Comparing Different Machine Learning Techniques for Classifying Multi Label Data.

D.O.I - 10.51201/Jusst12525
http://doi.org/10.51201/Jusst12525

Shriya Salunkhe
D.J.Sanghvi College of Engineering Mumbai, India.

Kiran Bhowmick
Department of Computer Engineering, D.J.Sanghvi College of Engineering
Mumbai, India.

***Abstract*—** In recent years, multi-label classifications have become common. Multi label classification is a classification in which a collection of labels is associated with a single instance, which may be a variation of the classification of a single label. The problem of huge data is the classification in which each instance is of different kind which further can be identified with more than one class. The various machine learning strategies for classifying multi-label data are discussed in this paper. Many researchers have been carried out that specify the grouping of multiple labels. Here we will compare various classification machine learning techniques that involve two approaches: the adapted algorithm approach and the method of problem transformation. Here we are using naive multinomial bayes and logistic regression. We use certain evaluation metrics to predict the differences as well. Better classification methods are discussed in this paper.

Keywords*— *Multi label classification problem, multi nomial naïve bayes, logistic regression, and evaluation metrics.

I. Introduction

Classification is a method of analysis of data that extracts models that define essential groups of data. In order to predict labels, the data analysis task is classification, where a model or classifier is made. Multi class is a classification of two classes. Multi class classification shows that each sample is assigned to a minimum of one label; a fruit is either an apple or a pear but not both at the identical time. It's one label classification. It contains multiple numbers of classes over two. They're mutually exclusive (independent). [1]

In multi label classification, each example contains different classes. Each class would have more than one label. Each label represents task, this tasks are related. The multiple labels are dependent. Each example depends on multiple labels. Each class can be categorized into different categories. So, these types of problems are called as Multi label classification problem. [2]

The Fig. 1 shows the diagram of multi class vs multi label.

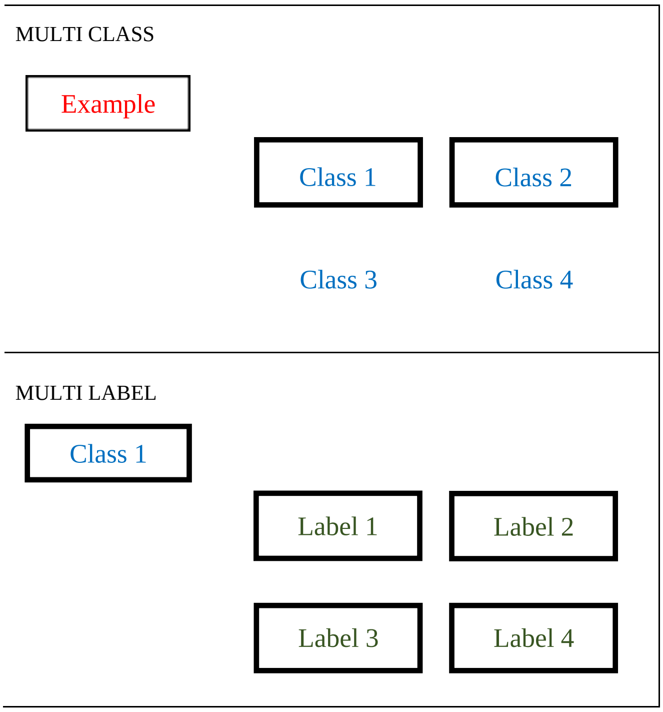
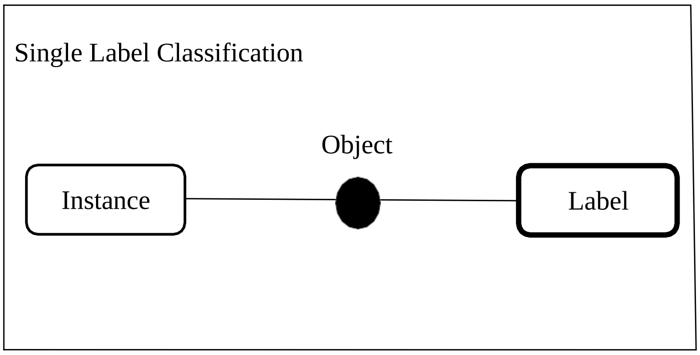


Fig. 1: Multi class vs Multi label

In Fig. 2, Single label classification is where each instance of a dataset is associated with just one class label. [3] Each instance can be associated with one or more class labels. This group of problems is known as Multi-Label Classification. [4]



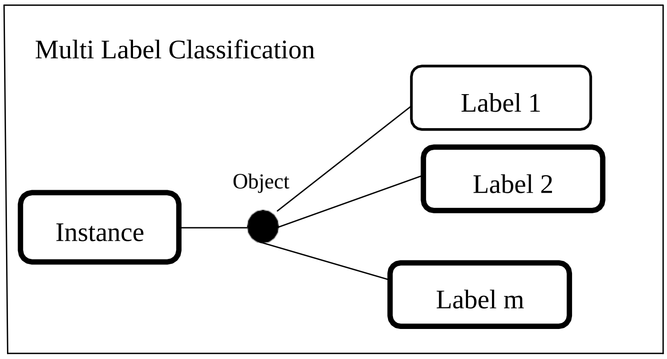


Fig. 2: Single Label vs Multi label classification.

A one label classifier isn’t always able to classify entire information. A technique has to be defined in which classification of data may be done correctly and provides reliable results. The aim is to compare the machine learning techniques for classification of multi label data in data streams. [6]

II. Literature Review

In this paper [1], for classifying multi label data a web variational inference method was used. During this usually to make the ensemble system, random projections are used. In this, Online Variational Inference (VIGO) for multivariate gaussians is employed, which is used to get better performance and multi label data classification. Several difficulties should be able to overcome, to adapt VIGO to multi label learning. Here the following evaluation metrics has been considered: F1 score, accuracy, precision, ranking loss.

Zhe Chu, Peipei Li, Xuegang Hu in [2] proposed COINS for multi label classification. In the world, it is not easy to get labels. The existing multi label data classification algorithms mostly deals with the classification using all labeled data and not with the emerging new classes. Few semi supervised methods are also there. Here the multi label semi supervised classification algorithm known as COINS which support co-training is used to train a base classifier on data. So an ensemble model which is used to adapt to the environment of an outsized number of unlabeled data is generated. Here the next evaluation metrics has been considered: hamming loss, one-error, coverage, ranking loss, and average precision.

Amani M. Alattas in [3] shows models and its challenges. SMLC (Stream Multi label Classification) model is used in many of the processes and events in which the net prediction and its decision is the most important work. This paper shows an AMLCM model (Adaptive Multi label Classification Model) that defines an aggregator concept. The aggregator works as an integrated place. In that a number of sub sections operate in parallel. This is done to update the heuristics and statics values.

Feng Qin, Jun Huang, Xiao Zheng, Zhixiang Yuan, Zekai Cheng, Weigang Zhang in [4] proposes LSML method (Label Specific features for multi label classification with Missing Labels). This is a replacement method. In this we first learn the label correlations. This label correlations can be exploited which gives an unfinished label matrix and can be replaced with a supplementary label matrix. Then, a label specific representation of data for every class label is learned.

Raphael Benedict G. Luta, Renann G. Baldovino, and Nilo T. Bugtai in [5] Paper proposed a system application for the multi-label classification of pH levels. The pH could be a measure of whether a substance is acidic or basic. The utilization of learning methods is less expensive and more reliable for pH level measurement. Here the subsequent evaluation metric has been considered: accuracy.

In this paper [6], Classification is the major issue. In this problem, each instance is of different kind and which is related to more than one class. This problem section is called classification. There are mainly two varieties of classification approaches for multi label: Algorithm dependent and Algorithm independent. Algorithm dependent approach consists of SVM and DT. Independent approach consists of instance and label based methods. The proposed method is AdaBoost and ADTBoost.

In this paper [7], K-labelsets ensemble method which supports mutual information and joint entropy has been proposed. The traditional random k labelsets method has drawbacks. This method have two main drawbacks: i) the imbalanced data may rise for any randomly selected label set, ii) There can be information redundancy and overlap due to dependency relation among labels with similar label set. Here, subsequent evaluation metrics has been considered: subset accuracy and label accuracy. Here the subsequent evaluation metrics has been considered: subset accuracy, label based accuracy.

In this paper [8], MLRBC (Multi Label Rule Based Classifier) for multi label classification is proposed. In this an algorithm combines LP with a rule based ML approach. This has an advantage of strong generalization capability of UCS and its robustness. Here the next evaluation metrics has been considered: hamming loss, recall, accuracy, one error, rank loss.

In this paper [9], the multi-label image classification is being considered where each image is often associated with multiple labels and labels are correlated with each other. In this paper, we propose a model called LMMAL. Here, we train a low-rank mapping matrix to point the mapping relation between the feature spaces. Here the following evaluation metrics has been considered: tuning parameter α.

In this paper [10], SVM (support vector machine) is proposed and also the new multi label learning algorithm is defined as RMLLA (representative multi label learning algorithm). Many approaches have been considered where first an affinity propagation algorithm is used to select respective features and their relationships. Then SVM is used to solve the problem. And then the RMLLA is used to solve the multi label classification problem. Here the subsequent evaluation metrics has been considered: ranking loss, hamming loss, cover measure, one error measure, and average precision measure.

In this paper [11], ELM and OS-ELM is used to make a multi label classifier which is web sequential. There are various world applications of multi label classification. The multi label classification includes the input sample which is related to a collection of target labels. Here high speed nature of ELM and OS-ELM is used. . Here the subsequent evaluation metrics has been considered: accuracy, hamming loss, F1 measure, precision and recall.

In this paper [12], MLC-LR has been used to solve MLC problem. The MLC problem is that the class labels are related to each data instance. The method uses clustering in feature space. Then FP growth algorithm is used to find link between labels. And then logistic regression is used over normalized data. Here the subsequent evaluation metrics has been considered: hamming loss, accuracy, precision, F1 measure and recall.

Understanding and problems to be focused:

Here during this literature review, there are many methods of classifying multi label data. The information sets used contains multiple labels. the issues already focused are exponential number of possible label sets, capturing dependencies between labels, limited time and limited computational resources, the size of data, missing labels, etc. that the main problem to be focused during this report is correct prediction of labels, to guage the accuracy of the techniques which we are using and to classify each labels within the given data set.

III. Implementation

Multi label classification techniques in data streams are as follows:

A. Problem transformation methods

Problem transformation methods change the multi-label problems into multiple single-label problems. Single label classifiers can be classified and then the results from the multiple single classifiers are combined together to get classification result. This adapts your data to algorithm. By limiting each instance to have one label, two-class and multi-class problems both can be emitted into multi- label ones. On the choice hand, the generality of multi-label problems inevitably makes it harder to look out.

Multi nomial Naïve bayes

In this algorithm, a binary mask is taken over multiple labels. One Vs Rest strategy could be used for multi-label learning. It is where a classifier is applied to predict multiple labels as an example. Naive Bayes supports multi-class, but we are in an increasing number of multi-label scenarios. Therefore, we wrap Naive Bayes within the One Vs Rest Classifier. Multinomial Naive Bayes calculates likelihood which is the total number of a word/token (random variable) and Naive Bayes calculates likelihood to be following:

$$P(x=v|c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Naive Bayes considers conditional independence of each feature. Multinomial Naive Bayes classifier is an instance of Naive Bayes classifier. It uses a multinomial distribution for each feature.

B. Algorithm Adaptation methods

AA methods use expansion of base classification algorithm to solve multi label problems. This adapts your algorithm to data. This can be able to adapt a single-label algorithm to provide multi-label outputs. Like specific classifier advantages (e.g., efficiency).

Logistic regression

One straightforward task to do multi-label classification with a multi-class classifier (such as multinomial logistic regression) is to assign each possible assignment of labels to its own class. Logistic regression is applied to data for each cluster per all labels. When an instance arrives in the testing phase, immediately the nearby cluster is identified by using Euclidean distance metric. If the predefined threshold is less than the calculated value, it is both antecedent and consequent labels.

Implementation steps

1. Data Gathering:

The process of collecting information is called as data gathering. The dataset is produced by Jigsaw and accessible at Kaggle [13].

2. Preprocessing:

It is the data mining technique which is used to transform the raw data in a useful and efficient format. First we've got transformed the comments into vector. So we've got applied pipeline to the comments.

3. Training data:

Training data is the part of your data which you use to help your machine learning model make predictions. The dataset is splitted into 80% training sets and 20% testing sets. 127656 instances are used for training out of 159571 instances. It is used for training 6 pipelines individually. The remaining 31915 instances are used for single and combination for performance measure and validation.

4. Classification:

Classification is a process of categorizing a given set of data into classes. At the moment we've got applied the multinomial naïve bayes and Logistic regression method to the information using One vs Rest strategy.

5. Evaluation metrics:

The metrics used in the report are accuracy, precision, recall, f1 score, support.

6. Results:

The result will be predicted using accuracy and other graphs.

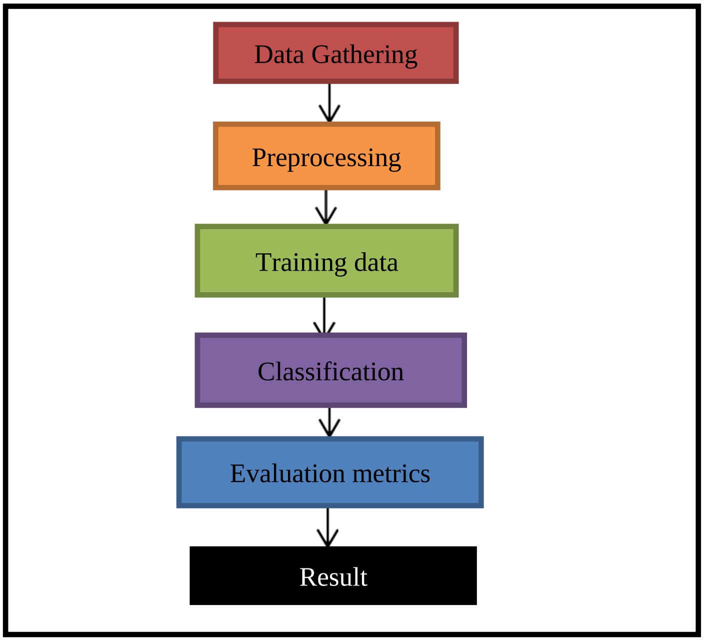


Fig. 3: Flow chart

The above figure Fig. 3 shows the flow chart of the multi label classification. After loading the data from csv file we take a count of number of comments coming under multiple labels. The below figure is of histogram and a bar graph which shows the distribution of comments under multiple labels.

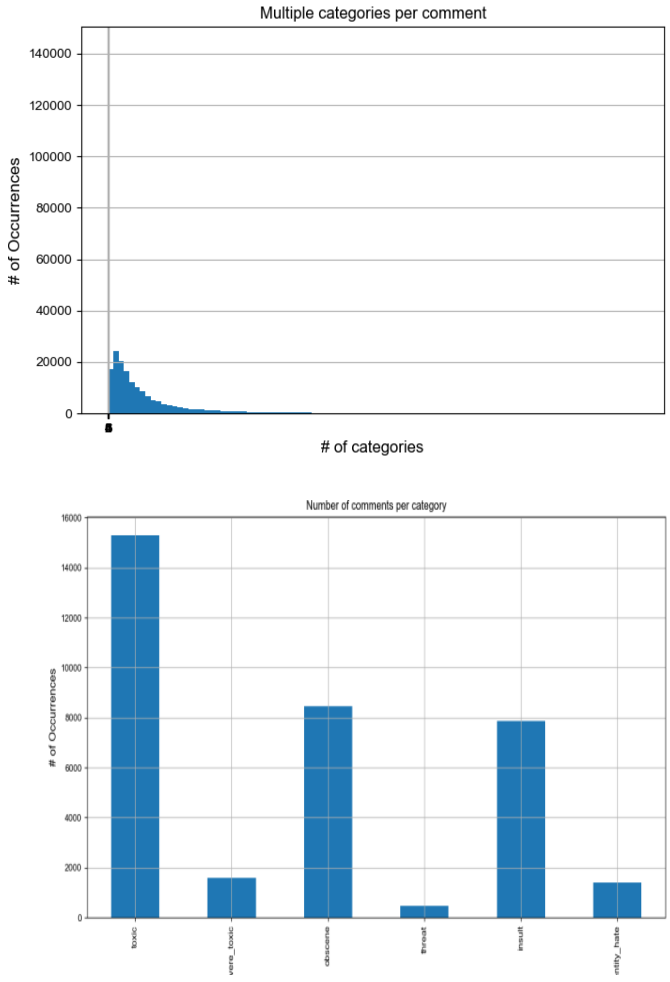


Fig. 4: Count of comments with multiple labels

Confusion matrix and classification chart are being used to describe the various metrics of the classification model. These metrics used are accuracy, precision, recall, f1 score, and support. The confusion matrix is shown as below:

TABLE I. CONFUSION MATRIX

		Predicted Class	
		A	B
Actual Class	a	TP	FN
	b	FP	TN

Where TN is the true negative, TP is the true positive, FP is the false positive, and FN is false negative. Using this confusion matrix we can determine the following metrics:
Accuracy (A): The accuracy is the percentage of correct predictions.

A = TN+TP / TP+FP+FN+TN

Precision (P): The correctly predicted positive observations divided by the total predicted positive observations is called as precision.

P = TP / TP+FP

Recall (R): The correctly predicted positive observations divided by all the observations in the class is known as recall.

R = TP / TP+FN

F-measure: The harmonic mean of precision and recall is called as F-measure.

F-measure = 2*P*R / P+R

Where P is the precision and R is the recall.

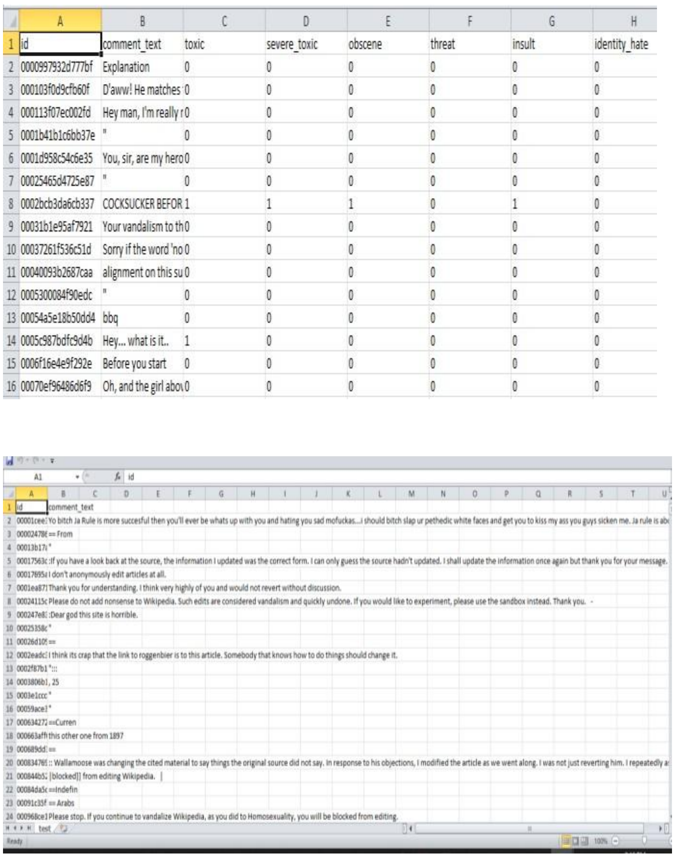
Support: Support is how frequently the items appear in the database.

The classification report is the report which shows the main classification metrics. The predictions which are True and which are False. The result of confusion matrix and classification report has been represented in the tabular form.

IV. DATASET

Dataset

This data set contains of total 159571 instances with comments. The dataset is produced by Jigsaw and accessible at Kaggle [13]. This contains toxic comments which have toxic, obscene, insult, threat and identity hate data.



	A	B	C	D	E	F	G	H
1	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
2	0000997932d777bf	Explanation	0	0	0	0	0	0
3	000103f0d9cfb60f	D'aww! He matches	0	0	0	0	0	0
4	000113f07ec002fd	Hey man, I'm really r	0	0	0	0	0	0
5	0001b41b1c6bb37e	"	0	0	0	0	0	0
6	0001d958c54c6e35	You, sir, are my hero	0	0	0	0	0	0
7	00025465d4725e87	"	0	0	0	0	0	0
8	0002bcb3da6cb337	COCKSUCKER BEFOR	1	1	1	0	1	0
9	00031b1e95af7921	Your vandalism to th	0	0	0	0	0	0
10	00037261f536c51d	Sorry if the word 'no	0	0	0	0	0	0
11	00040093b2687caa	alignment on this su	0	0	0	0	0	0
12	0005300084f90edc	"	0	0	0	0	0	0
13	00054a5e18b50dd4	bbq	0	0	0	0	0	0
14	0005c987bdfc9d4b	Hey... what is it..	1	0	0	0	0	0
15	0006f16e4e9f292e	Before you start	0	0	0	0	0	0
16	00070ef96486d6f9	Oh, and the girl abov	0	0	0	0	0	0

	A	B
1	id	comment_text
2	00001cee	Yo bitch Ja Rule is more succesful then you'll ever be whats up with you and hating you sad mofuckas...i should bitch slap ur pethedic white faces and get you to kiss my ass you guys sicken me. Ja rule is ab
3	00002478	== From
4	00013b17	"
5	00017563c	:If you have a look back at the source, the information I updated was the correct form. I can only guess the source hadn't updated. I shall update the information once again but thank you for your message.
6	00017695	I don't anonymously edit articles at all.
7	0001ea87	Thank you for understanding. I think very highly of you and would not revert without discussion.
8	00024115c	Please do not add nonsense to Wikipedia. Such edits are considered vandalism and quickly undone. If you would like to experiment, please use the sandbox instead. Thank you. -
9	000247e8	:Dear god this site is horrible.
10	00025358c	"
11	00026d10	==
12	0002eadc	I think its crap that the link to roggenbier is to this article. Somebody that knows how to do things should change it.
13	0002f87b1	"::::
14	0003806b1	, 25
15	0003e1ccc	"
16	00059ace	"
17	00063427	==Curren
18	000663aff	this other one from 1897
19	000689dd	==
20	00083476	:: Wallamoose was changing the cited material to say things the original source did not say. In response to his objections, I modified the article as we went along. I was not just reverting him. I repeatedly a
21	000844b5	[blocked]] from editing Wikipedia. |
22	00084da5c	==Indefin
23	00091c35f	== Arabs
24	000968ce1	Please stop. If you continue to vandalize Wikipedia, as you did to Homosexuality, you will be blocked from editing.

Fig. 5: Train and test data set[13]

V. RESULT

The following below are the evaluation metrics which are considered for the classification of the multi label data. The tables below show the evaluation metrics like accuracy, precision, recall, F1 score and support for the respective techniques used:

Multinomial Naïve Bayes:

TABLE II. EVALUATION METRICS FOR ALL THE MULTI LABELS USING MULTINOMIAL NAÏVE BAYES.

Label	Accuracy	Precision	Recall	F1 Score
Toxic	0.92	0.91	1.00	0.95
Threat	1.00	0.99	1.00	0.99
Insult	0.95	0.95	1.00	0.97
Identity hate	0.99	0.99	1.00	0.99
Obscene	0.95	0.95	1.00	0.97
Severe toxic	0.99	0.99	1.00	0.99

Logistic Regression:

TABLE III. EVALUATION METRICS FOR ALL THE MULTI LABELS USING LOGISTIC REGRESSION.

Label	Accuracy	Precision	Recall	F1 Score
Toxic	0.95	0.95	0.99	0.97
Threat	0.99	0.99	1.00	0.99
Insult	0.97	0.97	0.99	0.98
Identity hate	0.99	0.99	0.99	0.99
Obscene	0.98	0.97	0.99	0.98
Severe toxic	0.99	0.99	1.00	1.00

Thus, from these evaluation metrics we conclude the following:

TABLE IV. OVERALL EVALUATION METRICS

Parameters	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Multinomial naïve bayes	0.96	0.96	1.00	0.97
Logistic regression	0.98	0.97	0.99	0.99

The techniques identified very obvious straightforward insults and accordingly tagged them as insults. Racial and identity slurs were detected and was labeled as identity hate. Both the techniques were able to detect toxicity even through spelling mistakes. So both were able to classify complex sentence structures.

Here for multilabel classification, we achieved the best performance of 0.98 accuracy from logistic regression. In addition to recording the best f1 score, logistic regression performed best. The f1 score of logistic regression was 0.99. The multinomial naïve bayes also performed similarly well. In the future, we aim to achieve higher performance and accurate classifications using a more robust model.

VI. CONCLUSION

The matter of multi label classification has been studied within the context of data. The comparison between different methods for classification of multi label data has been represented. The methods compared are two approaches: adaptation algorithm approach and problem transformation based method. In these methods there are certain advantages and certain drawbacks. The downside of adaptation algorithm method is that the training process is irrelevant with the label information. The downside of problem transformation is it doesn't take into consideration the label correlations, so it's shortcoming by addressing the correlation between labels. More data increases model complexity, lowers accuracy.

VII. FUTURE SCOPE

As these methods have certain drawbacks. That the future scope is going to be to feature more multi label methods within the comparison and adapt some multilabel classification methods to accommodate multi label problems where the classes are hierarchically structured. The new methods can include online and active learning methods.

REFERENCES

[1] Thi Thu Thai Nguyen, Tien Thanh Nguyen, Alan Wee-Chung Liew, Shi-Lin Wang, Tiancai Liang, Yongjiang Hu, "An Online Variational Inference and Ensemble Based Multi-label Classifier for Data Streams" , Guilin, China, pp. 302-307, June 7-9, 2019.

[2] Zhe Chu, Peipei Li, Xuegang Hu, "Co-training based on Semi-supervised Ensemble Classification Approach for Multi-label Data Stream", 2019 IEEE International Conference on Big Knowledge (ICBK), pp. 58-65, 2019.

[3] Amani M. Alattas, "Adaptive Model over a Multi-label Streaming Data", IEEE, pp. 18- 22, 2018.

[4] Jun Huang, Feng Qin, Xiao Zheng, Zekai Cheng, Zhixiang Yuan, Weigang Zhang, "Learning Label-Specific Features for Multi-Label Classification with Missing Labels", IEEE, pp. 1-5, 2018.

[5] Raphael Benedict G. Luta, Renann G. Baldovino, and Nilo T. Bugtai, "Multi-label Classification of pH Levels using Support Vector Machines", IEEE, pp. 45-48, 2018.

[6] Mayank Budhiraja, "Multi Label Text Classification for un-Trained Data through Supervised Learning", 2017 International Conference on Intelligent Computing and Control (I2C2), IEEE, pp. 32-34, 2018.

[7] Ran Wang, Sam Kwong, Yuheng Jia, Zhiqi Huang, Lang Wu, "Mutual Information Based K-Labelsets Ensemble for Multi-Label Classification", 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 17-23, 2018.

[8] Shabnam Nazmi, Xuyang Yan, Abdollah Homaifar, "Multi-Label Classification Using Genetic-Based Machine Learning", 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 675-680, 2018.

[9] Anqian Guo, Jian Wu, Victor S. Sheng, Pengpeng Zhao, Zhiming Cui, "MULTI-LABEL ACTIVE LEARNING WITH LOW-RANK MAPPING FOR IMAGE CLASSIFICATION", IEEE, pp. 259-264, July 10-14 2017.

[10] JING-JING LI, FARRIKH ALZAMI, YUE-JIAO GONG, ZHIWEN YU, "A Multi-Label Learning Method Using Affinity Propagation and Support Vector Machine", IEEE Access, vol. 5, pp. 2955-2966, 2017.

[11] Rajasekar Venkatesan, Shiqian Wu, Mahardhika Pratama, "A Novel Online Real-time Classifier for Multi-label Data Streams", 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1833-1840, 2016.

[12] Prathibhamol C P, Jyothy K V, Noora B, "Multi Label Classification based on Logistic Regression (MLC-LR)", 2016 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, pp. 2708-2712, 2016.

[13] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data