International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 12 | December - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Tag Suggestion System Using Multi-Label Text Categorization

R. JOSHITHA[1], Dr. DURGESH KUMAR[2]

[1]SCOPE & Vellore Institute of Technology, Student
[2]SCOPE & Vellore Institute of Technology, Professor

Abstract - The study of text analysis is a relatively new discipline. The technique of evaluating and extracting information from textual data is already being used in a variety of fields, including marketing, product management, academics, and governance. The act of selecting acceptable categories from a predefined set and applying them to natural language texts is known as text classification or text categorization. Text classification, to put it simply, is the process of obtaining general tags from unstructured textual content. The generation of all generic tags is done using a list of pre-defined groupings. Users will find it easier to find and navigate your website or application if you categorize your information and items.

We'll be working on a text classification model in this article that examines the textual description of a query and generates numerous labels for it. Multi-Label Text Classification, a subset of multiple output models, will be used to construct a multilabel text classification strategy for a tag recommendation system.

Key Words: text classification, categorization, natural language processing, textual data, generic tags.

1. INTRODUCTION

Text analysis is a relatively recent area of study. In industries including marketing, product management, academics, and government, the process of analyzing and extracting information from textual data is already in use. Text classification, often called text categorization, is the process of assigning appropriate categories to natural language documents using a pre-established collection. In plain English, text categorization is the process of extracting generic tags from unstructured textual material. The source of all generic tags is a set of pre-defined groupings. Users will have a simpler time navigating your website or application if you categorize your content and things.

In Multi-Label Text Classification, one sample may be a member of several classes (MLTC). As seen, the majority of MLTC tasks include dependencies or correlations between labels. Current methods usually ignore the relationship between labels.

The dataset "StackSample:10% of Stack Overflow Q&A" will be adopted. The problem statement is for a multilabel text categorization method. We'll be developing a text classification model that analyses the textual explanation of a question and generates several labels for it. For a tag recommendation system, which is a subset of many output models, we'll employ multi-label text classification to construct a multilabel text classification approach. Text pre-processing is applied to the text data, and the cleaned data is then loaded for text classification. We'll model the data using text vectorization, encode the tag labels using multilabel binarization, and compare the results using classical classifiers like SGC, multinomial naive Bayes, random forest classifier, and others.

Classification is a type of supervised machine learning. A classification challenge is a predictive modeling issue where a class label is predicted for a certain input sample. It determines the data point category and is most helpful when the output includes both discrete and finite values.

There are four distinct categories to choose from:

➔ Binary classification

➔ Multiclass classification

➔ MultiLabel classification

➔ Imbalanced classification

The accuracy of the method could be increased by using multi-label text classification, such as Keras multi-label text classification. Multi-label text categorization may also be used with XLNET, GPT-2, and GPT-3.

2. LITERATURE REVIEW

A. Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization.

Guibin Chen1, Deheng Ye1, Zhenchang Xing2, Jieshan Chen3, Erik Cambria. (2017)

The job of classifying a textual document into one or more categories (or labels) is known as multi-label text categorization. This work may often be broken down into two sub-tasks: text feature extraction and multi-label classification.

Textual elements the majority of current research concentrates on the distributional representation of individual words, which is yet pre-processed and tokenized from the original raw text data, rather than expressing the entire text in one step, such as tf-idf weighting for a document.

Classification using multiple labels. The next step is to do multi-label classification for text characteristics because they may be properly represented as a feature vector. A high-dimensional feature vector called X E Rm is allocated to a subset of the label space Y with L potential labels representing each input occurrence. Learning a function from the training data is the problem at hand.

The two components of our method are the RNN for multi-label prediction and the CNN for extracting text characteristics. A global fix-length feature vector for the input text is extracted using CNN. The "initial state" or previous knowledge of the RNN is then established and used to predict a sequence of labels using these feature vectors.

B. Using Reddit Data for Multi-label Text Classification of Twitter Users' Interests

Angel Fiallos,Karina Jimenes (2019)

Model for multi-label text classification: The multi-label text classification has been used for a variety of jobs and projects, including classifying enterprises. Indexing document sets and analyzing text for the sentiment. A two-layer surface neural network-based prediction model called Word2vec is trained to recover the linguistic contexts of words. This approach is quite effective at extracting word embeddings from unprocessed text. Word2Vec creates a vector space from a big text corpus.

A method for categorizing data automatically is suggested, taking into account Reddit and Twitter data. To train a model with labeled data, a dataset of 42.100 publications from the well-known forums site Reddit is first gathered. The trained model is then used to forecast users' subjects of interest using a dataset of tweets from 1573 accounts, with an average of 100 tweets per user. Finally, with an average precision of 75.62 percent, we were able to automatically categorize data.

C. A Multi-Label Text Classification Method Based on Labels Vector Fusion

Yang Tao, Zhu Cui, Zhu Wenjun (2018)

The term "traditional text classification" refers to text that is classified using a single label, meaning that each document only falls under one category. A document frequently has several labels because of how complicated it is. The demands of text categorization nowadays cannot be satisfied by single-label classification. The dimension disaster, local optimal solution, and overlearning problems of text vectors are solved using CNN in a novel text multi-label learning method, and the underlying features are combined to provide a more abstract high-level representation. CNN is used to create the mapping relationship between the label vector and the text. The word embedding set of labels from the network output is utilized to obtain the nearest neighbor in the text, and the anticipated text's nearest neighbor serves as the multi-label.

By combining label word embedding, the multi-label text convolutional neural network (ML-TextCNN) is created. The semantic information and location information of neighboring words in the text are extracted by convolution and pooling operations from the text matrix created by word embedding and fed into the ML-Text CNN. The closest neighbor in the word embedding space of the original labels is then obtained using the output of ML-TextCNN as the semantic vector of the predicted labels. Finally, the text's numerous prediction labels are created using the closest neighbor labels. The text multi-label was evaluated using the experimental data set.

D. Ensemble multi-label text categorization based on rotation forest and latent semantic indexing

Yang Tao, Zhu Cui, Zhu Wenjun (2018)

Four essential concepts form the basis of an ensemble multi-label classification technique for text categorization: (1) executing Latent Semantic Indexing; (2) randomly partitioning the vocabulary; (3) document bootstrapping; and (4) using BoosTexter as a potent multi-label base learner. The Rotation Forest paradigm is naturally extended to multi-labeled data by MLRF.

The goal of MLRF is to create multi-label classifiers that are accurate and diverse. The key concept is the application of the feature extraction technique to various random feature set splits to create new characteristics for each base multi-label classifier in the ensemble. We picked LSI because it is thought to be successful at resolving lexical matching issues by creating conceptual indexes rather than specific concepts (or words) for retrieval in a database of texts.

E. An Empirical Study for Class Imbalance in Extreme Multi-label Text ClassificationBassam

Sangwoo Han, Chan Lim, Bonggeon Cha, Jongwuk Lee (2021)

Finding the most pertinent multi-labels from a text corpus with millions of labels is known as extreme multi-label text classification (XMTC). Existing research recommended several strategies to address the class imbalance issue utilizing various loss functions (such as the focal loss function) and data augmentation (i.e., mix-up).

On three datasets for the class imbalance problem, they studied focus loss and a mix-up with RNN-based and transformer-based deep XMTC models. To see improvements in tail label prediction as evaluated by propensity-scores accuracy, we run tests using the focused loss.

Two measurements were used to assess the experiment. The multilabel problem's class imbalance has a solution suggested for it. Given that labels seldom have large inverse propensity scores, multiplying P@k by the inverse of the propensity score might be seen as a way to benefit tail labels.

F. A Novel Method for Efficient Multi-Label Text Categorization of research articles

Rajni Jindal, Shweta (2018)

A prominent area of research in the field of text mining is text categorization. The bulk of documents in the actual world has several labels. In this paper, they have introduced a unique method for efficiently and automatically classifying multi-label text documents. The suggested method is founded on lexical and semantic ideas. Tokens in the text documents are identified using the accepted IEEE taxonomy. To study the semantic connections between tokens, WordNet, the standard lexical database, is used.

One of the most fundamental classification methods for data mining is KNN. Despite its efficacy and efficiency, it has several drawbacks that cannot be ignored. Real-world data is also by its very nature confusing. Fuzzy KNN, which is based on fuzzy membership, was developed to overcome this problem. However, it was quite time-consuming because the membership is established during the categorization step. They designed MFZ-KNN, a modified fuzzy-based KNN approach, to address this problem. In this method, fuzzy clusters are produced during the preprocessing stage, and the membership of the training data set is calculated using the centroid of the clusters. The complexity of time is therefore greatly reduced.

G. Feature ranking for enhancing boosting-based multi-label text categorization

Bassam Al-Salemi∗, Masri Ayob, Shahrul Azman Mohd Noah (2018)

Multiple labels can be learned well using boosting algorithms. Boosting algorithms, which function as ensemble learning algorithms, construct classifiers from a collection of tenuous hypotheses.

With a rank-and-filter approach to issue management, RFBoost was developed. In each learning iteration, it first ranks the training features before filtering and using just a subset of the highest-ranked features to create the weak hypotheses. Concerning AdaBoost, this step guarantees RFBoost will learn more quickly. This research examines seven feature ranking techniques to enhance the performance of RFBoost.

Numerous techniques have been developed and applied to address multi-label classification issues, including binary relevance (Boutell et al., 2004), classifier chains (Read et al., 2011), label powerset (Tsoumakas & Vlahavas, 2007), ranking by pairwise comparison (Hüllermeier et al., 2008), and calibrated ranking by pairwise comparison (Fürnkranz et al., 2008).

The maximum posterior principle is used in the multi-label kNN (MLkNN; Zhang and Zhou, 2007) method, which was modified from the conventional kNN algorithm for multiple-label classification.

H. Handling imbalanced datasets in multi-label text categorization using Bagging and Adaptive Boosting

Genta IndraWinata, Masayu Leylia Khodra (2015)

Multi-label text classification algorithms might not produce the best results because classifiers are constrained by the majority of the data and ignore the minority. The problem is addressed in this study by using the Bagging and Adaptive Boosting techniques to improve text categorization performance.

In this study, two approaches to deal with unbalanced datasets—bagging and adaptive boosting—are assessed and compared using multi-label classifiers. Both methods can improve categorization performance, particularly for J48 and SMO.

The best outcomes are obtained when using SMO with Bagging.ML-LP for subset accuracy and example-based accuracy. The best of all is the micro-averaged f-measure value for SMO with Bagging.ML-BR.

On the other side, the J48 algorithm with AdaBoost.MH has the lowest hamming loss value.

3. PROBLEM STATEMENT

Three components make up a Stack Overflow question: the title, description, and tags. Using the information in the title and description, we ought to be able to automatically suggest tags related to the topic of the query. These identifiers play a significant role in determining the question's type and field. Other Stack Overflow users who have already answered several questions on those tags are also recommended.

The success of the business depends on this. The more accurately Stack Overflow can predict these tags, the more effectively an ecosystem can develop to direct the right question to the appropriate audience.

Text categorization (sometimes referred to as text classification) is the process of assigning pre-defined categories to free-text writings. It can provide conceptual representations of document collections and has useful real-world applications. Academic papers are frequently categorized by technical fields and sub-domains, while patient reports in healthcare organizations are frequently indexed from multiple perspectives, using taxonomies of clinical conditions, different kinds of medical procedures, insurance coverage codes, and so forth. For example, news stories are frequently organized by subject categories (topics) or geographic codes. Another frequent use of text categorization is spam filtering, which divides email communications into two groups: spam and non-spam.

Depending on the blog platform you're using, specific tags might benefit your blog in several different ways. The most commonly used tags are displayed in a larger size in tag clouds, which are plugins that some blogs use to display every tag that has been applied to an article in the sidebar. Most blog systems often provide a web page for each weblog category. What does this mean, exactly?

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit the use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Imagine that one of your practice areas is auto accidents, and that "motorcycle accidents" is one of the tags you frequently employ. The probability that search engines will find your blog increases each time you write a blog entry with that tag since it is added to a page that displays all blog entries with the tag "motorcycle wrecks."

Weblog tags are mostly used to organize your content into categories. Google believes that everything on the internet should be neat and orderly. Think of blog tags as discrete buckets that categorize posts. You wouldn't want to add tags only for the sake of adding them, though.

A particularly successful social media tactic for visual marketing may involve the use of photo tagging. It can strengthen the context of your article and aid in its popularity. Image tagging is the process of adding descriptive information to a photo when it is submitted.

You gain a sizable competitive advantage when you incorporate popular keywords into your hashtags and picture labeling. You may compile a list of hashtags that are currently trending on social media for ease of use and rapid access. You'll have a better chance of connecting with a wider audience of individuals who are already interested if you use trending hashtags.

Social media is currently the best platform for interacting with your target audience. Its impact is being used by several educational institutions to enhance admissions and placements. They don't always fully utilize digital marketing aspects like the idea of tagging, though.

In this piece, we explained how hashtags function and how organizations can use them to successfully increase their social engagement. Nowadays, more and more people use hashtag

searches to get information. Scholarly brands should be aware of how to maximize this innovative research methodology.

4. ARCHITECTURE DIAGRAM

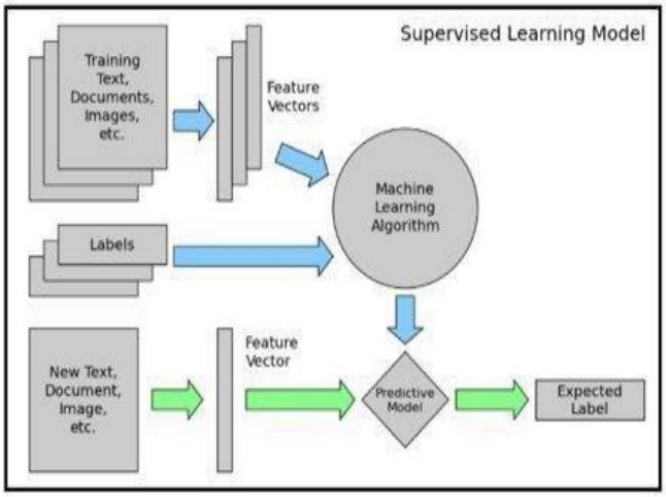

Fig -1: Architecture Diagram

5. FLOW CHART

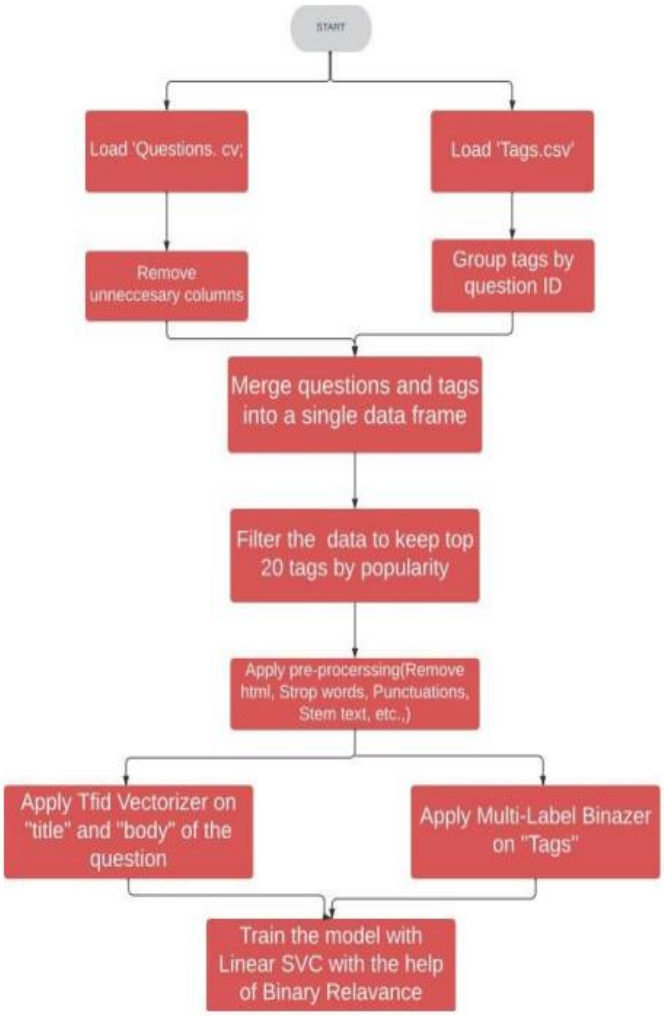

Fig -2: Flow Chart

6. METHODOLOGY

Load Questions.csv and Tags.csv. Group all tags given to the same question into a single string as the data is not grouped byID.

Drop unnecessary columns from questions and Merge questions and tags into a single data frame. Remove questions with scores lower than 5 to avoid outliers and split tags into lists. Get the most common 20 tags without the count and filter the dataset to only consider the data which has the top 20 tags.

Then Apply data pre-processing by:

- Remove HTML
- Remove stopwords
- Remove special characters
- Convert to lowercase
- Stemming

Html can be handled by the use of Regular Expressions and the rest can be done with the help of the nltk liberty.

Apply MultiLabelBinarizer on Tags and TfidfVectorizer on Questions. The most applicable machine learning algorithm for our problem is LinearSVC. The objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides or categorizes, your data.

From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.

For trying the model we decide to go with Linear SVC after a lot of testing as it constantly gave better results when compared to other classical classification methods like SGD Classifier and other Regression models.

We also used Binary Relevance from the scikit-multilearn lib. Binary Relevance transforms a multi-label classification problem with L labels into L single-label separate binary classification problems using the same base classifier provided in the constructor. The prediction output is the union of per-label label classifiers.

7. CONCLUSION

We first loaded the Text Pre-Processed Dataset using a Pandas Data Frame, after which we examined the String Tags.AST Module and encoded the tags using Multilabelbinarizer.

After that, we performed text vectorization on the question sort dataset using TfidfVectorizTor to evaluate the model's performance to actual data, we tested it agaiseveralr of classifiers, including SGDClassifier, LinearSVC, and Supplying Regression For Multi-Label Classification.

Among the categorization models we examined, we discovered that LinearSVC produced the best results.

8. REFERENCES

[1] 1 Y. Tao, Z. Cui, and Z. Wenjun, "A Multi-Label Text Classification Method Based on Labels Vector Fusion," 2018 International Conference on Promising Electronic Technologies (ICPET), 2018, pp. 80-85, doi: 10.1109/ICPET.2018.00021.

[2] G. Chen, D. Ye, Z. Xing, J. Chen and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383, doi: 10.1109/IJCNN.2017.7966144.

[3] S. Han, C. Lim, B. Cha, and J. Lee, "An Empirical Study for Class Imbalance in Extreme Multi-label Text Classification," 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), 2021, pp. 338-341, doi: 10.1109/BigComp51126.2021.00073.

[4] Al-Salemi, B., Ayob, M., & Noah, S. A. M. (2018). Feature ranking for enhancing boosting-based multi-label text categorization. Expert Systems with Applications, 113, 531–543. https://doi.org/10.1016/j.eswa.2018.07.024

[5] Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. Expert Systems with Applications, 57, 1–11. https://doi.org/10.1016/j.eswa.2016.03.041

[6] R. Jindal and Shweta, "A Novel Method for Efficient Multi-Label Text Categorization of research articles," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 333-336, doi: 10.1109/GUCON.2018.8674985.

[7] G. I. Winata and M. L. Khodra, "Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting," 2015 International Conference on Electrical Engineering and Informatics (ICEEI), 2015, pp. 500-505, doi: 10.1109/ICEEI.2015.7352552.

[8] G. Chen, D. Ye, Z. Xing, J. Chen and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2377-2383, doi: 10.1109/IJCNN.2017.7966144.

[9] Z. He, J. Wu and P. Lv, "Label correlation mixture model for multi-label text categorization," 2014 IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 83-88, doi: 10.1109/SLT.2014.7078554.

[10] A. Fiallos and K. Jimenes, "Using Reddit Data for Multi-Label Text Classification of Twitter Users Interests," 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG), 2019, pp. 324-327, doi: 10.1109/ICEDEG.2019.8734365.
https://ieeexplore.ieee.org/document/7078554

[11] https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5

[12] https://www.section.io/engineering-education/multi-label-classification-with-scikit-multilearn/

[13] https://www.coursera.org/learn/classification-vector-spaces-in-nlp?action=enroll

[14] https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[15] https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

[16] https://scikitlearn.org/stable/modules/generated/sklearn.svm.LinearSVC.html