# RESFIL: A Rest Microservice Based Spam Classification API on the Cloud

**Megha Patil[1] Divyani.E G.[2] Kandibedala Mounika[3] Shivaparasad K M[4]**
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Rao Bahadhur Y Mahabaleshwarappa Engineering College, Ballari, India

*Abstract*— RESFIL is a cloud-based microservice-api based on rest. The api must take as an input a message. A multi-tiered program must forward the message to a model of machine learning which will represent the outcome. A mobile application on android/iOS will be developed to show the use case. This model will be prepared on the basis of a comparative study of various classification models including methods of deep learning applied using tensor flow/pytorch. The product will be sent back to customer. RESFIL can be applied on various device types, such as a mobile app/site. RESFIL will be deployed to google cloud platform (GCP) so that any client with the right authentication can access the micro service.

***Keywords:*** Await, Asset, Json, On submit, Spam messages

## I. INTRODUCTION

Significant approaches to spam filtering include text analysis, image analysis, domain name white and blacklists, and group oriented approaches. Text content review of mails is a commonly used technique to spam. Many deployable applications are available on the server and client sides. Naive Bayes is one of the most widely used algorithms in those strategies. ResFil is a cloud-based rest API that assists in classifying text and images as spam or not spam. A mobile app for calling the appropriate document to the API will be created. The classification algorithms of machine learning will be used to train the model needed for the classification. A Deep learning-based approach will be implemented using Neural network, and a comparative analysis with classical algorithms such as Naïve Bayes, SVM (Support Vector Machine) etc. will be used to pick the algorithm based on this approach. ResFil is a cloud-based rest api that assists in classifying text and images as spam or not spam. A mobile app is being built with the correct document to call the api. The classification algorithms of machine learning will be used to train the model needed for the classification. A Deep learning based approach using Neural network will be applied and a comparative analysis with the classical algorithms such as Naive Bayes, SVM(Support Vector Machine) etc. will be used to pick the algorithm based on this approach.

## II. LITERATURE SURVEY

### *A. Study on the Effectiveness of Spam Detection Technologies*

Article: January2016
Published By: Muhammad Iqbal, Malik Muneeb Abid, Mushtaq Ahmad, Faisal Khurshid Muhammad Iqbal, Malik Muneeb Abid, Mushtaq Ahmad, Faisal Khurshid have proposed an effective spam detection technique for analysis and detection of spam. This work focuses on systematically evaluating the strength and weakness of existing spam detection technologies and presents taxonomy of known approaches.

### *B. Text Mining Approach to Detect Spam in Emails*

Article: Feb2016
Published By: Zahra Khan and Usman Qamar
Zahra Khan and Usman Qamar have proposed an effective spam detection technique for analysis and detection of spam. With technological advancement most of the modern day communication take splace through emails. This has made the contact process much quicker and easier as time is saved. After pre-processing the data, various algorithms are applied over the sample data set for classification.

### *C. A Study of Machine Learning Classifiers for Spam Detection*

Article: Sep2016
Published By: Shrawan Kumar Trivedi Shrawan Kumar Trivedi have proposed an effective spam detection technique for analysis and detection of spam.

Email correspondence is required in the present world but unsolicited emails ham per these communications. The high precision of the SVM, and the low false positiv e rate. SVM' straining time to build the model, however, is high, but as the results on other parameters are positive the time does not present such an issue.

## III. BACKGROUND CLASSIFICATION

Classification is a technique for categorizing our data into a desired and distinct number of classes, where each class maybe given name.

Classification applications are: speech recognition, recognition of the handwriting, biometric identification, classification of documents, etc.

### *A. Naive Bayes(Classifier)*

Naive Bayes is a probabilistic, Bayes theorem-inspired classifier. We are conditionally independent under a simple assumption which is the attributes.

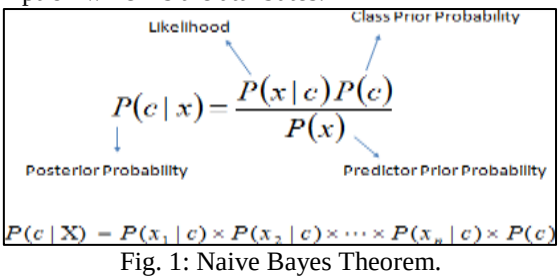



Fig. 1: Naive Bayes Theorem.

Classification is carried out by deriving the maximum posterior which is the maximum P(Ci) with the a fore mentioned statement that applies to the Bayes theorem. This assumption significantly reduces computational costs by calculating only the distribution of the groups. Naive Bayes is a very easy algorithm to implement, and in most cases good results have been obtained. Naive Bayes may be suffering from a problem called the zero probability question.

When the conditional probability for a given attribute is zero, a true prediction is not made. Use a Laplacian estimator this must be set specifically.

Advantages: For estimating the required parameters this algorithm requires a small amount of training data. Compared to more sophisticated methods Naive Bayes classifiers are extremely efficient. Disadvantages: The weak estimator is considered to be Naive Bayes.

### B. *Support Vector Machine*

Definition: Support vector machine is a representation of the training data as space points separated by a simple gap as large as possible into categories. *New* examples are then mapped into the same space, and predicted to belong to a category based on which side of the gap they fall.

Advantages: Effective in high dimensional spaces and using in decision method a sub-set of training points so it is also efficient in memory.

Disadvantages: The algorithm does not have clear estimates of probability, these are determined using a expensive five- fold cross-validation.

### C. *Decision Tree*

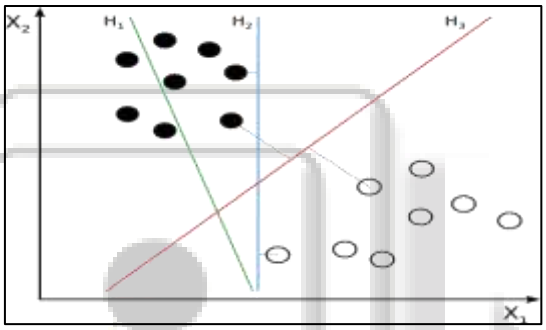



Fig. 2:

Decision Tree makes decisions using a tree-like pattern, as its name suggests. Centered on the most important differentiators in the input variables, it separates the sample into two or more homogeneous sets (leaves). The algorithm considers all features to choose a differentiator (predictor) and does a binary split on them. It will then pick the one with the least cost, and repeat recursively, until it successfully splits the data across all leaves.

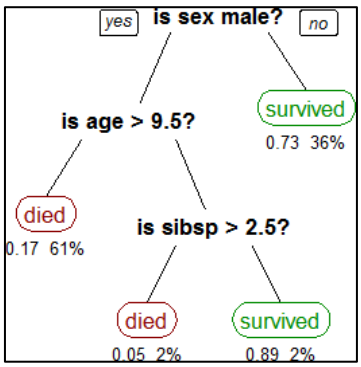



Fig. 3: Tree like representation of data in Decision tree

Advantages: Tree decision is easy to grasp and visualise.

Disadvantages: Decision tree can construct complex trees which are not well generalized.

### D. *Random Forest*

Random forest is an ensemble model that grows multiple tree species and classifies artefacts based on all the trees' "votes." E.g. A class that has the most votes from all the trees is allocated an entity. In doing so, it could mitigate the issue of high bias.

Image result for diagram for random forest

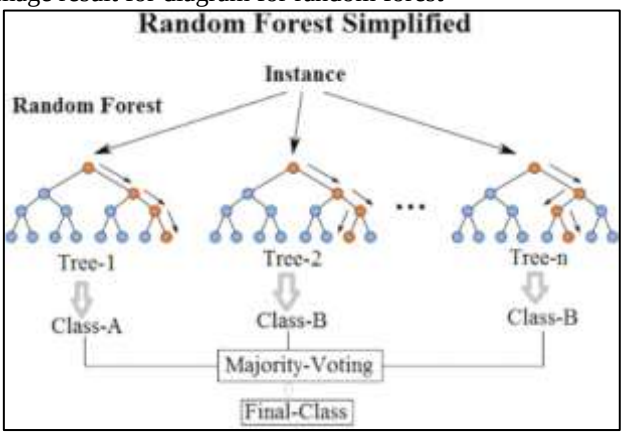


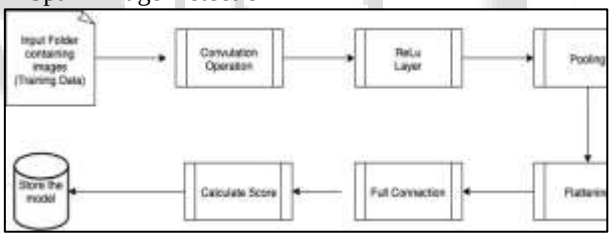

Fig. 4: Random forest

Random forest classifier is a meta- estimator which fits a number of decision trees on different dataset sub-samples and uses average to improve model predictive accuracy and over- fitting controls. The size of the sub-sample is always the same as the

## IV. METHODOLOGY

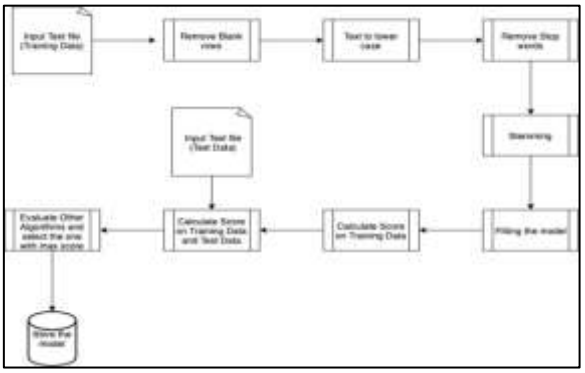
Data Flow Diagrams: Level 1:

– Spam Image Detection




– Spam Text Detection

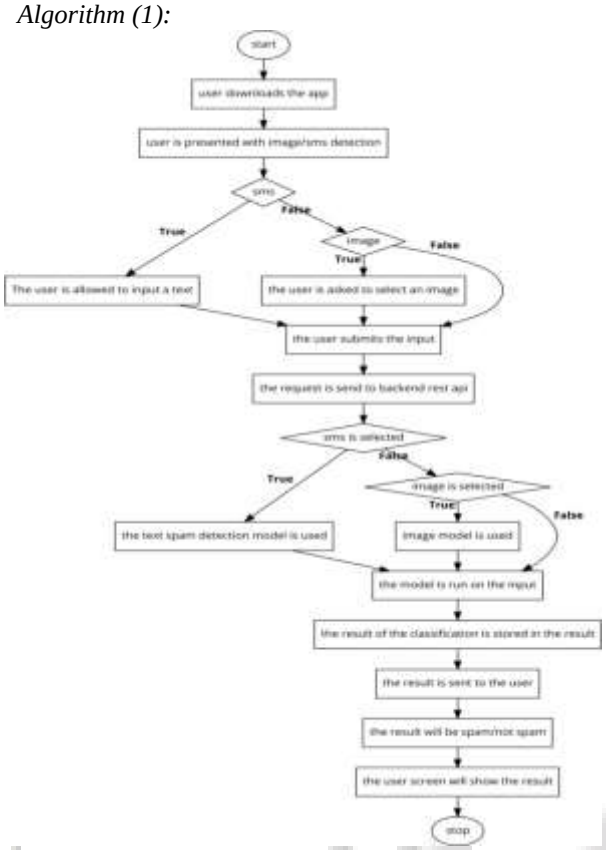
- Level 2:



*A. Algorithms*

*1) Algorithm (1):*




Explanation:-

The user logs into the application. The user is given 2 options to select either an image or a text. If the user selects text then then input area is provided in the next screen where the user will have an option to select the text. The user can input the text. the user must press the submit option and the backend api with the provided input will be called.

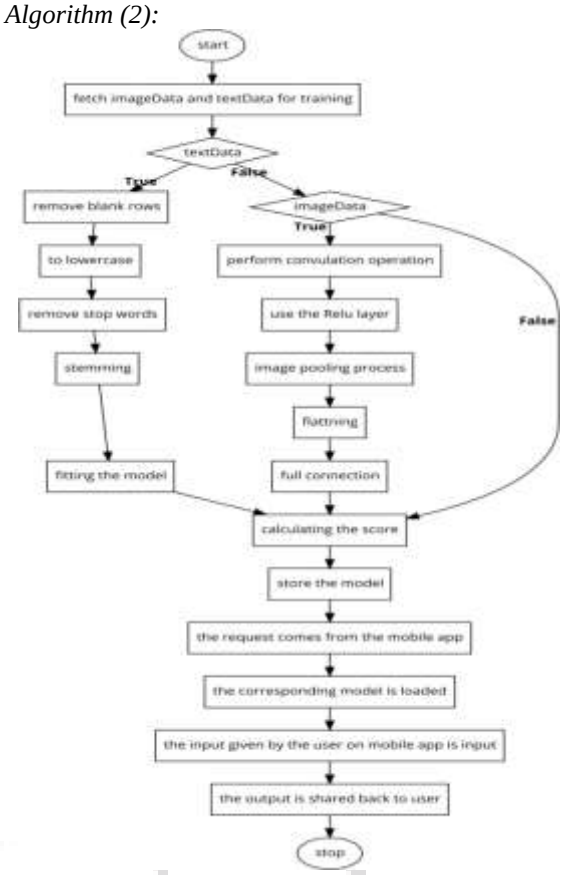
Once the api is called the back end will load the sms Model and pass in the given input. The model will classify the given input as spam or not spam and sent it back to the user.

In case of an image, the user is asked to select an image from the library. Upon selection of the image the user must send.

The api passes this picture into the api backend flask. The request will then be submitted to the image Model that was previously trained and either the image is a spam image or not will be the model's output.

This result will be forwarded back to the mobile app and told accordingly.

*2) Algorithm (2):*




Explanation:-

The cycle begins with getting the data needed to train the model. There are two types of data that are appropriate for processing the model. The first data is a list of sms marked as spam, and others not marked as spam. Additionally we have a collection of pictures that are spam and those that are not.

Let's start processing the data about the text. There is a collection of pre- processing steps which perform operations such as removing blank spaces, converting all text to lower case, removing stop words, stemming etc. And the text data is cleaned for further processing. Now the processed text is sent to various classification algorithms such as Naive Bayes, Random Forest, Regression logistics etc.

As seen in the results section, Naive Bayes gives the training data and test data a pretty good effectiveness.

Similarly, we can POST / spam / detect/text for the document, the flask server will load the document model and run the process of classification.

Therefore, the final result is given in the form of json, and the result is interpreted by the mobile app and informs the user accordingly.

## V. DATASET

Spam Detection dataset and file format dataset for text spam classification

Columns:

| Type | Number of images |
|---|---|
| Spam Images | 931 |
| Not spam images | 812 |

Image classification dataset:

There are two folders. One folder spam contains all the spam images. Another folder natural contains not spam images.

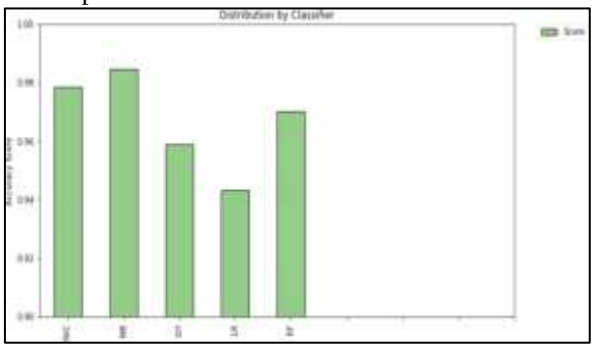
*A. Example of a spam images are as follows:-*


The score for various algorithms is as below:

| label | Algorithm |
|---|---|
| SVC | SVC (Support Vector Classifier) with the sigmoid kernel |
| NB | Multinomial Naive Bayes |
| DT | Decision Tree Classifier |
| LR | Logistic Regression Classifier |
| RF | Random Forest Classifier |

| | Score |
|---|---|
| SVC | 0.978469 |
| NB | 0.984450 |
| DT | 0.958732 |
| LR | 0.943182 |
| RF | 0.970096 |

When we plot in the bar charts it is as below:-



## VI. EXPERIMENTAL RESULT

The results of spam classification using mobile app

Fig. (a):

Fig. (b):

Fig. (c):

## VII. CONCLUSION

We did spam detection on text as well as photos in this article. We performed a comparative study of various classical machine learning classification algorithms for the text in order to find the best fit for the given function. We conclude after conducting experiments that Naive Bayes provides the best classification for the dataset used in research. We used traditional CNN methods to identify the images as spam rather than spam. We have created a mobile app that uses respond native that can run in both in and android that takes an input as image or text and sends the data to the cloud via server where the model is used and the user is sent back the response.

## REFERENCES

[1] Xianghan Zheng, ZhipengZeng, ZheyiChen, YuanlongYu, ChunmingRong "Detecting spammers onsocial networks "journal Neurocomputing 159(2015)27–34.

[2] Atefeh Heydari, Mohammad Ali Tavakoli, NaomieSalim, Zahra Heydari "Detection of review spam: A survey"journal Expert Systems with Applications 42 (2015) 3634–3642.
[3] Fahim A., Mutahira N. Naseem "Facebook Spam and Spam Filter Using Artificial Neural Networks" International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:1, 2015.
[4] Mohammed N. Al- Kabi, Izzat M. Alsmadi, Heider A.Wahsheh "Evaluation of SpamImpact on Arabic WebsitesPopularity" Journal of King Saud University – Computerand Information Sciences (2015) 27, 222–229.
[5] Iqbal, Muhammad, et al. "Study on the effectiveness of spam detection technologies." IJInformation Technology and Computer Science 1 (2016): 11-21.
[6] Khan, Zahra, and Usman Qamar. "Text Mining Approach to Detect Spam in Emails." TheInternational Conference on Innovations in Intelligent Systems and Computing Technologies (ICIISCT2016). 2016.
[7] Xu, Hailu, Weiqing Sun, and Ahmad Javaid. "Efficient spam detection across online social networks." 2016 IEEE International Conference on Big Data Analysis (ICBDA). IEEE, 2016.
[8] Meda, Claudia, et al. "Spam detection of Twitter traffic: A framework based on random forests and non-uniform feature sampling." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.
[9] Trivedi, Shrawan Kumar. "A study of machine learning classifiers for spam detection." 2016
[10] 4th international symposium on computational and business intelligence (ISCBI). IEEE, 2016.
[11] Liu, Pingchuan, and Teng-Sheng Moh. "Content based spam E-mail filtering." 2016
[12] International Conference on Collaboration Technologies and Systems (CTS). IEEE, 2016.
[13] Jain, Gauri, Manisha Sharma, and Basant Agarwal. "Spam detection on social media text." Intern. Journal of Computer Sciences and Engineering 5 (2017).