A Robust Approach for Effective Spam Detection Using Supervised Learning Techniques

ABSTRACT

In this age of popular instant messaging applications, Short Message Service or SMS has lost relevance and has turned into the forte of service providers, business houses, and different organizations that use this service to target common users for marketing and spamming. A recent trend in spam messaging is the use of content in regional language typed in English, which makes the detection and filtering of such messages more challenging. In this work, an extended version of a standard SMS corpus containing spam and non-spam messages that is extended by the inclusion of labeled

text messages in regional languages like Hindi or Bengali typed in English has been used, as gathered from local mobile users. Monte Carlo approach is utilized for learning and classification in a supervised approach, using a set of features and machine learning algorithms commonly used by researchers. The results illustrate how different algorithms perform in addressing the given challenge effectively..

**EXISTING SYSTEM**

Back in 2015, Agarwal et al. [5] utilized the comprehensive data corpus consolidated

by [6] and extended it by adding a set of spam and ham SMS collected from Indian mobile users. They demonstrated how different learning algorithms like Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) performed on the Term Frequency–Inverse Document Frequency (TF-IDF)–based features extracted from the corpora. Starting at around this time, a plethora of research works have used the same corpus and similar set of features and learning algorithms for designing spam detection systems. In the following set of similar works, it is observed that a set of learning and classification algorithms are used for a performance comparison study. Also, there is a paradigm shift toward neural network-based learning algorithms in more recent times.

In such a work in 2017, Suleiman et al. [7] demonstrated a comparative study of the performance of MNB, Random Forest, and Deep Learning algorithm–based models by using the H2O framework and a self-determined set of novel features on the same SMS corpus. Using word embedding features, Jain et al. [8] showed in 2018 how Convolutional Neural Network (CNN) can be utilized to achieve a better performance than a number of other baseline machine learning models in determining the spam messages from the corpus of [6].

In the same year, Popovac et al. [9] illustrated how CNN algorithm performs on the

same SMS corpus using TD-IDF features.

In 2019, Gupta et al. [10] proposed a voting ensemble technique on different learning algorithms, namely, MNB, Gaussian Naïve Bayes (GNB), Bernoulli Naïve Bayes (BNB), and Decision Tree (DT) for spam identification using the same corpus.

The trend of classifier performance comparison continues till recent times in 2020,

where the work by Hlouli et al. [11], illustrated how Multi-Layer Perceptron (MLP), SVM, k-Nearest Neighbors (kNN), and Random Forest algorithms perform on the same SMS corpus for detecting spam and ham using Bag of Words and TF-IDF–based features. In a similar contemporary work, GuangJun et al. [12] highlighted the performance of kNN, DT, and Logistic Regression (LR) models on SMS spam corpus, though the feature extraction techniques were not discussed.

A recent but different type of work by Roy et al. [13] shows how the same SMS corpus by Hidalgo et al. [6] is classified using Long Short Term Memory (LSTM) and CNN-based machine learning models with a high accuracy. The authors also noted that dependence on manual feature selection and extraction results often influences the efficacy of the spam detection system and consequently utilized the inherent features determined by the LSTM and CNN algorithms.

**Disadvantages**

* The system is not implemented Inverse Document Frequency (IDF).
* SMS data is to be finally used by the mathematical model–based supervised learning algorithms. These algorithms fail to deal with textual content in the data and are more comfortable with numeric values.

Proposed System

It is observed that in spite of the comparative study of classification performance undertaken by the aforementioned state-of-the-art works, none of them have attempted to determine and establish the robustness of the classification techniques in spam identification. Also, the abundance of spam messages in regional language is largely ignored in such works.

1. The system introduces the novel context of identifying spam and ham SMS in regional languages that are typed in English, along with the general English corpus of

spam and ham by extending it.

2. The system employes a Monte Carlo approach and ML Classifiers to repeatedly perform classification using different machine learning algorithms on different combinations of spam and ham text from the extended corpus (with k-fold cross-validation for a large value of k = 100) in order to determine the efficiency of baseline learning algorithms in comparison to the CNN-based model.

**Advantages**

* The proposed system is more effective due to presence of many ml classifiers.
* The proposed system implemented with an accurate prediction for the corresponding dataset.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).