iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5102

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: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

News Article Category Predictor

Navya Y[1], Apoorva N[2], Sudarashan K[3]
[3]*Associate Professor,* [1, 2]*Computer Science and Engineering Department, Srinivas institute of technology, Valachil*

Abstract: ***News article category predictor focuses on designing and developing an application to predict the category of news article intended to upload in the newspaper. This paper presents the algorithm for classification of articles into different genres based on the information retrieval from the article. The algorithm proposed here helps to classify the topic and discover the new topic as they appear in the content or the report provided. The algorithm explained here basically uses keyword extraction algorithm that is applicable to any of the languages.***
Keywords: News, Category classification, Information retrieval, Genre predictor, Article classifier

I. INTRODUCTION

Every newspaper or the digital news applications that we use, sort news according to its genre. Categories are high level groupings that allow easier navigation of the articles. The prediction technique makes easier the work of categorizing news articles. If a specific topic is related to more than one category then the algorithm must predict the relative percentage match to each category.
The combination of topics and categories create a hierarchical structure. For example an article about baseball can be put under sports category also under the achievement's category. So, we can say that there is no one to one relationship between the topic and the article. Its always a one to many relationship, meaning that one topic can belong to many categories.
The category classification problem can be seen in text classification or the document classification problems. But dealing with news is different than dealing with a document classification. Here the new documents must be processed as they appear. The new report document may contain information that is never seen before. Hence news genre classification requires a dynamic classification which is adaptive to latest news and predicts if it belongs to a new category.
This paper proposes algorithm for category classification that seems to be more effective and more precise. They meet the requirements mentioned that is classification, discovery and relative percentage to each category if they have one to many relationships. In addition to this the algorithm can be developed and implemented to deal with different languages. The paper will further continue as follows: background information and relative work, then the algorithm for category classification and finally present the conclusion and the future scope.
The below given figure describes the categorization hierarchy.

Fig 1.Categorization Hierarchy

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

II. BACKGROUND

Background details of the category Classification algorithm is as follows:

A. Category Classification

In news article classification, multi-label text classification is a problem. The goal is to assign one or more category label to a news articles. For each category, a classifier is used to give either "yes" or "no" answer on which the category should be assigned to a test. It's the example of using binary classifier. Some of the standard algorithms for text classification are Naïve Bayesian Classifiers [1] and support vector machines [2]. Some other Approach to multi-label classification includes boosting [3] and mixture models trained by the em algorithm [4].

A category classification algorithm for news, besides having the required high precision it should also be easily updated. This is because continuously there will be change in the category and events occurring at real world. These will be added to the classifier. By easily updatable, we mean that updating the classifier requires a simple non-exhaustive retraining or no retraining at all.

The previously used methods typically require both positive and negative examples for training data. The initial set of selected training data requires that each article is assigned to at least one positive label. Support Vector Machines offer performance, but they are slow to train and update the training data is not really viable. Category classification deals with broad grouping and such categories are classified on primitive set basis. So the first step we need to initiate in this algorithm is identifying the primitive sets. Since news is not just related to one particular country or culture we must assign categories that is applicable to all the country and culture.

B. Algorithm Overview

The proposed algorithm builds a category model to describe a category. The category model is made up of a category name, total number of documents, document counter and a list of associated keywords. Each entry in the keyword list contains stemmed keyword, the shortest non-stemmed version of the keyword and the number of training documents it appeared in. The keywords are extracted using the keyword extraction algorithm and can extract high quality keywords from a single document without a document collection or corpus statistics. Moreover, it is able to work on any language that has basic morphological analysis tools. The algorithm extracts noun phrases instead of unigrams to use as keywords. It uses in document statistical information about the noun and the individual words to weigh the extracted keywords. It was found that this approached had some advantages over using surrogate corpora when there was no existing document collection to use. A classifier is trained for each category. Each classifier can be trained independently of each other, which allows for easy updating of category information. The classifiers are not binary, meaning they do not give a "yes" or "no" answer. Instead they give an estimate of the likelihood that the article is in the category. The likelihoods from all the categories are used to determine which of the categories should be assigned to the article.

C. Training

To create a predictor which classifies the news articles based on the category first it needs to be trained. From these trained articles keywords are extracted using the keyword extraction algorithm previously mentioned. The number of keywords and the number of documents trained need to be kept in track. The numbers need to be recorded. This is the only required information for the classifier. Only the documents which can be assigned as positive to a particular category need to be trained. Updating the classifiers can be done by increasing the counter.

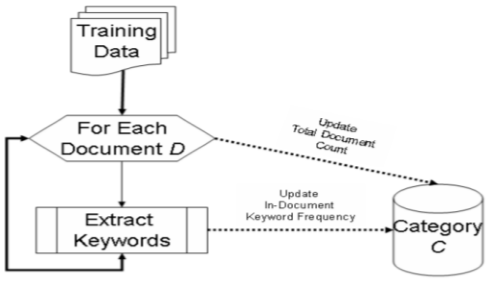


Fig 2.Training overview

Figure 2 shows the process of training a data. A set of documents are provided as data sets to train. Each time a article is given as input the total number of documents that are trained needs to be incremented. After that the classifier extracts the keyword from the article. The frequency of the keyword's found in the article needs to be recorded. If a keyword is extracted and if it is found in the keyword set the count need to be incremented. If its not found in the keyword set it need to be added to the keyword set. This will help to easily correct and update the misclassified categories.

D. Classification

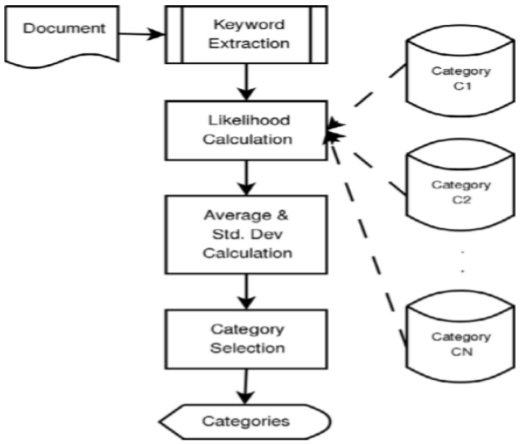


Fig 3. Classifying process

The above figure defines the process used in classification. The whole process of classifying involves 4 steps. The document whose category must be predicted is given as the input document. Keyword extraction will be the second step. Depending upon the extracted keyword the category likelihood is calculated, and then a dynamic threshold is created. Based on the result a category is selected and assigned to that document.
Likelihood can be calculated as follows:

$$(1) \qquad Likelihood(c_j|A=\{k_1,k_2,\cdots,k_n\}) = -\sum_{i=1}^{n} P(k_i|c_j)\log\left(P(k_i|c_j)\right)$$

Fig 3. Formula to calculate the liklihood

In the equation, cj is a category, A is the given article defined by a set of keywords and P(ki|cj) is calculated using the "In-ocument" and the "total number of documents" count. After calculating the likelihood the dynamic threshold is calculated where threshold is the mean and standard deviation of all liklihoods.

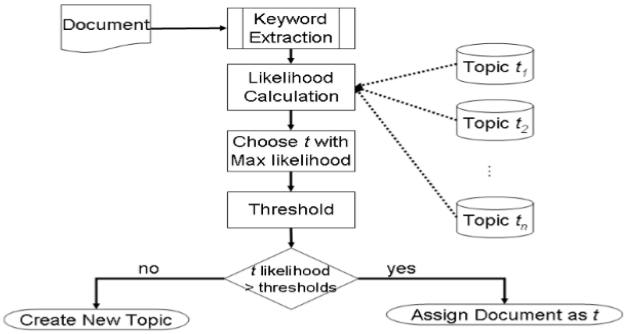


Fig 4.Topic discovery

If a new topic is found while classifying then a new topic is created. The above figure shows the flow while discovering an new topic.

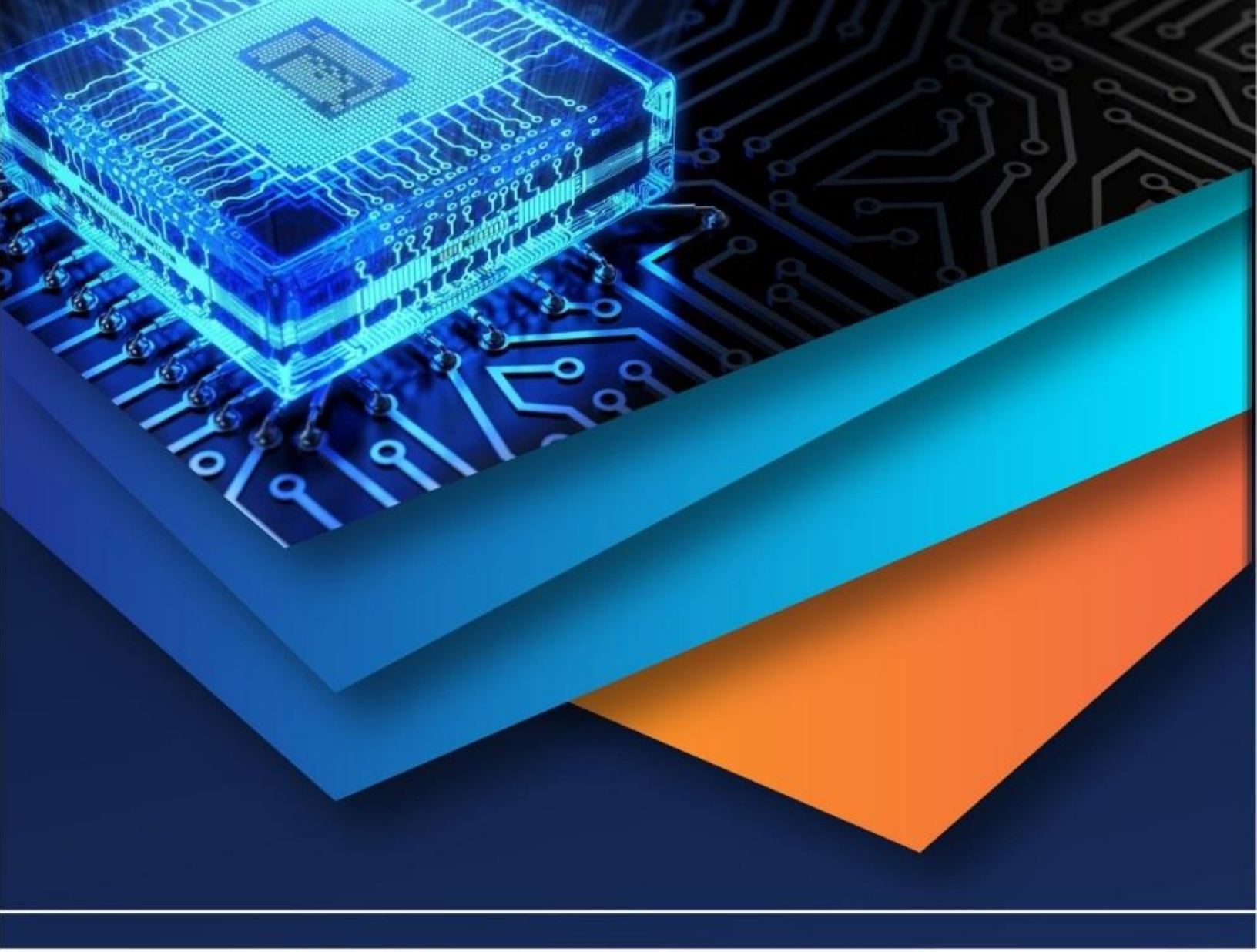
III. CONCLUSION AND FUTURE WORK

This paper presents algorithm for categorizing the article into different genres using topic discovery and classification. The news domain has a lot of challenges. Dealing with online news demands online classification and using this classification in digital application requires more precision and better performance. The algorithm presented in this paper is based on keyword extraction that is capable of dealing with multiple languages. This paper proposes that even the simple algorithms can be used to develop a better results. The category classification algorithm can train oits(optimised image segmentation) classifiers independent of each other and can be easily updated.

The future scope for this project is to test the algorithm on a large corpora so that efficient use can be made from this. In addition we can always think of improvising the algorithm so that the fragmentation and the categorization becomes more precise and acceptable.

REFERENCES

[1] McCallum, A. and K. Nigam, A comparison of event models for naive bayes text classification, in: AAAI/ICML-98 Workshop on Learning for Text Categorization, 1998
[2] Tong, S. and D. Koller, Support vector machine active learning with applications to text classification, in: P. Langley, editor, Proceedings of ICML-00, 17th International Conference on Machine Learning (2000), pp. 999–1006.
[3] Schapire, R. E. and Y. Singer, Boostexter: A system for multiclass multi-label text categorization, Machine Learning 39 (1998), pp. 135–168
[4] McCallum, A., Multi-label text classification with a mixture model trained by em, in: AAAI'99 Workshop on Text Learning, 1999.
[5] Category classification and topic discovery of Japanese and English news articles by David B. Bracewell



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY