Spam Detection using NLP Techniques

BollamPragna, M.RamaBai

Abstract*—Natural Language Processing is a vital field of research having applications in different subjects. Text Classification is a part of NLP where the text is converted into a machine-readable form by performing various methods. Tokenizing, part-of-speech tagging, stemming, chunking are some of the text classification methods. Implementing these methods on our data gives us a classified data on which we will train the model to detect spam and ham messages using Scikit-Learn Classifiers. We proposed a model to solve the issue of classifying messages as spam or ham by experimenting and analyzing the relative strengths of several machine learning algorithms such as K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, Logistic Regression, SGD Classifier, Multinomial Naive Bayes(NB), Support Vector Machine(SVM) to have a logical comparison of the performance measures of the methods we utilized in this research. The algorithm we proposed achieved an average accuracy of 98.49% with SVM model on 'SMS Spam Collection' dataset.

I. INTRODUCTION

We get hundreds of messages from unknown sources and our inbox is filled with unwanted emails. These unwanted messages are called spam and essential messages are called ham mails. We will prepare a model that will categorize messages in mobile devices as spam or ham. In order to achieve this, data from the messages is to be collected first and natural language processing techniques are to be applied on it.

The spam filtering among messages helps the mobile user to have a good visualization of the inbox. Unnecessary messages will be marked as spam so users need not waste their time reading them. In this paper, we propose to classify data in the messages as either spam (unwanted) or ham(wanted) messages. We devised our own spam detector.

II. RELATED WORK

Identifying spam messages from inbox has been done by various methodologies. Below are some of the approaches.

1. Sethi, P., Bhandari, V., &Kohli, B. (2017). "SMS spam detection and comparison of various machine learning algorithms." 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN).

2. DelviaArifin, D., Shaufiah, &Bijaksana, M. A. (2016)."Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier."2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob).

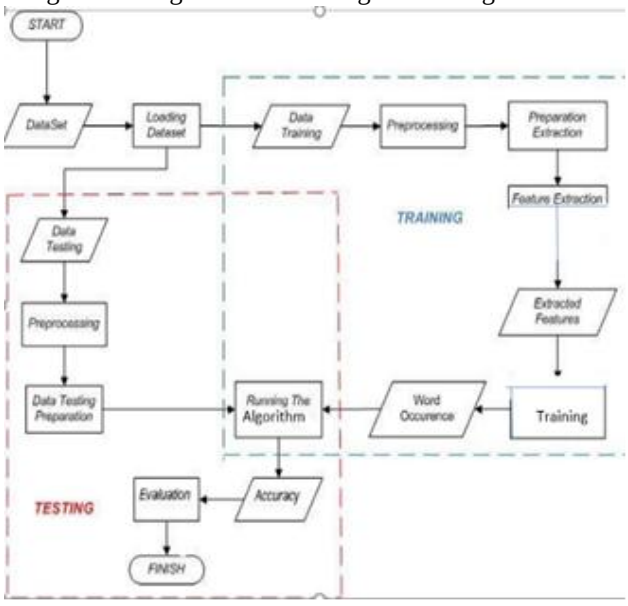
3. Agarwal, S., Kaur, S., &Garhwal, S. (2015). "SMS spam detection for Indian messages." 2015 1st International Conference on Next Generation Computing Technologies (NGCT).

4. Gupta, M., Bakliwal, A., Agarwal, S., &Mehndiratta, P. (2018). "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers." 2018 Eleventh International Conference on Contemporary Computing (IC3).

III. ALGORITHM

Spam is unsolicited bulk messages that are not required for the users but are forced into their inbox. Spams are sent mostly by advertisers, tricksters or by fraud people. Understanding if a message is spam or not can be easily done by reading it once. Our purpose is to detect spam by using various algorithms and measuring their accuracy to find the best fitting algorithm. The common classical approaches which use white-lists and black-lists methods do not work properly as they are only capable of blocking an entire server (source) from sending messages, that may contain some important messages (false positives). Therefore, spam filtration is to be done using text classification techniques.

In our experiments we performed text pre-processing in the first part which gives the annotated data that is split into two parts called training set and testing set to check the accuracy. The below figure represents the flow of our training and testing on the data using various algorithms.



Various types of spam filters used are given below.

1. Blatant Blocking-

Revised Version Manuscript Received on 10 September, 2019.
BollamPragna, (email: bpragna98@gmail.com)
Dr.M.RamaBai(Professor), (email: rama@mgit.ac.in)

Emails are deleted even before they reach the inbox. This blocking is done blatantly.

2. Bulk Email Filter-

This filters out those emails which are passed on through other categories but are unnecessary or spam messages.

3. Category Filters-

According to the specific content like email addresses etc, a user is allowed to define their own rule to enable the filtering of the messages. There can be one or more user-defined rules.

4. Null Sender Disposition-

Messages are disposed if they do not have an SMTP envelope sender address.

5. Null Sender Header Tag Validation-

All the messages are validated by checking security digital signature of each message in the inbox.

IV. FRAMEWORK

Various tools, techniques and data set used in this work is described in this section. As cellular messages repeatedly have a number of acronyms, efficiency of the filters is affected by it. So a large and valid message dataset is used in this process.

4.1 Tools and Algorithms

The machine learning algorithms used in the work are described in detail in this section.

4.1.1 Naïve Bayes (NB)

Naive Bayes Classifier uses Bayes Theorem, which determines the occurrence probability of an event considering the probability of an occurred event. Linearly separable problems are solved extremely well by Naïve Bayes classifier and for non-linearly separable questions, it performs reasonably good.

Multinomial Naive Bayes classifier uses a multinomial distribution for each one of the features generated on data. This is a particular instance of a Naive Bayes classifier.

4.1.2 Stochastic Gradient Descent

An objective function is optimized iteratively with suitable smoothness properties (e.g. subdifferentiable or differentiable) in Stochastic Gradient Descent Algorithm (SGD).

4.1.3 Support Vector Machine (SVM)

The SVM classifier creates an N-dimensional hyperplane which divides the data into two categories. SVM models are similar to a Neural Network. SVM classifier usually takes the input data and for every input taken it outputs the class to which this input belongs. Two class problems are solved by SVM which is a non-probabilistic binary linear classifier.

4.1.4 Logistic Regression

Although many sophisticated statistical models exit, Logistic regression(or logit regression) uses a logistic function to model a binary dependent variable. In regression analysis, it estimates the parameters of a logit model which is in the form of binary regression. In mathematical words a binary logistic model has a dependent variable with two possible outcomes. These outcomes can be labelled as “0” and “1” which usually represent two opposite classes such as pass/fail, win/loose.

4.1.5K-Nearest Neighbours

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input is same but the output varies. Input contains the closest k training examples in the feature space whereas, the output is decided depending on the k-NN usage for classification or regression:

1. The output of a k-NN classifier has a class membership. Classification of objects is done by a plurality vote of its neighbours and the object is assigned to the class that is most common among its k nearest neighbours. When ‘k’ has a value one, the object is directly given to the class which has a single nearest neighbour.
2. In k-NN regression, property value gives the output for the object. This value is the average of the values of k nearest neighbours.

4.1.6Random Forest Classifier

Ensemble learning method for classification, regression and other tasks is Random forests(or random decision forests) which operate at training time by constructing a multitude of decision trees and giving the class as output which is the mean prediction of the individual trees or mode of the classes.

4.1.7 Decision Tree Classifier

A decision tree has a flowchart-like structure containing nodes. Each internal node is a "test" on an attribute which branches into two nodes. These two nodes are the outputs of the decision in the test. The tree ends with leaf nodes which represent a class label (decision taken after computing all attributes). The path from the root node to reach one leaf node is a single classification rule. Three types of nodes are in a decision tree which are given below.

1. Decision nodes – generally given by squares
2. Chance nodes – generally given by circles
3. End nodes – generally given by triangles.

4.1.8 Ensemble Methods

The meta-algorithms combine various machine learning methods into one model which predicts the output and either reduces the variance (bagging), bias (boosting), or enhance predictions (stacking). We used the voting classifier in our experiment.

Voting Classifier

The Ensemble Vote Classifier is a meta-classifier that combines machine learning classifiers for classification by referring to majority or plurality voting which are either similar or conceptually different. (For simplicity, we will refer to both majority and plurality voting as majority

voting.) It implements "hard" and "soft" voting. In hard voting, the most frequent prediction done by classification models is considered as final class label. In soft voting, the averaging of class-probabilities is taken as output(only recommended if the classifiers are well-calibrated).

4.2 Dataset

Model is trained by giving a complete data on which supervised learning can be done. To achieve this, user message data has to be collected and mark them as either spam or ham. The data set used is given by the UCI Machine Learning Repository. It has over 5000 SMS messages which are labelled and are collected for message spam research.

The below image is a screenshot of the dataset that has been collected for Spam research which contains all the mobile messages. These messages are tagged accordingly as legitimate(ham) or spam. There are 5,574 rows with two columns in the dataset.

	0	1
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
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 aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
10	ham	I'm gonna be home soon and i don't want to tal...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...
12	spam	URGENT! You have won a 1 week FREE membership ...
13	ham	I've been searching for the right words to tha...
14	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
15	spam	XXXMobileMovieClub: To use your credit, click ...
16	ham	Oh k...i'm watching here:)
17	ham	Eh u remember how 2 spell his name... Yes i di...
18	ham	Fine if that□s the way u feel. That□s the way ...
19	spam	England v Macedonia - dont miss the goals/team...
20	ham	Is that seriously how you spell his name?

V. EXPERIMENTS

Various experiments are applied on the dataset which were based on Natural Language Processing(NLP) concepts like label encoding, tokenization, stemming, stop word removal, generating features and then applied ensemble method – voting classifier. All these experiments performed in the model classify the data set accurately.

5.1 Pre-Processing

The messages have to be pre-processed for the removal of unwanted punctuation, grammar, stop words etc.

5.1.1 Label Encoding

Label Encoder encode labels with values between '0' and 'n-1' where n represents the number of distinct labels for the classes. Same value is as assigned to the labels which are repeated earlier. In our experiment, we convert the class labels to binary values, where '0' is ham and '1' is spam.

5.1.2 Stop Word Removal

When using Natural Language Processing(NLP), our goal is to perform some analysis or processing so that a computer can respond to text appropriately.

A machine cannot understand the human readable form. So, data has to be pre-processed in order to make it machine-readable. This is "pre-processing" of which one of the major forms is to filter out useless data. This useless data (words) is generally referred as 'stop words' in Natural Language Processing(NLP).

5.1.3Stemming

Stemming is another pre-processing step that normalize sentences. Stemming is a way to account for the variations of words and sentences which often have a same meaning; furthermore, it will help us shorten the sentences and shorten our lookup. For example, consider the following sentence:

1. I was taking a ride on my horse.
2. I was riding my horse.

These sentences mean the same thing, as noted by the same tense in each sentence; however, that isn't intuitively understood by the computer. To account for all the variations of words in the English language, we can use the Porter stemmer, which has been around since 1979.

5.2 Feature Generation

Feature engineering is the process of constructing features for machine learning algorithms by using the knowledge of that specific domain. The words in each text message are the features on which the algorithm will predict the output. For this purpose, it will be necessary to tokenize each word. The most common 1500 words that are generated in feature generation will be used as our features. Then the data is split in training and testing datasets with a test size of 25%.

5.3 Implementation of Algorithms

We need to import each algorithm from scikit-learn library along with performance metrics. We require accuracy score and classification report metrics to predict the accuracy and give a classified report on the output.

VI. RESULT ANALYSIS

In this part, we compare the performance and accuracy of one algorithm with other machine learning algorithms. The algorithms applied in this work gave out high accuracy values. But among the various experiments done, the spam message is best predicted by Support Vector Classifier with an accuracy of 98.49%. Other algorithms have similar accuracy with a variation of about 3%. The below picture represents the accuracy values of various algorithms applied on spam dataset.

```
K Nearest Neighbors Accuracy: 95.19023689877962
Decision Tree Accuracy: 96.12347451543431
Random Forest Accuracy: 97.63101220387652
Logistic Regression Accuracy: 98.56424982053123
SGD Classifier Accuracy: 97.98994974874373
Naive Bayes Accuracy: 98.1335247666906
SVM Linear Accuracy: 98.63603732950466
```

The ensemble vote classifier applied on above algorithms gave an accuracy of 98.6%. The classification report below gives a detailed description of precision values.

Our model predicted legitimate messages as ham 1198 times but failed 17 times and it correctly predicted spam 176 times but failed twice. This is represented by the classification report below.

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1200
1	0.99	0.91	0.95	193
micro avg	0.99	0.99	0.99	1393
macro avg	0.99	0.96	0.97	1393
weighted avg	0.99	0.99	0.99	1393

		predicted	
		ham	spam
actual	ham	1198	2
	spam	17	176

VII. CONCLUSION AND FURTHER WORK

The previously collected mails are taken as dataset and for each input in the set, a class is predicted and given as output. The messages are first tagged correctly to apply algorithms on them. Applying various classifiers helps us to know the best and the worst algorithms for a problem.

The SVM algorithm was very effective outputting a high success percentage, up to 98%. This confirms that SVM is one of the best model for filtering spam messages in the inbox. This model need to be improved to understand sarcasm, context on the whole which could be essential while detecting spam.

VIII. ACKNOWLEDGMENTS

I would like to thank Dr.M.RamaBai, Professor, Dept. of CSE, for extending her help, support and guidance during this work.

REFERENCES

1. Navaney, P., Dubey, G., &Rana, A. (2018). "SMS Spam Filtering Using Supervised Machine Learning Algorithms." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).
2. Mathew, K., &Issac, B. (2011). "Intelligent spam classification for mobile text message." Proceedings of 2011 International Conference on Computer Science and Network Technology.