Predictive Modeling for Multi-Labels Tagging in Algorithmic Challenges

~Aayush Jaiswal (aajais@iu.edu)
~Meet Palod (mpalod@iu.edu)

Abstract:

This project seeks to advance user interaction with CodeForce, a hub for algorithmic contests, by devising a predictive framework capable of assigning multiple descriptive labels to each posted problem based on its textual content. The study sets out to meticulously parse problem descriptions, fine-tune text processing methods, and critically assess several computational models, including logistic regression and random forest classifiers, as well as a BiLSTM network with an embedding feature. The end goal is to refine the tagging mechanism to significantly uplift the accuracy, exhaustiveness, and reduction of misclassification in the system to aid users in pinpointing relevant problems that match their proficiency and search requirements.

Introduction:

The domain of competitive programming has long recognized the importance of effectively categorizing problem sets to enhance educational outcomes and user satisfaction. On platforms such as CodeForce, where the corpus of problems is extensive, the deployment of a predictive model that can accurately tag problems based on their textual descriptions is of paramount importance. This research delineates the development of a multi-label classification model that employs natural language processing and machine learning techniques to automate the assignment of tags to programming challenges. Drawing upon existing research that underscores the complexity of multi-label classification tasks (Zhang & Zhou, 2014; Tsoumakas et al., 2009), this study aims to bridge the gap between problem statements and their appropriate categorical tags, thereby refining the navigation and searchability of problems on the CodeForce platform.

Methods:

The methodology employed in this research involves gathering a dataset from Kaggle's public repository and using various critical data pre-processing techniques to prepare text-based classification tasks. These techniques, including text normalization, tokenization, and vectorization, are crucial for analyzing textual data. Our approach to model selection is to conduct a comparative assessment of machine learning algorithms, such as decision trees, random forests, and neural networks. To ensure accuracy in measuring the performance of our predictive models, we use evaluation metrics specially tailored to the unique demands of multi-label classification, with a particular focus on precision, recall, and the F1 score.

The study aims to achieve the below objectives:

1. To conduct an Exploratory Data Analysis (EDA) on the problem tags. This will involve examining the distribution of tags per question, the lexical complexity of tags, as well as the frequency of unigrams and bigrams within the problem statements.

2. To implement a rigorous Data Pre-processing routine on the problem statements. This involves converting text to lowercase, removing Unicode characters, purging HTML tags, eliminating stop words, and implementing lemmatization and stemming techniques.
3. After completing the EDA and pre-processing, we will proceed to implement and compare various machine-learning algorithms for multi-label classification. The models under consideration include Multi-label Logistic Regression, a Multi-label Random Forest Classifier, and a Deep Learning algorithm (BiLSTM with an Embedding layer). Our focus will be on hyper-parameter optimization to enhance the precision, recall, f1-score, and hamming loss.

Author Contribution Statement:

Project Phase	**Team Members**
Data Preparation and Pre-processing	Meet Palod
Exploratory Data Analysis	Aayush Jaiswal
Model Implementation and Training	Meet Palod
Performance Evaluation of Model, Model Selection, and Results	Aayush Jaiswal
Final Report and Presentation	Meet Palod, Aayush Jaiswal

References:

1. Zhang, M.-L., & Zhou, Z.-H. (2014). A synthesis of the art in multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819-1837.
2. Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Exploration in the domain of mining multi-label datasets. Data Mining and Knowledge Discovery Handbook, 667-685.
3. Bischl, B., et al. (2012). mlr: An R package for machine learning. Journal of Machine Learning Research, 17(170), 1-5.
4. Gibaja, E., & Ventura, S. (2015). A guide to multi-label learning. ACM Computing Surveys (CSUR), 47(3), 1-38.
5. Madjarov, G., et al. (2012). Comprehensive empirical comparison of multi-label learning methods. Pattern Recognition, 45(9), 3084-3104.
6. Read, J., et al. (2011). Classifier Chains: A multi-label classifier methodology. Machine Learning, 85(3), 333-359.
7. Sorower, M. S. (2010). A survey report on multi-label learning algorithms. Oregon State University, Corvallis.
8. Li, J., et al. (2016). An overview of feature selection techniques in data science. ACM Computing Surveys (CSUR), 50(6), 1-45.