An Efficient Approach For Text Classification Using Ensemble Models in Machine Learning

AXEL GORIS RAFAT KHAN
goris | rafatk @kth.se

September 7, 2022

1 Allocation of responsibilities

Axel Goris is responsible for conducting research, finding datasets and analysing the results of the models. Rafat Khan is responsible for building, training and interpreting the models and the source of difference.

2 Organization

The project will be organized as a two-person project. We have found some relevent prior works and there are plenty of test data available in Kaggle [1], from where we can select suitable dataset and divide it into training data, validation data and test data. In parallel, we will start developing the experiments and implement the models.

3 Background

Machine learning algorithms have been used to classify texts for a long time [2]. Support Vector Machines [3], Naive Bayes Classifier[4], XGBOOST[5] and KNN[6] algorithms can classify texts. Recently, Sreemathy JayaprakashP et al. published a paper proposing text classification based on features selection and pre-processing thereby reducing the dimensionality of the Feature vector and increasing the classification accuracy [7]. It seems like an interesting approach that works in a more efficient way. From this paper, we got an idea to use Multiple-Model Techniques [8] to further enhance the performance. This concept of ensembling model is being used for different purposes. The ensemble forecasting approach combines different data sources, models of different types, assumptions, and/or pattern recognition methods. A more accurate ensemble model can be achieved by combining information from multiple sources, analyzing it with different techniques, and considering different sources of uncertainty in the real world [9]. In this project, we will implement the hypothesis and ensemble KNN, Naive Bayes and XGBOOST to determine if this approach enhance the accuracy and performance further. We will also conduct a comparison study of the accuracy and performance.

4 Problem statement

Examine if ensembling KNN, Naive Bayes and XGBOOST can be a more efficient way for text classification in terms of efficiency, accuracy and performance.

5 Problem

The project will investigate the accuracy and performance of KNN, Naive Bayes and XGBOOST on a specific dataset for text classification. Then we will ensemble the models by aggregating the output from each model with two objectives: reducing the model error and maintaining its generalization by using some pre-defined metrics.

6 Hypothesis

Our hypothesis is that ensembling machine learning models will improve performance and accuracy significantly for classifying texts.

7 Purpose

The purpose is to see, whether or not, ensembling machine learning models improves performance and accuracy and if if it does, to what level?

8 Goal(s)

Implementing models for text classification. Ensemble the models by aggregating the outputs. Then conduct a comparison survey of the accuracy and overall performance. If all the experiments are designed correctly, this project will provide conclusive decision on the performance of ensemble models for text classification.

9 Tasks

We will implement models for text classification using three different algorithms. We will measure the accuracy and overall performance for each algorithm using experimental data. Then we will ensemble the models by aggregating the output from each model. We will define a metric for aggregating the output. Then we will do a comparison of the accuracy and overall performance.

10 Method

The project will use a quantitative analysis to produce an accurate evaluation of the performance difference between different models.

11 Milestone chart (time schedule)

The project will start on 6 September and end at 10 October. There will be the following milestones and deliverable:

09 September - Select experiment data

10 September - Start building the models

15 September - Start running the experiments

20 September - Define aggregation metrics for ensembling models

25 September - Evaluation of model for additional data points to be check by new experimental runs to check that the model accurately produces the expected outcomes.

Before 10 October - Submit final report (the report will have been written in parallel with each of the above steps)

References

[1] *Text Classification*. en. URL: `https://kaggle.com/code/matleonard/text-classification` (visited on 09/07/2022).

[2] Fabrizio Sebastiani. "Machine Learning in Automated Text Categorization". In: *ACM Computing Surveys* 34 (Apr. 2001), pp. 1–47. DOI: `10.1145/505282.505283`.

[3] Chen Shan. "Research of Support Vector Machine in Text Classification". In: *Advances in Intelligent and Soft Computing* 119 (Jan. 2011), pp. 567–573. ISSN: 978-3-642-25537-3. DOI: `10.1007/978-3-642-25538-0_79`.

[4] Duan Li-guo, Di Peng, and Li Ai-ping. "A New Naive Bayes Text Classification Algorithm". In: *TELKOMNIKA Indonesian Journal of Electrical Engineering* 12 (Sept. 2014). DOI: `10.11591/telkomnika.v12i2.4180`.

[5] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. Pages: 794. Aug. 2016. DOI: `10.1145/2939672.2939785`.

[6] Gongde Guo et al. "Using kNN model for automatic text categorization". In: *Soft Computing* 10 (Mar. 2006). DOI: `10.1007/s00500-005-0503-y`.

[7] Sreemathy Jayaprakash and P Balamurugan. "An Efficient Text Classification Using KNN And Naive Bayesian". In: *International Journal on Computer Science and Engineering* 4 (Mar. 2022), pp. 392–396.

[8] Jason Brownlee. *A Gentle Introduction to Multiple-Model Machine Learning*. `https://machinelearninc` `com/multiple-model-machine-learning/`. 2022.

[9] Hao Wu and David Levinson. "Ensemble Models of For-Hire Vehicle Trips". In: *Frontiers in Future Transportation* 3 (Apr. 2022), p. 876880. DOI: `10.3389/ffutr.2022.876880`.