An Approach to Multiclass Text Classification

Neha Sharma[1], R. K. Gupta[2]
[1, 2] Department of CSE & IT
[1, 2] Madhav Institute of Technology & Science, Gwalior (India)

Abstract- *Text categorization is the task of arranging a set of documents into preordained categories. This has several applications including spam filtering, email filtering, authorship attribution, content classification and news monitoring etc. Classifying text automatically has become the need in recent times due to the huge availability of text on the internet. It is attractive because it is efficient to extract information from categorized documents. There are two types of classification schemes: binary classification and multiclass classification. Binary classification deals with only two classes and used in applications such as sentiment analysis [positive, negative], spam filtering [spam, not spam], customer service message classification [urgent, not urgent] etc, whereas multiclass classification deals with more than two classes and used in applications such as identifying disease of a patient [tuberculosis, jaundice, diabetes], deciding whether temperature is [low, medium, high] etc. This paper is a review on multiclass classification methods as it has several benefits over binary classification such as it is used to solve real world problems which include soil classification, document classification, agricultural land area classification etc.*

Keywords- Text preprocessing, Feature Selection, Feature extraction, Text categorization.

I. INTRODUCTION

Traditionally documents were stored in paper files, within folders and filing cabinets and a major disadvantage of this type of document file organization was the time it takes to access the document file particularly when documents are in large quantity. Due to the advancement of technology documents are being stored digitally, but at the same time it is difficult to retrieve relevant document from the stored document if they are not categorized properly. Text categorization has recently gained the importance. It is the task of classifying text automatically [1]. Before performing the task of assigning categories to documents, it is necessary to carry out two main processes [3]. Firstly: documents must be transformed into a suitable format, for example: text present in all the documents should be converted in lowercase, white-spaces and non-informative words such as pronouns, articles and tags must be removed then important words are selected from the document. This reduction in vocabulary improves overall performance. Secondly: categories are assigned to documents. Categories may be derived from a large collection of very specific content identifiers; categories can be expressed as numerically, individual words or phrases [2]. Since documents can be classified using two classes or more, the system can be designed to perform binary classification in which classification is done only for two classes or it can be designed to perform multiclass classification in which classification is done for more than two classes so later technique is more beneficial than the previous one because it can be used to solve the real world problems such as document classification, soil classification, agricultural land area classification etc, So the purpose of this paper is to have a review on multiclass classification methods. The paper is organized as follows: Section 2 presents the literature review. Section 3 presents various text preprocessing methods. Section 4 presents various feature selection methods. Section 5 presents the feature extraction method. Section 6 presents various multiclass classification methods and finally, conclusion is given in section 7.

II. LITERATURE REVIEW

Aswini et al. [1] discussed the pattern mining techniques which are used to find the text patterns and also showed that how these patterns are useful to find interesting information. Sukanya et al Ghosh et al. [2] discussed stages of text mining process which includes information retrieval, natural language processing, data mining, information extraction. Agnihotri et al. [3] focuses on pattern and cluster mining on text data because text data is growing day by day in electronic format so there is a need to find clusters of similar kind of words and also finding the patterns. Cavnar et al. [4] discussed the text categorization using the N-gram technique. Chrystal et al. [5] has discussed the TF-IDF method as a feature extraction. Gomaa et al. [6] discussed different similarity measures which is useful in topic detection, clustering, machine translation, and text summarization etc. Mehra et al. [7] has conducted a survey on multiclass classification methods, and discussed three techniques to solve the problem of multiclass classification, techniques discussed are: First method are an extension of binary methods, second method converts multiclass classification problem into binary problem, third method discusses the hierarchical classification methods. Ramasundaram et al. [8] has used artificial neural network for the text categorization because there are some problems which cannot be solved sequentially. Jian-Fang et al. [9] has given various classification techniques such as naïve bayes, support vector machines and KNN.

III. TEXT PREPROCESSING

Before assigning categories to documents it is necessary to carry out the preprocessing of text documents because unprocessed documents consists unwanted information and removing such information will reduce the processing time required during categorization. Text preprocessing seems to be the most time consuming phase in the whole process of knowledge discovery [4]. Knowledge discovery from unprocessed documents produces inconsistent and noisy results [5]. There are four common preprocessing steps including tokenization, stop-word removal, and lowercase conversion and stemming.

1. Tokenization

In text preprocessing, the process to split text into words, phrases or other meaningful parts is known as tokenization. These splitted parts are named as tokens. For splitting text delimiters such as punctuation, white-spaces are observed because whenever these delimiters are present between texts it is broken into individual words.

Example: Input- Family, Country, Apple, Laptop, Jack Mac.

Output-

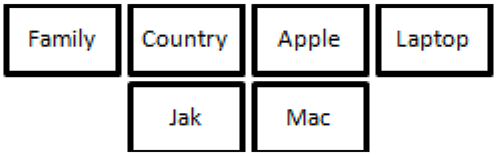

Fig 1: Tokenization

2. Stop-Word removal

Stop-words are the words that are commonly encountered in texts and are known to be non-informative words because they add noise in a document and found to be irrelevant in the text classification process and therefore needed to be removed prior to the classification.

Example of such words are 'the',' in',' a',' an' etc.

3. Lowercase Conversion

Converting the documents into lowercase is a preprocessing step. This is a necessary step because storing all the documents in a same format provides convenience in text classification process.

4. Word Stemming

Stemming is a technique that reduces the words into their root or stem [6]. The aim of stemming is to obtain stem, or root, forms of derived words. Since derived words are semantically similar to their root form so reduction of a word into its root form will be beneficial for the computation of the word occurrences.

Example: achievable, achievement, achieving belongs to root word 'achieve'.

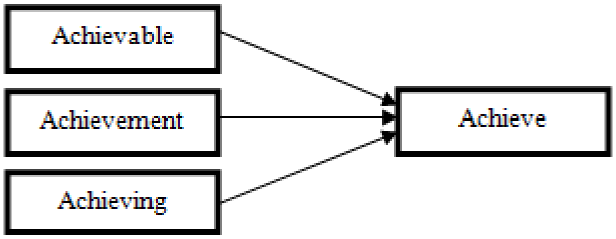

Fig 2: Word-Stemming

IV. FEATURE SELECTION

Feature selection is the phenomenon to select important features from the documents by reducing the dimensions of a document and removing irrelevant features which do not help in text classification process. There are various methods explained below.

1. Word Occurrences

To represent the document in vector form, single words found in the training corpus are selected as features ignoring the sequence in which words occur. If a word occurs in a document, it is assigned with value 1 or if a word is not present document is assigned with value 0. This type of representation is known as Boolean representation.

Example:
Text 1: good, fan, king, pot
Text 2: lion, jug, good, king

Table 1: Boolean representation

	good	Fan	king	pot	lion	Jug
Text 1	1	1	1	1	0	0
Text 2	0	0	1	0	1	1

2. Latent Semantic Indexing

Latent semantic analysis (LSA) is an indexing and retrieval method that uses mathematical technique called singular value decomposition (SVD). LSA records keywords which a document contains as well as it examines the document collection as a whole, to see which other documents contain some of the same words.

3. N-Gram

An n-gram is word sequences of length up to n [5]. These are basically a set of co-occurring words within a given window and when we compute n-grams you typically move one word forward. These are extensively used in natural language processing and text mining.

For example: The sentence is “sun rises in the east”.

If N=2 (known as bigrams), the n-grams would be:

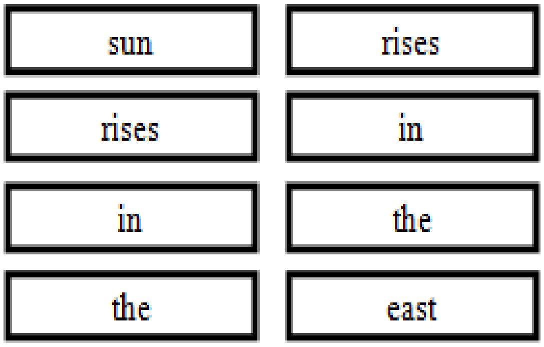

Fig 3: bi-gram representation

In above figure we have 4 bi-grams. If N=3 then n-grams would be:

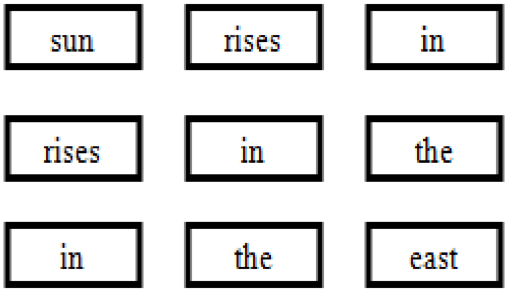

Fig 4: tri-gram representation

4. Parts of Speech

Assigning the part-of-speech (POS) to each word in a sentence is known as part of speech tagging. A part of speech is used in information retrieval.

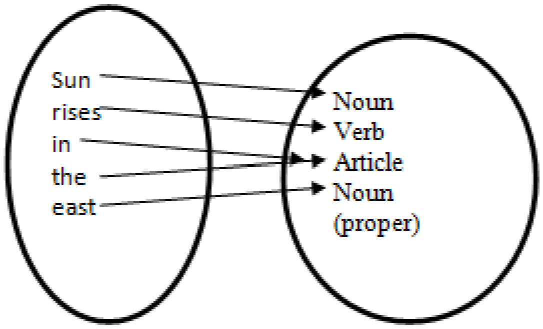

Fig 5: POS tagging representation

V. FEATURE EXTRACTION

The process of identifying and combine certain features is known as feature extraction. It is the preprocessing step of pattern recognition. For extracting features from text term frequency-inverse domain frequency method can be used. This method is abbreviated as TF-IDF in which term frequency is defined as the number of times the term‘t’ appears in a document ‘d’. Document frequency is defined as the number of documents that contains the word‘t’ whereas IDF defines the importance of a term‘t’. The way to calculate TF-IDF is given below.

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

$$\text{IDF (t)} = \log\ \frac{1+|D|}{|d|}$$

Table 2: A term frequency matrix showing the frequency of terms per document.

d/t	t1	t2	t3	t4	t5	t6
d1	0	4	10	8	0	5
d2	5	19	7	16	0	0
d3	15	0	0	4	9	0
d4	22	3	12	0	5	15
d5	0	7	0	9	2	4

TF (d4, t6) = 1+log (1+log (15)) = 1.3377

$\text{IDF (t6)} = \log \frac{1+5}{3} = 0.301$

TF-IDF (d4, t6) = 1.3377*0.301 = 0.403

VI. MULTICLASS TEXT CLASSIFICATION

Classification done for more than two classes is known as multiclass classification [7].

For example: if we want to classify name of the fruits which may be an apple, orange or pear. Multiclass classification makes an assumption that each sample is assigned to only and only one label that is name of the fruit can be either an apple, or orange, or pear but not all at the same time. There are many application areas of multiclass classification which includes identifying the disease of a patient, deciding whether temperature is low, medium, or high. In these areas multiclass classification is a big problem. To solve the problem of multiclass classification methods of binary classification are extended. These include neural networks, decision trees, k-Nearest Neighbor, Naïve Bayes, and Support Vector Machines, but in the following paper we are only discussing the first three.

1. K Nearest Neighbor

The most simple and intuitive pattern classification method is the classification based on the distance function [9]. Its main idea is to use same class focus to represent this class, calculate distance from the classifying samples to the center of gravity, and this class is included in the nearest class.Euclidean distance is used as a distance metric to measure proximity between two pints or two tuples.

$$\boxed{\mathrm{Dist}(X1, X2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2}}$$

KNN is known as lazy learner because it do not uses the training set to do any generalization. As there is an absence of generalization KNN needs to keep all the training data. KNN is easy to implement whereas it suffers from the curse of dimensionality.

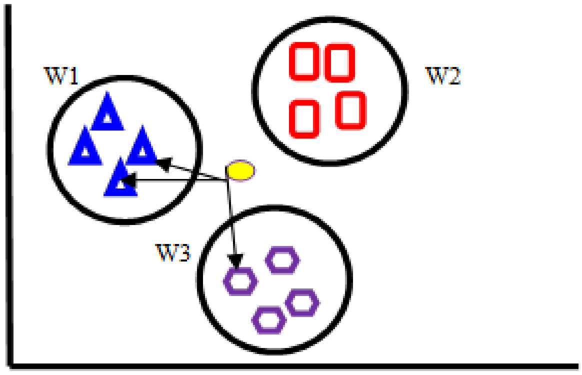

Fig 6: Typical example of KNN classifier

2. Bayesian Classification

It is the most popular technique for effective document classification. Since document can be viewed as the calculation of the statistical distribution of documents in specific classes, a Bayesian classifier first trains the model by calculating a generative document distribution p (d|c) to each class c of document d and tests which class is most likely to generate the test document [8].

3. Support Vector Machines

Support vector machines can be used to perform effective classification since they work on high dimensional data. SVM is used to classify linear and non-linear data. SVM consists of two cases.

Case 1: When data are linearly separable

Black dots represents class c1, y=+1 (sports= L)
Blue dots represents class c2, y=-1(sports =H)

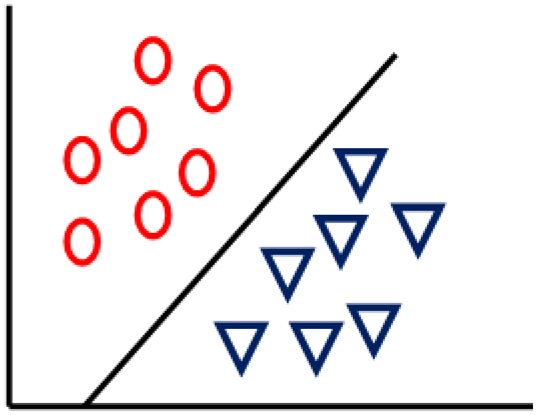

Fig 7: 2-D training case for linearly separable data

Case 2: when data are linearly inseparable

In linearly inseparable data there is no line which can separate the classes.
Red dots represents class c1, y=+1 (sports= L)
Green dots represents class c2, y=-1(sports =H)

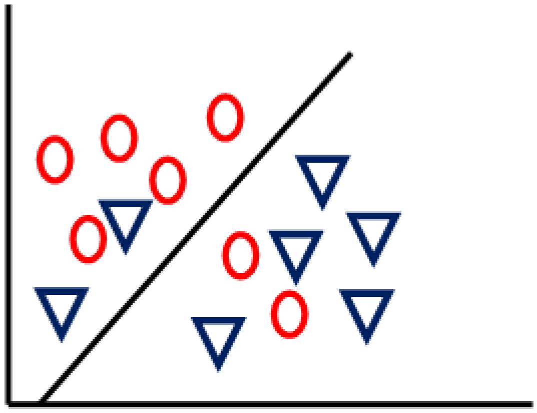

Fig 8: 2-D training case for linearly inseparable data

VII. CONCLUSION

This review presented different approaches to solve the multiclass classification problem in the areas of real world applications. It is seen that it can be solved by extending the methods of binary classification

REFERENCES

[1] Aswini, V., and S. K. Lavanya. "Pattern discovery for text mining." In Computation of Power, Energy, Information and Communication (ICCPEIC), 2014 International Conference on, pp. 412-416. IEEE, 2014.

[2] Sukanya, M., and S. Biruntha. "Techniques on text mining." In Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on, pp. 269-271. IEEE, 2012.

[3] Ghosh, Sayantani, Sudipta Roy, and S. Bandyopadhyay. "A tutorial review on Text Mining Algorithms." International Journal of Advanced Research in Computer and Communication Engineering 1, no. 4 (2012): 7.

[4] Agnihotri, Deepak, Kesari Verma, and Priyanka Tripathi. "Pattern and Cluster Mining on Text Data." In Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, pp. 428-432. IEEE, 2014.

[5] Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Ann Arbor MI 48113, no. 2 (1994): 161-175.

[6] Chrystal, Jincy B., and Stephy Joseph. "MULTI-LABEL CLASSIFICATION OF PRODUCT REVIEWS USING STRUCTURED SVM."

[7] Mehra, Neha, and Surendra Gupta. "Survey on multiclass classification methods." (2013).

[8] Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 3rd Edition, (2011).

[9] Jian-Fang, Cao, and Wang Hong-bin. "Text categorization algorithms representations based on inductive learning." In 2010 2nd IEEE International Conference on Information Management and Engineering, pp. 352-355. 2010.