# Stack Overflow Tag Recommendation System using One-VS-Rest Multi-label Classification

Trixia R. Belleza and Concepcion L. Khan

***Abstract*— Stack Overflow aims to organize questions by requiring users to categorize their posts with tags. However, this tagging process is done manually by users which can cause inaccurate classification of posts and disruption of the website′s overall user-experience. Hence, the objective of this research is to recommend tags based on the question content using a one-vs-rest approach multi-label classifier. To test the accuracy of the classifier, three metrics will be used: precision, recall, and F-score. A chrome extension will be created so that the application can be accessible to users.**

***Index Terms*— tags, classification, one-vs-rest, multi-label classifier, chrome extension**

## I. Introduction

### A. Background of the Study

A large population of users use online question-and-answer forums, like Stack Overflow and Quora, as a platform for discussing common topics of interest [1].

Unlike Quora, Stack Overflow is mainly for computer science majors and programmers [2]. The website is also strict with its standards for user interaction. In fact, according to the Stack Overflow website (2018) [3], questions are organized into well-defined categories based on its content in a form of tags. Tags are words that describe the question content in order to connect people that are knowledgeable in that field. With this, users can easily filter-search for questions.

In related studies, Guan et al [4] used a graph-based approach in dealing with tag-recommendation systems, in which, tags are being chosen according to their relevance to the content and preference of the user using a graph-based ranking algorithm. In addition, collaborative tagging data was used as an input to their algorithm.

However, in this study, the problem has been taken from another perspective and a different strategy. Instead of using collaborative tagging, the questions posted on the website will be used as input data.

Generally, using questions as the dataset, the machine will be able to classify them based on their topic category.

### B. Statement of the Problem

Tagging of questions are done manually by a user, which disrupts the website′s user-experience [5]. In fact, the researchers, Short, Wong, and Zeng (2014) [6], mentioned that some inexperienced users are having a hard time identifying which category their question belongs to. Moreover, manual tagging is prone to human error, which may result to improperly tagged questions. Hence, this can form a disarray of information that could overturn its purpose.

### C. Significance of the Study

The burden of manual tagging of questions on Stack Overflow calls for the need of an automated text-classification machine.

In order to classify texts, a multi-label classification technique, called the one-vs-rest approach, will be used in this study. This is especially used in classifying text-content to multiple categories according to Liu and Chen (2015) [7].

### D. Objectives of the Study

Generally, the study aims to create an application that recommends tags to Stack Overflow questions using a multi-label classifier. Specifically, it aims:

1) To gather ten thousand training and testing data from Stack Overflow using the Stack Exchange Application Programming Interface;
2) To create one-vs-rest multi-label classifiers using three different machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, and Logistic Regression;
3) To test the accuracy of the SVM, Naive Bayes, and Logistic Regression algorithms with a 60% threshold in classifying questions according to their topic categories by computing precision, recall, and F-score; and
4) To create a chrome extension that recommends tags to Stack Overflow questions using the algorithm with the highest accuracy among the three that were used, namely, SVM, Naive Bayes, and Logistic Regression.

## II. Review of Related Literature

Stack Overflow, an online forum, has been widely used for discovering new technologies and analyzing topics in computer science. In this website, posts are organized according to their topic category. In order to achieve this, users must manually classify categories that describe the question they are posting. However, manual tagging of questions degrades the website′s user experience which led researchers to improve the website′s features.

### A. Stack Overflow: A Question and Answer Forum

Stack Overflow is an online question-and-answer forum especially made for programmers. It is a website for discovering new trends and technologies through the discussions of its users [2]. According to Short, Wong, and Zheng (2014) [6], the website offers a platform for users to discuss about the question posted by another user. In posting a question, the user must provide the title and the body of the question including

tags as a means for describing what the topic is all about. These tags are essential in providing filtered searches and organized posts.

However, manual tagging of posts can be quite tedious, which is inefficient in terms of user-experience [5].

### B. Machine Learning

Machine Learning is the science of getting computers to detect patterns from a trend, and use it to predict a possible outcome using computational methods.

One type of machine learning is supervised learning. Supervised learning is a method in which target labels are known. In other words, there is a desired output based on a given input [8]. This type of machine learning is commonly used in automated text classification given a set of labels [9]. To demonstrate, in order for the machine to classify a sample, it should be given a set of labeled data to work on first. Then, the machine will be given new unlabeled texts for the testing phase, in which the machine will automatically classify them based on the model it has learned [10].

According to Alpaydin (2010) [11], the results of machine learning algorithms can become reliable for analysts, data scientists, and businessmen to create decisions and uncover hidden insights. With this, machine learning has been widely used because of its many applications. One application would be the automatic detection of acromegaly, a hormonal disorder, from facial photographs [12]. Another would be an automatic classification of app reviews which is essential for app developers to improve their app [13].

Machine learning has been the cause of changes in people′s day-to-day activities and it still continues to prove its potential up to this day.

### C. Natural Language Processing Toolkit

The natural language processing toolkit (NLTK) is a most commonly used platform by researchers, teachers, and programmers. It is a suite of libraries and programs for classification, tokenization, stemming, parsing, etc., especially made for processing the human language [14].

Over the years, NLTK has continued to improve with more complex data structures and processing tasks [15].

### D. Multi-label Classification

A multi-label classification assigns an instance to multiple target labels. In the study of Wang, Chen, and Suis (2018) [16], multi-label classification was applied in order to assign texts to multiple categories based on conceptual content. Thus, the training set contains the texts associated with a set of categories. The goal is to predict the categories of unseen text data based on the model created from the training set.

Another study was conducted by Almeida et al [17] using multi-label classification as a means of analyzing the sentiments found in text. In fact, this methodology was able to categorize multiple emotions in the same text.

It is often confused with the definition of multi-class classification. To differentiate, a multi-class classification is mutually exclusive, wherein, each instance is only mapped to one label. In a multi-label classification, two or more labels can be assigned to each instance.

According to Nooney (2018) [18], "Most traditional learning algorithms are developed for single-label classification problems." Because of this, solutions to multi-label classification problems are transformed into multiple single-label problems, so that the existing single-label algorithms can be used.

According to Boutell et al in 2004 (as cited by Tsoumakas & Katakis, n.d.) [19], two of the techniques in multi-label classification decomposes the problem into multiple single-label classification problems. One is the one-vs-rest approach where samples that are associated with the label are positive and the rest, which represent other labels, are negative. In other words, this approach distinguishes one class from the rest of the other classes [20]. On the other hand, the one-vs-one approach trains each sample with only two classes in which, it learns to distinguish each pair of classes [21].

### E. Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning algorithm that is based on the concept of decision planes. Decision planes are the boundaries that separates classes of objects. According to Gholami and Fakhari (2017) [22], the goal of the SVM is to find a decision function which can classify unseen data with a good accuracy.

Decision planes classify data linearly. However, most classification problems are not always linearly separable and more often, complex structures are needed in order to separate one class from the other [23].

To solve this problem, a kernel machine was introduced. The idea is that, the data that is not linearly separable in a two-dimensional space may be linearly separable in a higher dimensional space by computing the inner products of the data points [24].

Support Vector Machines have been widely used in text-classification systems because of its outstanding classification accuracy compared to other classification algorithms [25].

### F. Naive Bayes

Naive Bayes algorithm is a classification technique that uses the probability theory, Bayes theorem, which relates conditional probabilities. Conditional probabilities are basically based on prior knowledge of conditions that are related to the event. Bayes theorem states that if A and B represent two events, the probability of A given B, P(A|B) denotes the conditional probability of A occurring, given that B occurs. On the other hand, the probability of B given A, P(B|A), denotes the probability of B occurring, given that A occurs. In addition to Bayes theorem, it states that these two conditional probabilities are related to each other [26].

To illustrate how Naive Bayes algorithm works, Techopedia (2018) [27] uses sorting of fruits as an example. A classifier that categorizes fruits into apples and oranges would use shape and color as input features but would not assume all these

things at once since apples and oranges have a similar round shape.

According to Wei et al (2011) [28], Naive Bayes algorithm has an advantage for multi-label classification problem. In their study, they mentioned that McCallum [29] used this approach to classify multi-labeled documents and Zhang et al [30] used the same algorithm in dealing with multi-label problems that incorporated feature selection mechanisms that gave the machine learning technique a higher accuracy.

### G. Logistic Regression

Logistic regression is a statistical analysis technique which is used when the dependent variable is dichotomous, where 1 represents the value of TRUE and 0 denotes FALSE [31].

In order to define logistic regression more clearly, Saedsayad [32], a website that explains numerous machine learning algorithms, compared linear regression from logistic regression based on what the linear regression cannot do that the logistic regression can. It was stated that a linear regression has a large range of values that goes outside the acceptable range 0 to 1. Since binary or dichotomous datasets are limited to two possible values only, the residuals will not be normally distributed on the predicted line.

Unlike linear regression, logistic regression forms a logistic curve that limits the values from 0 to 1. The curve is formed using the natural logarithm of the odds of the target labels.

In related studies, Fujino and Isozaki (2008) [33] used logistic regression for a patent mining task. The logistic regression models were trained by providing existing patent documents as dataset. They designed the feature vectors of the documents with weighting and component selection methods to avoid overfitting.

### H. Tag Recommendation System

Tagging is a form of categorizing texts using keywords to describe its content. This mechanism has become popular on the Web because it promises to provide organization and easy retrieval of online content [34].

According to Guan et al [4], a tag recommendation system is necessary to help users decide tags more accurately by suggesting suitable and relevant tags to them. It can also enrich the index of resources.

Moreover, the focus of their study lies on the documents that are annotated by users. With tags, users can bookmark the Web URLs they visited.

The researchers used a collaborative filtering strategy that explores collaborative knowledge and not on the content of the documents.

However, considering the diverse interests of users, the performance of the tag recommendation system provided a low-level accuracy in finding tags directly related to the current document.

In another study, the tag recommendation system can be modeled by focusing on the most frequent tags assigned to a text-content and the tagging history of the user, in which, tag recommendation algorithms can be classified into user-centered and resource-centered ones [35].

## III. Materials and Methods

### A. General Flowchart

The flow of the study is shown below:

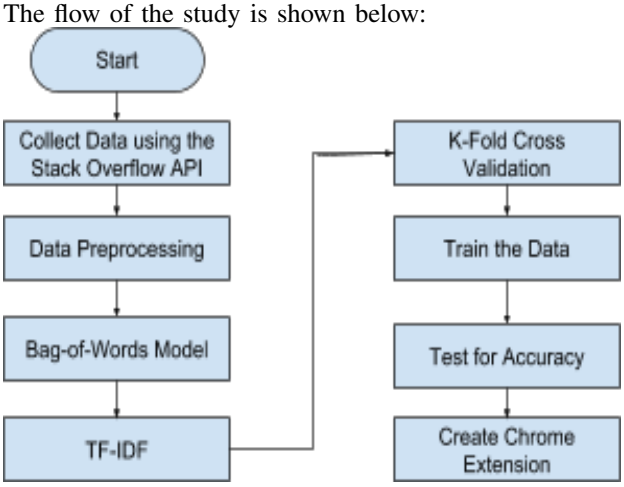


Fig. 1. A flowchart showing the summary of the process.

### B. Frameworks and Libraries

Python3, together with the libraries NumPy, and Scikit Learn will be used in this study. The JSON library will also be used for extracting the scraped data from Stack Overflow that is stored in a JSON-format file. The NumPy library will be used for mathematical computations involving multi-dimensional arrays. The Scikit Learn Library will be used for importing a Support Vector Machine (SVM) model.

NPM is a package manager for JavaScript and Rapydscript which converts Python syntax into JavaScript. NPM and Rapydscript will be used for creating a chrome extension.

### C. Data Collection

Stack Overflow provides an Application Programming Interface (API) that outputs a JSON-format list of questions with tags.

JavaScript will be used in retrieving JSON-format data to push it onto the database management system, MySQL. [3].

### D. Data Preprocessing

The language of the text-data that will be used in this study is English.

In text data, there are words called stop words. These words are ignored because they are irrelevant and unnecessary for processing text analysis. Punctuation marks are also disregarded [36]. Thus, in data preprocessing, stop words (e. g. , the, of, a, but) and punctuation marks are put aside in a stop list. A Natural Language Toolkit (NLTK) library will be used in extracting stop words and stemming the text data. All text data will also be converted in a lower-case form to make it case-insensitive to the program.

The actual data labels that will be used in this study are the five commonly used tags in Stack Overflow. These are JavaScript, Java, C, R, and Python [3]. Control flow statements if-then, if-then-else, for-loop, and while-loop will also be used. Thus, the total number of actual data labels for this study is nine.

### E. Bag-of-Words Vocabulary Model

The bag-of-words model is a representation of data in natural language processing. It contains the set of words, excluding stop words, and the number of times the word has occurred [37].

### F. Term Frequency-Inverse Document Frequency (TF-IDF)

According to Erra et al (2013) [38], the Term Frequency-Inverse Document Frequency (TF-IDF) provides the weights of a word depending on its significance to a text. TF-IDF is basically the product of Term Frequency (TF) and Inverse Document Frequency (IDF). TF is computed by

$$tf(t,d) = \frac{f_{t,d}}{max\{f'_{t,d} : t' \in d\}}$$

**TF** is the frequency f of the term t in the document d.

On the other hand, **IDF** is computed by

$$idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|}$$

Where N is total number of documents and $|d \in D : t \in d|$ is the number of documents d in which the term t appears. If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |d \in D : t \in d|$.

Finally, TF-IDF can be computed by multiplying TF and IDF.

$$tf(t,d) * idf(t,d,D) = tf(t,d) \times idf(t,D)$$

### G. K-Fold Cross Validation

To avoid overfitting, the data set will be divided into k groups (or folds) of equal size. In this study, the value of k will be 10. The first fold is the validation set and the method is fit on the remaining k-1 folds.

The training data set contains the questions from Stack Overflow and the tags of the questions. The test data set will be used to test for accuracy by giving the machine a set of data without labels.

### H. Training the Model using a Multi-label Classifier

In this study, the researcher will use a one-vs-rest multi-label classifier approach. To classify the N-class data in a one-vs-rest approach, N binary classifiers must be created, $\{a_1, a_2, \ldots a_N\}$. For the i-th classifier, let all the points in the class $a_i$ be one, and let all the points not in the class $a_i$ be zero [5].

Basically, multi-label classification outputs a model that maps x features to y binary vectors by assigning a binary value, 0 or 1, for each label in y. Then, a machine learning algorithm will be applied to train each model that will be produced by the one-vs-rest classifier.

Three machine learning algorithms, SVM, Naive Bayes, and Logistic Regression, will be applied to the one-vs-rest classifier and will be evaluated as to which among them will provide the highest accuracy.

### I. Performance Evaluation

According to Lingras and Butz (2007) [39], to measure the performance of the classifier, the researcher will be using three metrics: precision, recall, and F-score.

Precision is computed by finding the ratio of correctly classified samples to the total classified positive samples.

$$Precision = \frac{TP}{TP + FP}$$

Recall is computed by finding the ratio of correctly classified samples to the total samples in the actual class.

$$Recall = \frac{TP}{TP + FN}$$

Finally, to compute for F-score that will determine the accuracy of a classifier is as follows:

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

### J. Stack Overflow Chrome Extension

In order to apply this study to the Stack Overflow website, the researcher will be using a chrome extension.

In this study, Python will be used for the main program. However, in order to make a chrome extension, the code should be written in JavaScript. Hence, the researcher will be using a RapydScript, a pre-compiler of JavaScript, that converts Python syntax to JavaScript [40].

### K. Gantt Chart

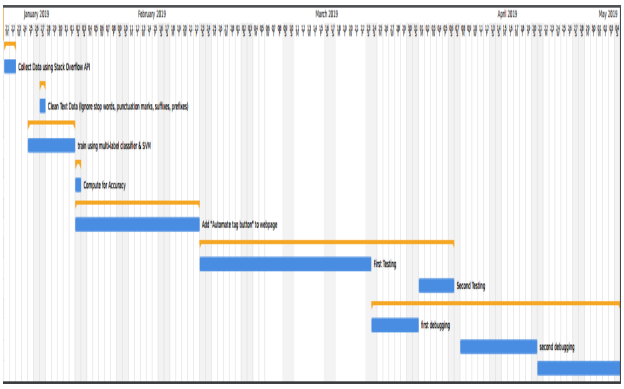


Fig. 2. A gantt chart showing the planned period for each task.

## REFERENCES

[1] V. S. Rekha and S. Venkatapathy, "Understanding the usage of online forums as learning platforms," *Procedia Computer Science*, vol. 46, pp. 499–506, 2015.

[2] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[3] S. E. Inc., *Stack Exchange API v2.2*, (accessed October 1, 2018), https://api.stackexchange.com/docs.

[4] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang, "Personalized tag recommendation using graph-based ranking on multi-type interrelated objects," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 540–547.

[5] M. Eric, A. Klimovic, and V. Zhong, "{meric, anakli, vzhong}@ stanford. edu december 8, 2014," 2014.

[6] L. Short, C. Wong, and D. Zeng, "Tag recommendations in stackoverflow," *CS224W Project*, 2014.
[7] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1083–1093, 2015.
[8] T. W. Edgar and D. O. Manz, *Research Methods for Cyber Security*. Syngress, 2017.
[9] C. Keming and Z. Jianguo, "Research on the text classification based on natural language processing and machine learning," *JOURNAL OF THE BALKAN TRIBOLOGICAL ASSOCIATION*, vol. 22, no. 3, pp. 2484–2494, 2016.
[10] S.-L. Developers, *An introduction to machine learning with scikit-learn*, (accessed October 1, 2018), http://scikit-learn.org/stable/tutorial/basic/tutorial.html.
[11] E. Alpaydın, "Introduction to machine learning 2nd edition ed," 2010.
[12] X. Kong, S. Gong, L. Su, N. Howard, and Y. Kong, "Automatic detection of acromegaly from facial photographs using machine learning methods," *EBioMedicine*, vol. 27, pp. 94–102, 2018.
[13] W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik, "On the automatic classification of app reviews," *Requirements Engineering*, vol. 21, no. 3, pp. 311–331, 2016.
[14] N. Project, *Natural Language Toolkit*, May 06, 2018 (accessed November 8, 2018), https://www.nltk.org/#natural-language-toolkit.
[15] S. Bird, "Nltk-lite: Efficient scripting for natural language processing," in *Proceedings of the 4th International Conference on Natural Language Processing (ICON)*, 2005, pp. 11–18.
[16] R. Wang, G. Chen, and X. Sui, "Multi label text classification method based on co-occurrence latent semantic vector space," *Procedia computer science*, vol. 131, pp. 756–764, 2018.
[17] A. M. Almeida, R. Cerri, E. C. Paraiso, R. G. Mantovani, and S. B. Junior, "Applying multi-label techniques in emotion identification of short texts," *Neurocomputing*, 2018.
[18] K. Nooney, *Deep dive into multi-label classification! towards data science*, 2018, https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff.
[19] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.
[20] B. Krawczyk, M. Galar, M. Woźniak, H. Bustince, and F. Herrera, "Dynamic ensemble selection for multi-class classification with one-class classifiers," *Pattern Recognition*, vol. 83, pp. 34–51, 2018.
[21] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.
[22] R. Gholami and N. Fakhari, "Support vector machine: Principles, parameters, and applications," in *Handbook of Neural Computation*. Elsevier, 2017, pp. 515–535.
[23] T. S. Inc, *Support Vector Machines*, 2018 (accessed November 2, 2018), http://www.statsoft.com/textbook/support-vector-machines.
[24] H. Kandan, *Understanding the kernel trick*, August 31, 2017 (accessed November 2, 2018), https://towardsdatascience.com/understanding-the-kernel-trick-e0bc6112ef78.
[25] N. Shafiabady, L. H. Lee, R. Rajkumar, V. Kallimani, N. A. Akram, and D. Isa, "Using unsupervised clustering approach to train the support vector machine for text classification," *Neurocomputing*, vol. 211, pp. 4–10, 2016.
[26] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.
[27] T. Inc., *Naive Bayes*, 2018 (accessed December 1, 2018), https://www.techopedia.com/definition/32335/naive-bayes.
[28] Z. Wei, H. Zhang, Z. Zhang, W. Li, and D. Miao, "A naive bayesian multi-label classification algorithm with application to visualize text search results," *International Journal of Advanced Intelligence*, vol. 3, no. 2, pp. 173–188, 2011.
[29] A. McCallum, "Multi-label text classification with a mixture model trained by em," in *AAAI workshop on Text Learning*, 1999, pp. 1–7.
[30] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.
[31] S. Solution, *What is Logistic Regression?*, 2018 (accessed December 1, 2018), https://www.statisticssolutions.com/what-is-logistic-regression/.
[32] Saedsayad, *Logistic Regression*, 2018 (accessed December 1, 2018), https://www.saedsayad.com/logistic_regression.htm.
[33] A. Fujino and H. Isozaki, "Multi-label classification using logistic regression models for ntcir-7 patent mining task." in *NTCIR*, 2008.
[34] Y. Wu, S. Xi, Y. Yao, F. Xu, H. Tong, and J. Lu, "Guiding supervised topic modeling for content based tag recommendation," *Neurocomputing*, vol. 314, pp. 479–489, 2018.
[35] R. Krestel and P. Fankhauser, "Personalized topic-based tag recommendation," *Neurocomputing*, vol. 76, no. 1, pp. 61–70, 2012.
[36] D. Munková, M. Munk, and M. Vozár, "Data pre-processing evaluation for text mining: transaction/sequence model," *Procedia Computer Science*, vol. 18, pp. 1198–1207, 2013.
[37] J. Brownlee, "A gentle introduction to the bag-of-words model," *Machine Learning Mastery*, vol. 21, 2017.
[38] U. Erra, S. Senatore, F. Minnella, and G. Caggianese, "Approximate tf–idf based on topic extraction from massive message stream using the gpu," *Information Sciences*, vol. 292, pp. 143–161, 2015.
[39] P. Lingras and C. J. Butz, "Precision and recall in rough support vector machines," in *Granular Computing, 2007. GRC 2007. IEEE International Conference on*. IEEE, 2007, pp. 654–654.
[40] P. S. Foundation, *Rapydscript*, (accessed October 1, 2018), https://pypi.org/project/rapydscript/.

**Trixia R. Belleza** is a BS Computer Science student. She enjoys seeking knowledge about various algorithms that are in-line with her chosen degree.