected from the IMDb archives, labelled with respect to their overall sentiment polarity (positive or negative) or subjective rating (e.g., two stars). Features are extracted by using list of stop words for commonly occurring words and stemming words with similar context. The terms document frequency is computed. In a set of documents x and a set of terms, a each document can be modelled as a vector v in the a dimensional space, R^a this is a vector space model. Let the term frequency be denoted by, freq (x, a) this expresses the number of occurrence of term a in the document x. The term-frequency matrix $TF(x, a)$ measures the association of a term a with respect to the given document x. is assigned zero if the document does not contain the term, and a number otherwise. The number could be set as $TF(x, a) = 1$ when term a occurs in the document x or uses the relative term frequency. The **relative term frequency** is the term frequency versus the total number of occurrences of all the terms in the document. The term frequency is generally normalized by:

$$TF(x, a) = \begin{cases} 0 & \text{freq } (x, a) = 0 \\ 1 + \log (1 + \log (\text{freq } (x, a))) & \text{otherwise} \end{cases}$$

Another measure used is the **inverse document frequency (IDF)**, it represents the scaling factor. If term a occurs frequently in many documents, then its

importance is scaled down due to its reduced discriminative power. The $IDF(a)$ is defined as follows:

$$IDF(a(= \log \frac{1+|x|}{x_a}$$

x_a is the set of documents containing term a.

Similar documents have similar relative term frequencies. The similarity can be measured among a set of documents or between a document and a query. Cosine measure is generally used to find similarity between documents; the cosine measure is got by:

$$sim(v_1, v_2) = \frac{v_1 \cdot vsub}{|v_1|\,|v_2|}$$

where v_1 and v_2 are two document vectors, v_1, v_2 defined as $\sum_{i=1}^{a} V_{1i}\, V_{2i}$ and $\left| v \backslash V_1 \right| = \sqrt{V_1 \cdot v_1}$.

After Calculating the Inverse Document Frequency , the following Machine learning models for Classification are applied and the results are analysed for varied number of text documents.

A. Naïve Bayes classifier

Naïve Bayes classifiers are statistical classifier based on the Bayes theorem. It uses probabilistic approach for predicting the class of given data, by matching given data to the class having highest posterior probability. Following are the algorithms used in Naïve Bayes:

$$P(C_i | V) = \frac{P(V | C_i)\, P(C_i)}{P(V)}$$

where $V = (V_i, \ldots\ldots, v_n)$ is the document represented in n-dimensional attribute vector and $C_i, \ldots\ldots C_m$ represents m class. But is it computationally expensive to compute $P(V | C_i)$. In order to reduce computation, the naïve assumption of class conditional independence is made. Thus,

$$P(V | C_i) = \prod_{k=1}^{n} P(X_k | C_i)$$

B. K-Nearest Neighbour Classifier

K- Nearest neighbour classifier is based on the premises that the vector space model will be similar for documents which are similar. The training documents are indexed and each is associated with its corresponding label. When a test document is submitted, it is treated like a query and retrieves from the training set documents that are most similar to the test document. The class label of the test document is assigned based on the distribution of its k nearest neighbors. The class label can further be refined by adding weights. Thus by tuning k, higher accuracy is obtained. Nearest neighbor method are simple to understand and easy to implement .

IV. EXPERIMENT

A total of sixty reviews with 30 positive and 30 negative are chosen in this work. Total of 25 words were selected and their IDF computed. Naive Bayes and K Nearest neighbour classifier are used to compute the classification accuracy of the proposed method. Classification accuracy is appreciable irrespective of the number of text documents..

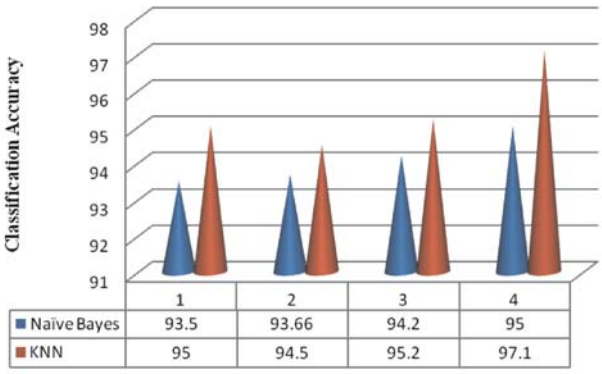

Fig. 1

Figure 1 shows the comparison of the classification accuracy using different classifiers to different number of text documents .

V. CONCLUSION

In this paper it is proposed to extract words from reviews and feature set is reduced by applying IDF. Irrespective of the number of Text documents used ,the classification accuracy is quiet promising Further work needs to be done on multiple dataset pertaining to different areas of text classifications.

REFERENCES

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

[2] Zhang W., Yoshida T., and Tang X. 2007. Text classification using multi-word features. In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 - 3524.

[3] Hao Lili., and Hao Lizhu. 2008. Automatic identification of stopwords in Chinese text classification. In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718 - 722.

[4] M. M. Saad Missen, and M. Boughanem. 2009. Using WordNet s semantic relations for opinion detection in blogs. ECIR 2009, LNCS 5478, pp. 729-733, Springer-Verlag Berlin Heidelberg.

[5] Balahur A., and Montoyo A.. 2008. A feature dependent method for opinion mining and classification .In proceedings proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7.

[6] Pang B, Lee L. (2004) "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", *Proceedings of the ACL, 2004*.

[7] Turney P. (2002) 'Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews', *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics - ACL*, 2002.

[8] Xiaowen Ding, Bing Liu,Philip S. Yu. (2008)A holistic lexicon-based approach to opinion mining. WSDM '08 Proceedings of the international conference on Web search and web data mining.

[9] Wiebe J, Bruce R, Martin M, Wilson T, Bell M. (2004) 'Learning Subjective Language', *Computational Linguistics*, Vol. 30, No. 3, pp. 277-308, January 2004.

[10] Pang B, Lee L. (2008) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1-2, pp. 1-135, 2008.

[11] Pang B, Lee L. (2002) 'Thumbs up? Sentiment Classification using Machine Learning Techniques', *Proceedings of EMNLP,* 2002.

[12] Salvetti F, Lewis S, Reichenbach C. (2004) 'Automatic Opinion Polarity Classification of Movie Reviews'. *Colorado Research in Linguistics, June 2004, Volume 17, Issue 1. Boulder: University of Colorado.*, 2004.

[13] Hall, M. A. & Smith, L. A. (1998). Practical feature subset selection for machine learning. In C. McDonald(Ed.), Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth, 4-6 February, 1998(pp. 181-191). Berlin: Springer.

[14] Kulkarni S, Lugosi G, Venkatesh S. (1998) 'Learning Pattern Classification - A Survey', *IEEE Transactions on Information Theory*, Vol. 44, No. 6, October 1998.

[15] Cody W, Kreulen J. T, Krishna V, Spangler S W, (2002) 'The Integration of Business Intelligence and Knowledge Management', *IBM Systems Journal,* Vol. 41, No. 4, pp. 697-713, 2002.

[16] Horrigan J. (2008) 'Online Shopping', *Pew Internet and American Life Project - Research Report*, February, 2008.

J.Isabella is a Research Scholar from Sathyabama university doing research in the field of Webmining. She received **"Best Research Student Paper award"** in the International Conference of Software Engineering and Applications, December 2011 held at Singapore.