International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

An Improved Hybrid Stacked Classifier for Multi Label Text Categorization

P. Sree Lakshmi , Kavitha

Abstract: ***Nowadays, the applications of multi label classification are increasing very rapidly with the growth of information technology. One among it is, multi label text categorization which deals with the automatic categorization of documents or comments posted in a social site. Because of the exponential growth of digitization of unstructured categorical data, there is an emerging need for text categorization in particular with multiple labels. Conventionally, it has been solved by either transforming the problem into single class or extending the existing classifiers. An improved hybrid stacked classifier has been proposed to address the challenges in multiple label assignment for text document. The model has been built with three classifiers stacked together with Label Power set by taking class probabilities and a Meta classifier. The experimental results show that the proposed method outperforms well than the existing methods.***

Index Terms: Multi Label Text Categorization, Multi Label Classification, Stacked Classifier, Label Power set, Hybrid Classifier

I. INTRODUCTION

Conventional classification problems associate a single class at a time for an instance. But there are some scenarios in real world where there is a need to relate many class labels for a single case which is called as Multi Label Classification. A common circumstance where this type of classification can be applied is Document Classification where a document is related with more than one class label at a time.

As the availability of the text documents in digital form is abundant with the growth of Internet and information technology, the subsequent need to organize, analyze them has become more focus in research point of view. Text categorization is a process that automatically categorizes text into predefined categories has gained significant attention in recent years.

Due to the fast growing of internet contents, over tens of thousands, even hundreds to thousands of labels can be found in various domains. To list a few, product categorization in e-commerce, web page tagging, medical subject heading, indexing of biomedical documents, Wikipedia and social media sites comment categorization follows multi label text categorization.

Hence, to address the challenges faced by multi label text categorization, a hybrid stacked classifier model ML-HSC is proposed in this research. In this research, different conventional classification models have been stacked by taking class probabilities and Label Power set method which is a problem transformation method for multi label classification. Then a Meta classifier Logistic Regression is used to predict the set of labels to be assigned for an instance. The proposed method has been implemented in Python and evaluated for the performance. The evaluation metrics shows that the proposed method performs well than existing multi label classification methods.

The organization of this paper is as follows. Section 2 discusses the steps involved in Multi Label Text Categorization. Section 3 reviews the studies on multi label classification and text categorization. The proposed methodology is discussed in Section 4 and the results are analyzed in Section 5. Finally, the conclusion section delivers the summary of this research study.

II. MULTI LABEL TEXT CATEGORIZATION

The purpose of Text categorization is to categorize the documents into predefined categories. Each document can be classified into various, just one, or no category at all. The process of text categorization includes the following steps.

Document Representation: As the real-world document mostly available in unstructured form, Document representation is the fundamental phase of representing the document in a different format which is appropriate for analysis. It can be transformed as records with a countable number of attributes. Bag-Of-Words (BOW) is frequently used word-based representation technique [1].

The next step to be done is removal of stop words which are some common words that are irrelevant for analysis. After removing those words, root stem word be find out for each word. Word stemming reduces words to unify all through the document.

Feature Selection or Feature Transformation: Even though a document is represented using Bag-Of-Words (BOW) representation, the number of words will be huge. So, to reduce the dimensionality of feature set, feature selection methods are applied [2]. Thereby classification accuracy can be improved and overfitting can be avoided.

Some of the common methods of feature subset selection for text categorization uses evaluation measures such as Document Frequency, Term Frequency, Mutual Information, Information Gain, Odds Ratio, Chi- square statistic and Term Strength.

Manuscript published on 30 September 2019
* Correspondence Author
P.Sree Lakshmi , School of Computer Science and Applications, REVA University, Bangalore, India.
Kavitha , School of Computer Science and Applications, REVA University, Bangalore, India.

Constructing a Vector Space Model (VSM) After the common preprocessing task, an algebraic model called Vector Space Model can be devised to represent the document as vectors of index terms. Each term in the document has its own weightage which shows the importance of the term in that document.

The entire set of vectors of all the text documents that has been taken for analysis is called a VSM.

Building a Classifier: Now the document is in suitable form for analysis. There are many classification techniques available for text categorization. Some of the commonly used classification algorithms are Decision trees, Naive-Bayes, Rule induction, Neural networks, K- nearest neighbors and Support vector machines.

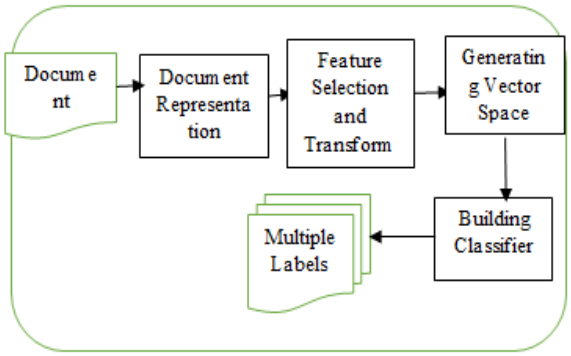

Fig . 1. Multi Label Text Categorization Model

Multi Label classification methods can be classified in two different ways as Problem Transformation method and Adaptation method. The problem transformation methods convert the multiple label problems into one or more than one single-label classification problems. The adaptation methods extend the existing classification algorithms to handle multiple label data straight.

III. LITERATURE REVIEW

Multi Label Classification has its application in many domains as including Semantic Scene classification, Music and Audio categorization, Image categorization, Classification of genes in Bio informatics, and importantly Text Categorization. Different methodologies, techniques and approaches have been developed because of its diverse nature of classification with assigning multiple labels for a single instance.

Conventional text categorization methods taken only terms as Unigrams features [1]. The text was represented as Bag-of-Words (BOW) using unigrams. But in this representation, the association between unigrams will not be considered.

Soumya George K, Shibily Joseph [2] recommends a way to discover co-occurrence feature from reference text of Wikipedia pages.

Koster and Seutter [3] used feature induction methods which involve a combination of single words and word pairs. In the method, nouns are extracted with their modifiers. Phrases are represented by an abstraction called Head/Modifier pairs.

Maciej Janik and Krys Kochut [4] proposed a method which uses semantic graphs for classification task. It deliberates the conversion of a text document into graph of entities which occurs in the document, and then determining the complete categorization of those graphical elements.

Alberto Lavelli et al [5] presented an investigational comparison of two distributional terms representations like the document occurrence representation and the term co-occurrence representation. The results show that the TCOR always outperforms the DOR.

In [6] Juan Manuel et al proposed a procedure, where distributional term representations are used for short-text categorization.By using document occurrence and term co-occurrence statistics the DTR's preserves the contextual information of the terms.

In [7], Harrag et al. directed a proportional study for three text preprocessing techniques, Dictionary Lookup Stemming, Root Based Stemming and Light Stemming for the Arabic text categorization.

Zakaria et al, in [8] proposed a novel method that excerpts broad concepts from Word Net for every term in the document and associates them with the terms in different ways to generate a representative vector.

H. Nezreg and, H. Lehbabin [9] recommend a theoretical representation for text which uses Word net for document categorization. This is achieved based on terms disambiguation.

Shereen Albitar et al, in [10] focus on the approaches for semantic enhancement of abstracted text representation, semantic kernel method and enriching vectors method. They estimated the impact of these strategies on the supervised classification of conceptualized text.

There are two straightforward problem transformation methods that helps to transform force the learning difficult multi label problems into conventional single label classification [11]. Initially it randomly chooses one of the multiple labels from every instance and removes the rest, the second method simply removes each multi label instance from the data set.

The other transformation method reflects each varied set of labels that present in multiple label data set as a single label. Hence, it learns a single label classifier which functions on Power set.

MMAC [12] is another method which follows the model where in classification rule sets are developed using association rule mining.

Jincy and Stephy [13] proposed a ranking algorithm for multi-label classification. This algorithm uses support vector machines which is a linear model that minimizes the objective function (ex: Ranking loss) by preserving a large margin.

In [14-15], they presented different variations of the Support Vector Machine classifier concerned with multi-label classification. Here the margin is improvised in either of the two ways. First the similar negative training instances that is inside the threshold. Second method is removing the negative training instances of the entire class if they are similar to positive class.

IV. PROPOSED ML-HSC METHOD

As followed in the Text categorization process, there are mainly 2 phases involved as Data Representation as pre-processing phase and Text Categorization as Classification phase.

A. Document Representation

Initially the stop words are removed from the document which is followed by word stemming to get only the relevant bag of words for analysis.

As this research study involves in applying stacking technique, multiple classification techniques which should be adopted for the initial step are to be identified.

This study has taken three classification algorithms as k-nearest neighbor, Naïve Bayes classification and Random Forest. All these individual classification methods are trained based on the complete training data set i.e., bag of words of the concerned document.

The class membership probabilities are taken from the first level classifiers. So, all the features are now converted to class probabilities making the length of each record as 2 * number of class labels. This has been taken as input for the next level which is Label Power set method, a transformation method for multiple label classification problems.

Hence, the multi label classification problem is converted to single label classification problem. And then, by taking the Meta features from each individual classification models in the form of ensemble, a Meta classifier Logistic Regression is applied.

Finally, the prediction generated is transformed back to match the multiple class labels so that original problem prediction can happen.

The algorithmic steps for the proposed methodology are as follows:

B. Text Categorization

Let D be the Document
Let Nc be the number of Classes

1. Document Representation
   Step1: SWD ← Stop Word Removal (D)
   Step 2: TD ← Stemming (SWD)

2. Text Categorization
   Let Classifier C = {'KNN', 'NB', 'RF'}
   Where KNN = K-Nearest Neighbor
   NB = Naïve Bayes, RF = Random Forest
   Number of Classifier Models Nm = | C |

Step 1: $\forall c_i \in C$ *Apply* $C_i(TD) \rightarrow P$
Where P = {Pr (KNN), Pr (NB), Pr (RF)} and Class Probabilities from each Classifier

Step 2: Apply Label Power set Transformation (P) → LPD
Where LPD = {X, Y'} Y' is the Transformed Class Label

Step 3: Apply Logistic Regression (LPD) → M
Where M is the Meta Classifier Model

Step4: $\forall d \in TD$ *Transform*$(Y') \rightarrow ML$
Where ML is the set of Multiple Class Labels

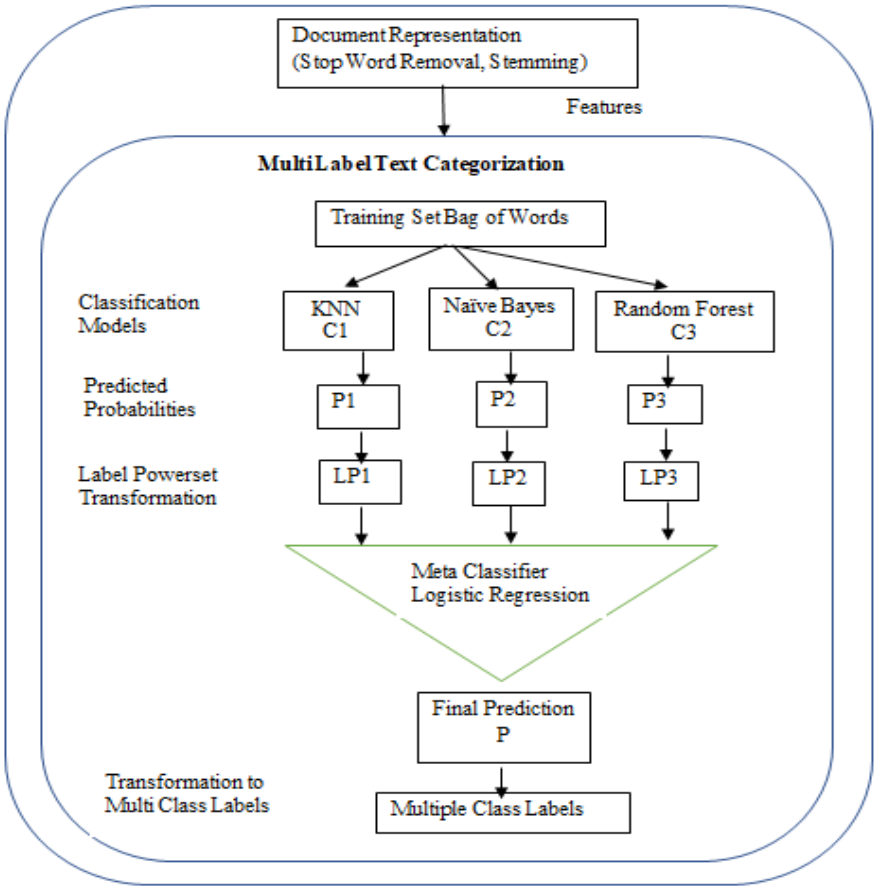

Fig .2 Proposed Hybrid Stacked Classifier for Multi Label Text Categorization

V. RESULT AND DISCUSSION

The performance of the proposed Hybrid Stacked Classifier was analyzed by implementing it in Python. The dataset taken for this study is "Toxic Comment Classification". The dataset consists of Wikipedia comments which are already associated with labels toxic behavior. There are 6 class labels in dataset which are the types of toxicity as toxic, identity, hate, severe toxic, insult, obscene, threat.

The size of training data set is 153164 which are adequate for analysis. The comment length varies from 500 to 5000 in characters and hence, the maximum number of features is kept as 5000.

In order to analyze the performance of the proposed method, it has been compared with two other present algorithms , Label Power set for Multi label classification and ML-KNN (Multi Label k-Nearest Neighbor) algorithm.

Since, in Multi label classification the prediction is finite set of labels, the prediction may be completely correct, partially correct and completely wrong. A different measure called Exact Match Ratio can best evaluate the multi label prediction as follows:

$$Exact\ Match\ Ratio\ EMR = \frac{1}{n}\sum_{i=1}^{n} I\,(Y_i = Z_i)$$

Where, n is the number of instances
I is the Indicator Function
Y is the Actual Labels
Z is the Predicted Labels

Following is the table which shows the performance of the proposed method using EMR as well as conventional classifier accuracy measures.

TABLE 1. Evaluation Metrics Comparison

Evaluation Metrics / Algorithm	Label Power set	MLK-NN	ML-HSC
Exact Match Ratio	0.5754	0.748	0.7913
Accuracy	0.8246	0.8913	0.9257
Precision	0.6198	0.6719	0.7434
Recall	0.6021	0.4502	0.6147
F1-Score	0.4709	0.6185	0.6482

From the above table it is very much evident that the proposed method ML-HSC outperforms well than other common Multi label classification methods Label Power set and MLk-NN.

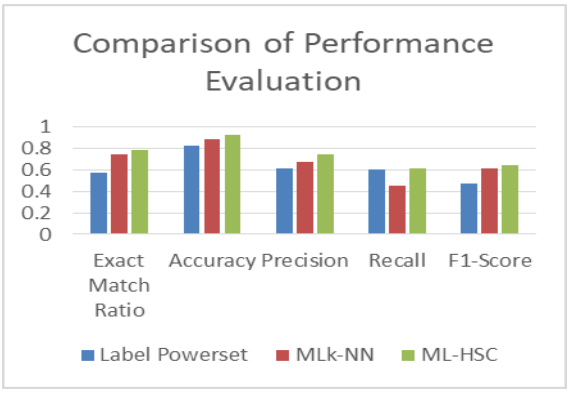

Fig.3. Comparison of performance evaluation

VI. CONCLUSION

Text categorization is one of the critical research issues in several domains as there is a boom in digitalization of the documents and growth of social network. The problem becomes worse when it involves with association of multiple labels. There are different methods and techniques existing for multi label text categorization. To address the nature of assigning multiple labels, a hybrid stacked classifier method is proposed in this study. A stacked classifier with KNN, Naïve Bayes and Random Forest has been developed with Label Power set method by taking class probabilities. Logistic Regression is used as meta classifier to predict the multiple labels finally. The proposed method performs well than the conventional methods in comparison with the evaluation metrics.

REFERENCES

1. Fabrizio Sebastiano, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (2002), no. 1, 1–47.
2. Sou mya George K, Shibily Joseph, Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. V (Jan. 2014), PP 34-38.
3. Cornelis HA Koster and Mark Seutter, Taming wild phrases, Advances in Information Retrieval, Springer, 2003, pp. 161–176.
4. Maciej Janik and Krys J Kochut, Wikipedia in action: Ontological knowledge in text categorization, Semantic Computing, 2008 IEEE International Conference on, IEEE, 2008, pp. 268–275.
5. Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations: an experimental comparison. Thirteenth ACM international conference on Information and knowledge management (CIKM'04). New York, NY, USA, 2004
6. .Juan Manuel Cabrera, Hugo Jair Escalante, Manuel Montes-y Gómez. Distributional term representations for short text categorization .14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2013).Samos, Greece,2013
7. Harrag, F., El-Qawasmah, E., & AL-Salman, A.M.S. (2011). Stemming as a Feature Reduction Technique for Arabic Text Categorization. 10th IEEE International Symposium on Programming and Systems (ISPS)
8. Zakaria Elberrichi1, Abdelattif Rahmoun2, and Mohamed Amine Bentaalah1 Using WordNet for Text Categorization, The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008.
9. H. Nezreg, H. Lehbab, and H. Belbachir, Conceptual Representation Using WordNet for Text Categorization, International Journal of Computer and Communication Engineering, Vol. 3, No. 1, January 2014.
10. Shereen Albitar – Bernard Espinasse – Sébastien Fournier, Semantic Enrichments in Text Supervised Classification: Application to Medical Domain, Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference 2013
11. Ying Yu Witold Pedrycz, Duoqian Miao "Multi-label classification by exploiting label correlations" Elsevier ltd. 2014.
12. Thabtah, F., Cowling, P., Peng, Y.: Mmac: A new multi-class, multi-label associative classification approach. In: Proceedings of the 4th IEEE International Conference on Data Mining, ICDM '04. (2004) 217–224
13. Jincy B Chrystal, Stephy Joseph, "Multi Label Classification of Product Reviews using structured SVM ", International Journal of Artificial Intelligence and Applications, Vol 6, No. 3, May 2015.
14. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence 176(1), 2012
15. 15. Zhang, M.L., Zhou, Z.H.: A Review On Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 2014

AUTHORS PROFILE

P. Sree Lakshmi is a Research Scholar, pursuing her PhD in Computer Science, REVA University, Bengaluru.
Her research interests include Datamining, Machine learning, Artificial intelligence and Evolutionary computing

Dr. Kavitha has done her doctorate in Computer Science. She has published more than 30 papers in her research area of interest in Data Mining, Swarm Intelligence. She is a member in GSTF, IDES, UACEE, IACSIT and AIRCC. She is reviewer for Journals and Technical Programme Committee Member for many international conferences. She has given key note speech in various conferences, seminars and workshops. Her research interest includes Data Mining, Data Analytics and Swarm Intelligence.