Addis Ababa University
አዲስ፡አበባ፡ዩኒቨርሲቲ
SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

Addis Ababa University

College of Natural and Computational Science

School of Information Science

Documentation on the SMS Spam Detection for Amharic Texts using python

Submitted to: - Wondwossen Mulugeta (PhD)

Done By:

1. Amina Abdulkedir….GSR/6604/12
2. Henon Mengistu……GSR/0995/12

Sep 12, 2020

Introduction

Spam detection is an example of document classification task, which involves classifying an email or SMS as spam or non-spam. Spam box in our Gmail account is the best example of this. In this project, we build a text classifier in python3 using Amharic SMS texts with 1002 text messages. Our model tries to detect Amharic spam SMS messages based on machine learning models.

We first collected the messages, and then labeled them on the basis of their content which is either ***ham*** (non-spam) or ***spam***. To make a distinction between the spam and non-spam messages, we have taken Spam messages as the unsolicited texts that are received from organizations a user has not confirmed a subscription to and ham messages as message received from individuals

To achieve this goal we go through some steps :

Importing libraries and Load the data

The spam data was collected from our archive of spam messages found on our phones.The ham messages were collected consensually from our personal networks We collect the data using primary source from our friends.

We import necessary libraries such as:

- pandas , numpy , nltk, string…..etc

After importing the Excel , the first few entries of our data set looks like this:

```
                                                     sms label
0               አይ እዛ ሚማር አይመስለኝም . እዛ አካባቢ ነው ግን የሚኖረው   ham
1    ወንድሜ እንኳ አያናግረኝም።  እንደ ኤች አይ ቪ በሽተኛ አድርገው ይይዙኛል   ham
2                                                     NaN   ham
3  በቅርቡ ወደ ቤት እመለሳለሁ እናም ዛሬ ስለዚህ ነገር ድጋሚ ማውራት አልፈ...   ham
4  ለዚህ እስትንፋስ ላመሰግናችሁ ትክክለኛዎቹን ቃላት እየፈለግኩ ነው ። እ...   ham
5                                  ከዊል ጋር እሁድ እራት አለኝ!!   ham
6                                     ኦ እሺ ... እዚ አያለው :)   ham
7            ስሜትሽ እንደዚህ ከሆነ ጥሩ ነው፤ እንደ ስሜትሽ ነው የሚሆነው   ham
8                           ስሙን የሚትጽፈው በቁም ነገር እንደዚ ነው?   ham
9                       ለ 2 ወር ለመሞከር እሞክራለሁ። ሃሃሃ  ስቀልድ ነው   ham
```

Data Preparation and Representation

Cleaning textual data is a little different from regular data cleaning. There is a much heavier emphasis on text normalization than removing outliers or leverage points. We go through several methods of normalization

The target variable is categorical (ham, spam) and we needed to convert into a binary variable because machine learning models always take numbers as input and not the text. We replace ham with (meaning not a spam) and spam with 1 (meaning that the SMS is a spam)

- Removing punctuations both Local and general punctuations
- Normalizing letters such as, ሀ፣ሕ፣ኅ፣አ፣ኣ etc.
- Tokenization

Training

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

X_train & y_train are the training data. X_train contains all features to be used for training the model, and y_train contains its corresponding target variable. X_test & y_test, on the other hand are the data to be used for validating the model, hence, haven't been introduced to our model. X_test contains feature columns to be predicted. y_test as the target variable with actual values is contrasted against the predicted values,

Split the loaded dataset into two, 80% of to train, and 20% as a validation set. Training data in th e *X_train* and *Y_train* for preparing models and a *X_test* and *Y_test* sets to use for later.

Model Selection

By tokenizing the messages into various special tokens, vocabulary can be completed, however, we still need to rate the token until we get each token. After getting each token, we need to score for each token

Different terms have varying relevance when used to describe contents. This effect is captured through the assignment of numerical weights to each term of the sms messages

For the machine learning algorithm we need a way to represent text data and the tfidf vectorizer helps accomplish that goal. It is a way to derive characteristics from the text to be used in machine learning algorithms.

The problem statement was a good indicator to select the appropriate model, a binary classification problem determining if a given SMS text is spam or not,We tried two models' accuracy estimations, *logistic regression and MultinomialNb*, with logistic regression presenting the highest score

Evaluate the model on the training data set

For logistic regression the accuracy is

```
Accuraccy: 0.9601990049751243
```

For MultinomialNB the accuracy is

```
Accuraccy: 0.9203980099502488
```

- This means the model is good for this problem