SPAM/HAM CLASSIFICATION

Project Report

ALLA SWAPNIKA SRI
GROUP-7

2022

Table of Contents

Introduction

The file we are going to use contains a collection of more than 5 thousand SMS phone messages. Using labeled ham and spam examples, we'll **train a machine learning model to learn to discriminate between ham/spam automatically**. Then, with a trained model, we'll be able to **classify arbitrary unlabeled messages** as ham or spam.

Here I am going to develop an SMS spam detector using **SciKit Learn's Naive Bayes classifier algorithm**. However before feeding data to Machine Learning NB algorithim, we need to process each SMS with the help of Natural Language libraries.

Here the messages are in the human-readable language which computer can't understand, so we have to use the NLP to make it possible for computers to read human (natural) language SMS and determine which parts are important.

So, Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.

NLP makes it possible for computers to read the text, hear speech, interpret it, measure sentiment and determine which parts are important.

The input and output of an NLP system can be –

Speech
Written Text

Data Description

We use a database of 5574 text messages from UCI Machine Learning repositor gathered in 2012. It contains a collection of 425 SMS spam messages manually extracted from the Grumbletext Web site (a UK forum in which cell phone users make public claims about SMS spam), a subset of 3,375 SMS randomly chosen non-spam (ham) messages of the NUS SMS Corpus (NSC), a list of 450 SMS non-spam messages collected from Caroline Tag's PhD Thesis, and the SMS Spam Corpus v.0.1 Big (1,002 SMS nonspam and 322 spam messages publicly available). The dataset is a large text file, in which each line starts with the label of the message, followed the text message string.

After preprocessing of the data and extraction of features, machine learning techniques such as naive Bayes, SVM, and other methods are applied to the samples, and their performances are compared

Given data set:

3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives around here though
5	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send
6	ham	Even my brother is not like to speak with me. They treat me like aids patent.
7	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to c
8	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL3
9	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update C
10	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
11	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 i
12	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD
13	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil m
14	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
15	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com
16	ham	Oh k...i'm watching here:)
17	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
18	ham	Fine if that s the way u feel. That s the way its gota b
19	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20
20	ham	Is that seriously how you spell his name?
21	ham	I'm going to try for 2 months ha ha only joking
22	ham	So ü pay first lar... Then when is da stock comin...

Libraries installed:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- NLTK
- SkLearn
- Pickle
- Flask
- WordCloud

Html and CSS is used fot wed deployment interface for the project

Data Cleaning/Preprocessing

1) The Dataset had two columns: 1st containing the class of data ham or spam and 2nd containing a string which is the text message.

2) All English Stopwords were imported from NLTK and were removed if found in
the sentences.

3) We used a basic Count vectorizer from Sklearn library in order to tokenize and vectorize the string of text. The Count Vectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. It can be used as follows:

a) Create an instance of the CountVectorizer class.

b) Call the fit() function in order to learn a vocabulary from document.

c) Call the transform() function on the document as needed to encode each as a
vector. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document.

4) We got our numeric or real feature vector from string of text messages.

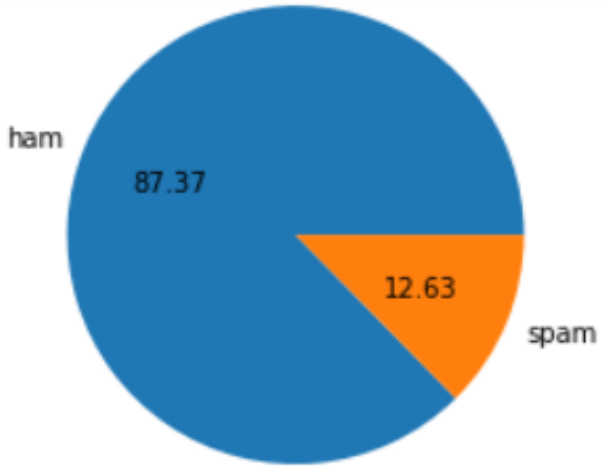
5) We next perform the test train split in the ratio of 70:30, we select the samples randomly.

6)We feed the X_train, X_test, y_train, y_test to different ML models namely,
Multinomial naïve bayes, Support vector machines, K nearest neighbours, Random
forest and AdaBoost.

After removining some of the un required attributes the resultant data set is as follow:

	Target	Text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

```
Target    0
Text      0
dtype: int64
```



Length of the comments after cleaning:

Acknowledgements

Most reports conclude with a page acknowledging the contributions of the people who worked tirelessly on the projects mentioned within. Below, list down the names of those committed to these projects, such as:

- Those responsible for concept and coordination
- Your group of researchers
- The writers behind the impact report
- The designers of the impact report
- Your colleagues from Local and Partner Organizations
- Your contributors and donors

We thank you for your continued support in our efforts to contribute to the SDGs.

Contact

Your NFP Name
123 Anywhere St., Any City, ST 12345
123-456-7890

ww.reallygreatsite.com
hello@reallygreatsite.com
@reallygreatsite

Message From Our Leaders

This section is an opportunity to demonstrate how top management is taking the lead and giving direction to your NFP's sustainability efforts.

An authentic, sincere and mission-driven statement from the chair, CEO or president signals commitment and sets the tone for the rest of the report. It can include an overview of the vision, direction and strategy the organization is committed to take, to help move the needle on the SDGs to meet the global 2030 deadline.

Catch your reader's eye by highlighting one of your main points in this space.

his message demonstrates the organization's understanding of its impact and responsibility to people and the planet. Importantly, it also builds the audience's trust and confidence in the organization.