VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Review Report - 1

Programme: B-Tech

Course: Natural Language Processing

Slot: E2+TE2

Faculty: Dr. M Premalatha

Component: J-Component

Title:

Email/SMS Spam Classification using NLP

Team Member(s):

1. Aayushi (20BCE1791)

2. Garvit Jakar (20BCE1838)

Abstract

Natural Language processing or NLP is a subset of Artificial Intelligence (AI), where it is basically responsible for the understanding of human language by a machine or a robot.
One of the important subtopics in NLP is *Natural Language Understanding (NLU)* and the reason is that it is used to understand the structure and meaning of human language, and then with the help of computer science transform this linguistic knowledge into algorithms of Rules-based machine learning that can solve specific problems and perform desired tasks.

Keywords
Natural Language Processing, Artificial Intelligence, Human Language, Machine Learning

Introduction

Most of us should be familiar with spam emails. Cisco defines it as unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers. Therefore, spam email filtering is an essential feature for email services such as Outlook and Gmail. Services providers are extensively using Machine learning techniques to filter and classify them successfully.
The best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more. Spam detection is one of a handful of NLP problems that experts consider 'mostly solved' .

1. **Data Set Description**:
   The SMS Spam Collection v.1 is a set of SMS messages that have been collected and labelled as either spam or not spam. This dataset contains 5574 English, real, and non-encoded messages. The SMS messages are thought-provoking and eye-catching. The dataset is useful for mobile phone spam research.
   The dataset we are using is a public text data, it contains two columns including text (email) and spam (label). The spam column has two values (0 and 1), the text is labelled 1 if it is a spam or 0 otherwise.

   Link for the dataset used:- https://huggingface.co/datasets/sms_spam

- This dataset could be used to train a machine learning model to classify SMS messages as spam or not spam.

- This dataset could be used to develop a tool that can automatically identify and block spam messages.
- This dataset could be used to study the characteristics of spam messages and develop strategies for identifying and avoiding them

2. Methodology and Algorithm used:

Different levels of NLP are:-

1. Morphological
2. Lexical
3. Syntactic
4. Semantic
5. Discourse
6. Pragmatic

Spam detection in Email/ SMS falls under the semantic level of nlp. The semantic level of linguistic processing deals with the determination of what a sentence really means by relating syntactic features and disambiguating words with multiple definitions to the given context. This level entails the appropriate interpretation of the meaning of sentences, rather than the analysis at the level of individual words or phrases.

The machine learning algorithms to be used further in the project to detect spam email/SMS are **Naive Bayes algorithm and J48 (Decision tree) algorithm**. In this process of detecting the spam mails, the dataset is divided into different sets and given as input to each of the algorithm.

Processes Involved

1. We need to import the data from Kaggle or some similar publicly available and accessible datasets using :-

   Import NumPy as np // library in Python
   Import pandas as pd

2. Data cleaning : *// to remove all the Nan values present in the dataset*

   Remove stop words to reduce the vocabulary.

3. EDA *// Exploratory Data Analysis*
4. Text Pre-processing (Tokenise):

   Given a plain text, we first normalize it and convert it to lowercase and remove punctuation and finally split it up into words, these words are called tokenizers.

5. Normalisation:

   In order to further simplify our text data, we can lemmatize or stem in this step. **Lemmatization and Stemming** usually refer to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma

6. Model Building
7. Model Evaluation
8. Improvements depending on the evaluation