**Performance Evaluation on Vector-Model Combinations**

***Project Proposal B***

GitHub link: <https://github.com/trela47/PEoVMC_nlp-project>

**Team Members:**

Yamuna Bollepalli

Saisri Teja Pepeti

Saikiran Yedulla

Blessy Kuriakose

**Goals and Objectives**

**Motivation:**

Exploring the effectiveness of differing vectorization-and-model combos in the text based prediction task of spam detection. By this project we will be able to accomplish which vectorization will work with best model accuracy and least time complexity by comparing combinations.

**Significance:**

There are multiple vectorization techniques available that but there is no proper model available to distinguish which technique and model suits the best for a particular dataset. In this project we will be able to provide the accurate model for a particular dataset along with the time complexity. So, this helps the organizations and individuals time as they are already aware of the model best suit for a particular dataset.

Moreover, we will try to improve the accuracy of Support Vector Machine classification and Naive Bayes along with the least time complexity. We will also work with Random Forest classification and Decision Tree classification and assuming that we will be able to bring slight variation in the accuracy compared to the existing. We target to achieve at least 85 % accuracy from one of the models.

**Objectives:**

* Clean the raw text.
* Get language-based vectors from the cleaned text.
* Make the models for training and testing and track the metrics (confusion matrix, time complexity, etc.)
* Try each vector for each model.
* Evaluate the metrics to conclude the best combination.
* Calculate the time complexity.

**Features:**

We will be using the spam detection dataset from kaggle, of which the direct features are available as direct text. Alterations of cleaning and vectorization (frequency-based understanding of the words present) will be completed prior to use with the models.Then evaluate the model performance based on time complexity and model metrics.

**Project lifecycle:**

****

**Figure 1: Project Lifecycle.**

**References**

<https://medium.com/analytics-vidhya/magic-of-tf-idf-202649d39c2f>

<https://neptune.ai/blog/vectorization-techniques-in-nlp-guide>

<https://www.sketchbubble.com/en/presentation-project-life-cycle.html>

<https://machine-learning.paperspace.com/wiki/machine-learning-models-explained>