iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50626

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Tag Recommendation System for Marathi News Articles by using Multi-label Classification

Parmesh Walunj[1], Krupa Shah[2], Rishi Tank[3], Atharva Mathure[4], Ritesh Shekhar[5], Ms. Deepali Kadam[6]
Information Technology, Datta Meghe College of Engineering Navi Mumbai, India

Abstract: ***Multi-label classification is the variant of a classification problem where multiple labels are assigned to each instance. In multi-label classification, the training set is composed of instances each associated with a set of labels, and the task is to predict the label sets of unseen instances through analyzing training instances with known label sets. This paper demonstrates the use of multi-label classification to determine tags for news articles written in the Marathi Language of India. The proposed study uses Binary Relevance (One vs Rest) technique of multi-label classification to establish the tags for the given input of a Marathi news article. Tag recommendation systems for Marathi news articles can greatly enhance the user experience for readers and help them find the articles that are most relevant to their interests.***
Keywords: Multi-label Classification, One vs Rest, Recommendation System, Marathi Language

I. INTRODUCTION

In this study, we present a novel approach to building a tag recommendation system for Marathi news articles using multilabel classification. We use a combination of Natural Language Processing and Machine Learning Techniques to create a model that can accurately assign relevant tags to Marathi news articles.

Multilabel classification is a type of machine learning problem where an object (e.g. an image, text document, or audio clip) can be assigned multiple labels or categories simultaneously. This is in contrast to traditional classification problems, where an object is assigned a single label from a fixed set of categories. Multilabel classification has a wide range of applications across different domains, including: *Image Classification* - in this context, an image can have multiple labels that describe its contents. For example, an image of a beach could be labeled with "ocean," "sand," "sun," and "beach umbrella" all at once.[1][2][3] *Text classification* - Multilabel classification can be used to automatically categorize text documents into multiple categories. For example, a news article could be classified as "politics," "sports," and "entertainment" all at once.[4][5][6] *Recommendation systems* - Multilabel classification can be used to recommend products, movies, or music based on multiple criteria. For example, a movie recommendation system could suggest movies that are both "dramatic" and "romantic" to a user. [7][8] *Bioinformatics* - In genomics research, multilabel classification can be used to identify multiple functional properties of a protein, based on its amino acid sequence. [9]

One approach to building a tag recommendation system is through multilabel classification with binary relevance (also called as One vs Rest Approach). This means that the system is trained on a dataset of Marathi news articles, where each article is labeled with one or more tags. The binary relevance part refers to the fact that each tag is treated as a separate binary classification problem, where the model predicts whether the tag should be assigned to the article or not.

By using this approach, the tag recommendation system can learn to predict multiple tags for each article, taking into account the relationships between different tags and the content of the article itself. This can lead to more accurate and relevant tag suggestions. Recommendation systems for Marathi news articles are becoming increasingly important in today's world of digital journalism. With the sheer volume of Marathi news articles being published every day, it can be difficult for readers to find the articles that are most relevant to them. A tag recommendation system can help solve this problem by automatically assigning relevant tags to Marathi news articles, making it easier for readers to find the articles they are interested in.

II. MULTILABEL CLASSIFICATION

Multi-label classification can be broken down into two types: i) Problem transformation, ii) Algorithm Adaptation methods. Using *problem transformation techniques*, the multi-label classification problem is reduced to a set of binary classification issues that can be solved by single-class classifiers. The second type *algorithm adaptation method, extends* a specific learning algorithm to handle multi-label data directly.

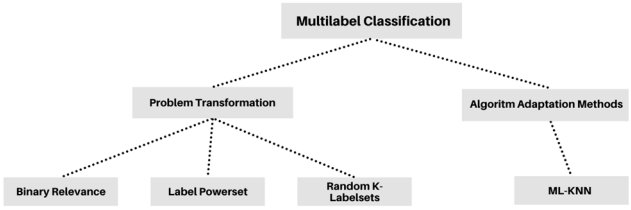


Fig. 1.Types of Multilabel Classification Techniques

A. Binary Relevance (One vs Rest)

Multi-label learning addresses the problem that each example is represented by a single instance and associated with multiple class labels at the same time. Binary relevance is the most intuitive solution for learning from multilabel examples. It works by decomposing the multi label learning task into a series of independent binary learning tasks, one for each class label. Binary related techniques are called One-vs-All (BR-OvA). The BR-OvA method transforms a dataset with k labels into a k single-label dataset and fits a binary classifier for each label. Another binary-related technique is called one-to-one (BR-OvO). BR-OvO transforms a multilabel dataset into multiple binary datasets. Each dataset contains two different labels.

B. Label Powerset

The label powerset method converts a multi-label dataset to a single multi-class dataset by considering each label combination as a unique class. It achieves multi-label classification by assigning an instance to a class that consists of a set of labels. Then a multi-class classifier is trained to assign an instance to one of the above classes. This approach considers correlations between class labels. This approach is commonly called the label powerset method because it treats each member of the power set of labels in the training set as a single label. This method requires a worst-case classifier (2^|C|) and has high computational complexity. However, as the number of classes increases, the number of unique label combinations can grow exponentially. This can easily lead to a combinatorial explosion and thus computation impracticality. Additionally, some label combinations have few positive examples.

C. Random K-Labelsets

The RAKeL(Random K-Labelsets) is another multi-label classification technique that combines various single-label learning models. Each single-label model is built using the label powerset (LP) technique based on randomly generated label subsets of size k. RAkEL can improve the generalization ability and reduce the complexity of the original LP method, but the quality of the randomly generated label subsets can be poor. RAkEL is superior to LP for the following reasons that the resulting single-label classification task is computationally simpler.

D. Adapted Algorithm – ML-KNN

Algorithm adaptation methods for multi-label classification usually focus on adapting single-label classification algorithms to multi-label cases by changing the cost/decision function. Here we use a multi-label lazy learning approach called ML-KNN, which is derived from the traditional K-Nearest-Neighbor (KNN) algorithm. The skmultilearn.adapt module implements algorithmic adaptation approaches for multi-label classification including but not limited to ML-KNN.

III. LITERATURE SURVEY

In recent years, there has been growing interest in developing tag recommendation systems for Marathi news articles to improve the user experience of readers. One popular approach for building such systems is through the use of multilabel classification, where multiple tags can be assigned to a single article. However, the use of one vs rest method in multilabel classification has not been extensively studied in the context of Marathi news articles.

To fill this gap, this study is introduced. In order to get a clear idea several research studies were explored related to the use of multilabel classification in Indian regional languages. A study conducted by Abbas, Syed Zain, et al. (2022) used a combination of Natural Language Processing techniques and neural networks to predict relevant tags for Urdu news articles but the system also supported Hindi and Marathi Languages. The authors reported that their proposed approach achieved high accuracy using the BERT-model.[10]

Another study by Ranasinghe, North et al. (2022) for Offensive Language Identification in Marathi used a mixture of traditional machine learning algorithms and deep learning algorithms to achieve the highest accuracy.[11]
Furthermore, a study by Chy and Siddique et al. (2021) proposed a tag recommendation system for Bangla news articles using Naive Bayes Classifier. The authors reported that their proposed system achieved high accuracy in classification of news articles but did not outperform the other models in terms of precision and recall.[12] Overall, the literature suggests that the use of one vs rest method in tag recommendation systems for Marathi news articles could lead to high accuracy and improved performance. However, more research is needed to explore the effectiveness of one vs rest method in different contexts and with different algorithms. This study aims to contribute to this literature by proposing a tag recommendation system for Marathi news articles using one vs rest method and evaluating its performance against other existing models.

IV. PROPOSED SOLUTION

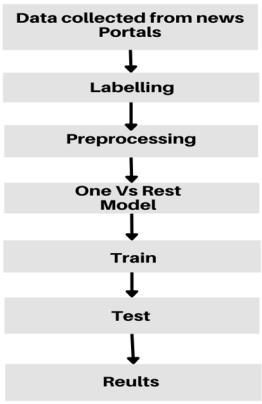


Fig. 2.System Architecture

1) *Data Preparation:* Collect a dataset of documents (e.g., blog posts, articles) and their associated tags. Preprocess the text data by removing stop words, stemming, and tokenizing. Split the dataset into training and testing sets.
2) *Feature Extraction:* Convert the preprocessed text data into numerical features using methods like TF-IDF (Term Frequency-Inverse Document Frequency) or Word Embeddings.
3) *Model Training:* Train two models using the training set: OneVsRest and Binary Relevance.
4) *OneVsRest:* The OneVsRest algorithm is a popular method for multi-label classification problems such as tag recommendation. In this algorithm, a binary classifier is trained for each tag, where the tag is treated as the positive class and all other tags are treated as the negative class. The algorithm trains multiple binary classifiers to predict the presence or absence of each tag in a document. Train a model for each tag, where the tag is treated as the positive class and all other tags are treated as the negative class. This method creates a binary classifier for each tag, predicting whether or not the tag is relevant to a document.

The training process involves the following steps:

a) For each tag, create a binary label vector indicating whether or not the tag is present in each document in the training set.
b) Train a binary classifier (e.g logistic regression, support vector machine) for each tag using the document features and the corresponding binary label vector for the tag.
c) For a new, unseen document, predict the presence or absence of each tag using the trained binary classifiers. The output of the classifiers can be combined using techniques such as thresholding or ranking to select the most relevant tags for the document.

The OneVsRest algorithm has several advantages for tag recommendation, including its ability to handle a large number of tags and its flexibility in choosing different binary classifiers for each tag. However, it may be less efficient than other methods like Binary Relevance for very large datasets or very high-dimensional feature spaces.
Overall, the OneVsRest algorithm provides a powerful approach for tag recommendation that can be adapted to a variety of problem settings and data types.

- *Model Evaluation:* Evaluate the performance of the two models on the testing set using metrics like Precision, Recall, and F1-Score. Compare the performance of the two models and select the one with the best results.

- *Model Deployment:* Deploy the selected model for tag recommendation on new, unseen documents. When a user inputs a document, the model predicts the most relevant tags for the document based on the trained model.
- *Model Monitoring and Improvement:* Monitor the performance of the deployed model and make improvements as needed. This may include retraining the model with updated data or fine-tuning the hyperparameters of the model.

Overall, this proposed solution provides a framework for building and deploying a tag recommendation system using OneVsRest and Binary Relevance methods. However, it is important to note that the specific implementation details and model hyperparameters may vary depending on the specific problem and data at hand.

V. METHODOLOGY

This section describes the experimental setup. Table 1 provides the dataset, and further descriptions of the measurements used to experimentally evaluate the performance of the proposed approach are stated.

A. Data Collection

The dataset comprises the news data gathered from various online Marathi newspapers like Lokmat, Maharashtra Times, and Sakal. Data consists of different categories including Crime (Gunha), Sports (Krida), Manoranjan (Entertainment), Weather (Havaman), Politics (Rajkaran).

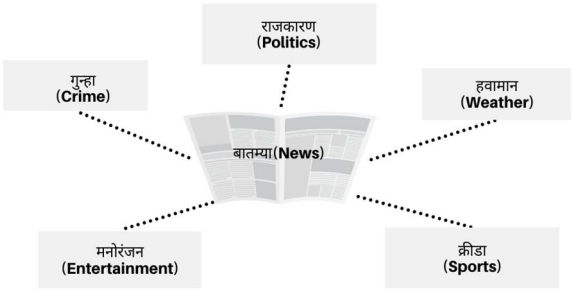


Fig. 3.Tags of Dataset

TABLE I. DATASET

Tags	Categories
गुन्हा (Crime)	खून (Murder), दरोडा (Robbery), सायबर क्राइम (Cyber Crime), बलात्कार (Rape)
क्रीडा (Sports)	क्रिकेट (Cricket), आयपीएल (IPL), टी-20 (T-20), फुटबॉल (Football), कबड्डी (Kabaddi)
हवामान (Weather)	पावसाळी (Monsoon), उन्हाळा (Summer), हिवाळा (Winter)
मनोरंजन(Entertainment)	मराठी सिनेमा (Marathi movies), बॉलीवूड (Bollywood), हॉलीवूड (Hollywood), सेलिब्रिटी (Celebrity), टेलिव्हिजन (Television), चित्रपट (Movie)
राजकारण (Politics)	भाजपा (BJP), शिवसेना (Shiv Sena), काँग्रेस (Congress), जागतिक राजकारण (International politics)

B. Data Preprocessing

1) *Missing Data Removal:* Remove Blank Rows: Removes rows that contain blank cells. Remove duplicate values: Repeated comments are removed as they can affect the response of machine learning algorithms

a) *.Stop word removal:* Stop words are the words which do not contribute to the meaning or to the sentiment orientation of the sentence. For the Devanagari dataset the Marathi stop words are extracted from [30], stored in the csv file. The following sentence explains the stop words removal process.

Sentence: हे एक छान पुस्तक आहे.

Stopwords: {हे, एक, आहे}

Sentence after removing stopwords: छान पुस्तक

b) Vectorization

Vectorization is a phase in feature extraction in Machine Learning. By translating text to numerical vectors, the goal is to extract some distinguishing features from the text for the model to train on.

C. One vs Rest Model

In this approach, we first load the marathi news article dataset and split it into training and testing sets. We then create a logistic regression classifier for each class in the dataset by looping over the possible target classes and training a binary logistic regression classifier on the training data. We use the astype(int) method to create a binary target variable where the positive class is the current target class and the negative class is all other classes. We then store the trained classifiers in a list.

Next, we loop over the instances in the test data and obtain a prediction from each of the trained logistic regression classifiers. We use the predict() method of the logistic regression classifier to obtain the probability of the instance belonging to the positive class, and we store these scores in a list. Finally, we assign the instance to the class that has the highest prediction score.

Finally, we compute the accuracy of the predictions by comparing the predicted class labels to the true class labels in the test data.

VI. ACCURACY & RESULTS

```
t = classifier.predict(x_test[:10])
print("predicted label")
for i in range(10):
  print(predictedLabels(t.toarray()[i]))
# print(predictedLabels(t.toarray()[0]))
# print(predictedLabels(t.toarray()[1]))
```

```
predicted label
['हवामान', 'उन्हाळा']
['हवामान', 'उन्हाळा']
['क्रिडा', 'कबड्डी']
['क्रिडा', 'कबड्डी']
['गुन्हा', 'खून']
['हवामान', 'उन्हाळा']
['क्रिडा', 'क्रिकेट', 'आयपीएल']
['क्रिडा', 'फुटबॉल']
['हवामान', 'उन्हाळा']
['क्रिडा', 'कबड्डी']
```

Fig. 4.Predicted Labels of Test Data

Based on the dataset on which the model was trained and tested the models accuracy comes out to be 0.64 (64%) based on the dataset which was 1000 data points having 30 categories. The accuracy of the model can be increased by extracting more news data by data scraping, manual scraping, news API etc.

The above figure shows the results for 10 news articles taken from the test data. The labels of those news articles are predicted in the output section respectively.

VII. FUTURE SCOPE

There is a significant scope for a tag recommendation system for Marathi news articles using multilabel classification. Here are some potential benefits and applications of such a system:

A. Improved user Experience

A tag recommendation system can help readers find articles that are relevant to their interests quickly and easily. This can improve user engagement and satisfaction with the news website.

B. Increased Engagement

By suggesting relevant tags, the system can encourage readers to explore more articles on the same topic, leading to increased engagement and time spent on the website.

C. Better Content Discovery

The system can help surface articles that might have been overlooked or buried in the website's archives, leading to better content discovery for readers.

D. Improved SEO

Tagging articles with relevant keywords can improve their search engine rankings and visibility, leading to increased traffic to the website.

E. Personalization

A tag recommendation system can be personalized for each user based on their reading history and preferences, providing a unique and tailored experience.

F. Cost-saving

With the help of tag recommendation systems, news organizations can save time and money by reducing the manual labor of tagging articles.

G. Improving Ad Targeting

By understanding user interests and preferences through their search and reading habits, the tag recommendation system can help improve the targeting of advertisements, making them more relevant to users.

Overall, a tag recommendation system for Marathi news articles using multilabel classification has significant potential to improve user experience, increase engagement, and drive traffic to the news website.

REFERENCES

[1] Nasierding, Gulisong, et al. “Clustering Based Multi-Label Classification for Image Annotation and Retrieval.” *2009 IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2009, https://doi.org/10.1109/icsmc.2009.5346902. Accessed 26 Feb. 2023.

[2] Boutell, Matthew R., et al. “Learning Multi-Label Scene Classification.” Pattern Recognition, vol. 37, no. 9, Sept. 2004, pp. 1757–1771, https://doi.org/10.1016/j.patcog.2004.03.009.

[3] Abdel Maksoud, Eman A., et al. “Medical Images Analysis Based on Multilabel Classification.” Machine Learning in Bio-Signal Analysis and Diagnostic Imaging, 2019, pp. 209–245, https://doi.org/10.1016/b978-0-12-816086-2.00009-6.

[4] Katakis, Ioannis, Grigorios Tsoumakas, and Ioannis Vlahavas. "Multilabel text classification for automated tag suggestion." ECML PKDD discovery challenge 75 (2008): 2008.

[5] Yang, Bishan, et al. “Effective Multi-Label Active Learning for Text Classification.” *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’09*, 2009, https://doi.org/10.1145/1557019.1557119. Accessed 5 June 2019.

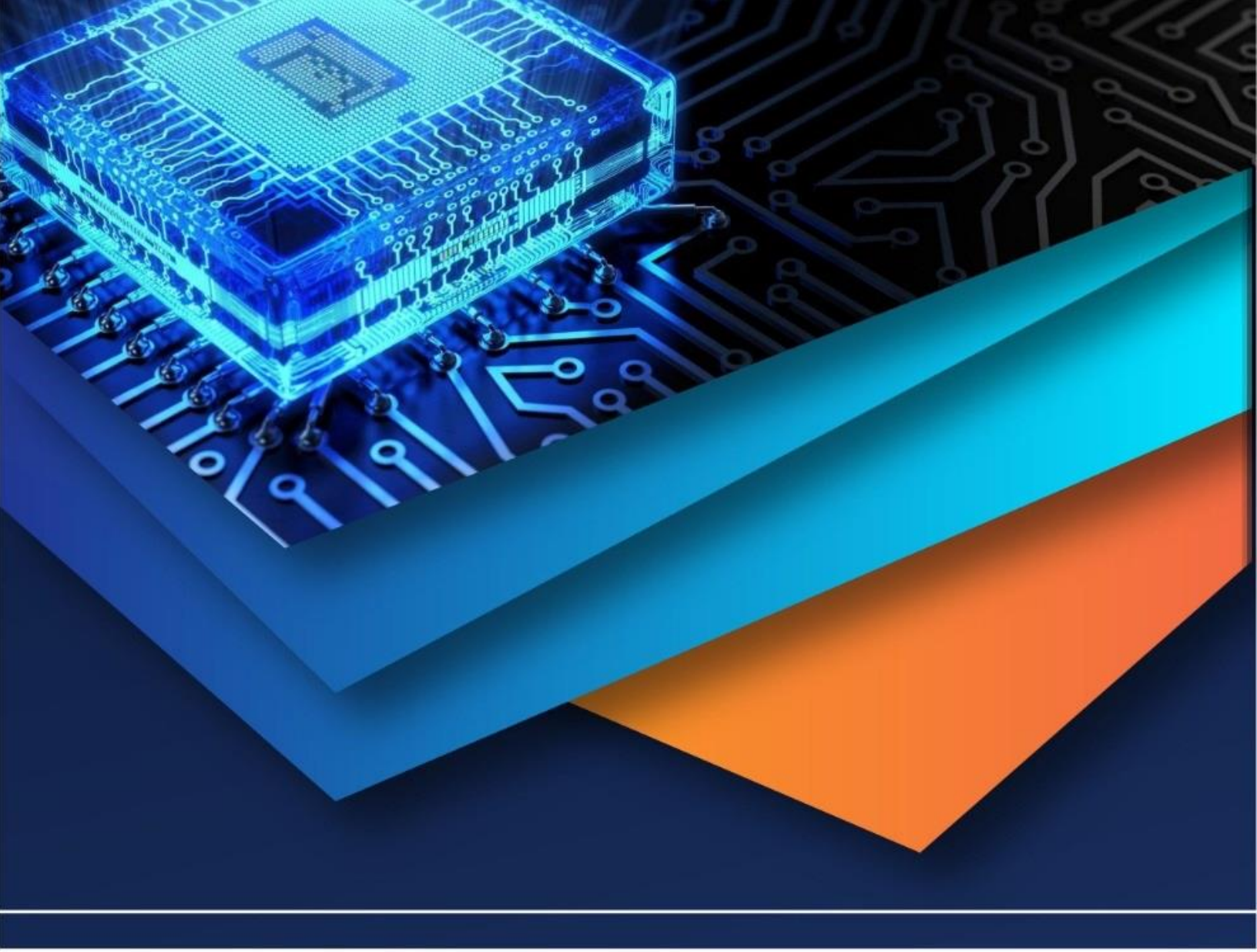
[6] S. Joshi and B. Nigam, "Categorizing the Document Using Multi Class Classification in Data Mining," 2011 International Conference on Computational Intelligence and Communication Networks, Gwalior, India, 2011, pp. 251-255, doi: 10.1109/CICN.2011.50.

[7] Carrillo, D., López, V.F., Moreno, M.N. (2013). “Multi-label Classification for Recommender Systems.” In: , et al. Trends in Practical Applications of Agents and Multiagent Systems. Advances in Intelligent Systems and Computing, vol 221. Springer, Cham. https://doi.org/10.1007/978-3-319-00563-8_22

[8] Rajput, N.K., Grover, B.A. “A multi-label movie genre classification scheme based on the movie’s subtitles.” Multimed Tools Appl 81, 32469–32490 (2022). https://doi.org/10.1007/s11042-022-12961-6

[9] Zafer Barutcuoglu, Robert E. Schapire, Olga G. Troyanskaya, Hierarchical multi-label prediction of gene function, Bioinformatics, Volume 22, Issue 7, April 2006, Pages 830–836, https://doi.org/10.1093/bioinformatics/btk048

[10] Abbas, Syed Zain, et al. "Urdu News Article Recommendation Model Using Natural Language Processing Techniques." *ArXiv:2206.11862 [Cs]*, 29 May 2022, arxiv.org/abs/2206.11862. Accessed 26 Feb. 2023.
[11] Ranasinghe, Tharindu, et al. "Overview of the HASOC Subtrack at FIRE 2022: Offensive Language Identification in Marathi." *ArXiv:2211.10163 [Cs]*, 18 Nov. 2022, arxiv.org/abs/2211.10163. Accessed 26 Feb. 2023.
[12] Chy, A. N., et al. "Bangla News Classification Using Naive Bayes Classifier." *IEEE Xplore*, 2014, ieeexplore.ieee.org/abstract/document/6997369. Accessed 14 Apr. 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY