International Journal of Research in Engineering, Science and Management
Volume-1, Issue-10, October-2018
www.ijresm.com | ISSN (Online): 2581-5792

Intelligent Spam Detection Micro Service with Server Less Computing

Akshay Shelatkar[1], Neal Yadav[2], Abhijit Karve[3]
[1,2]*Student, Department of Information Technology, Zeal College of Engineering and Research, Pune, India*
[3]*Assistant Professor, Dept. of Information Technology, Zeal College of Engineering and Research, Pune, India*

***Abstract*—Today for personal and business purpose most of the users uses email as one of the most important source for communication. The use of email is increasing day by day without being affected by alternative ways of communication such as social networking sites, SMS, mobile applications, electronic messages. As frauds using email classification is increasing due to extensive use of emails, it becomes very important issue to classify mails as fraud or normal mails. The Intelligent Spam Detection System (ISDS) provide automatic way to classify emails as SPAM i.e. fraud mail and HAM i.e. normal mail using multiple machine learning algorithms.**

***Index Terms*—intelligent spam detection system, machine learning, naive bayes algorithm, dataset.**

I. Introduction

The internet has become necessary part of our lives and email is key application for communication. User spends much time on classifying the mails, so it is necessary to classify them automatically. And due to increase in users of email system the frauds are also increasing. To reduce these frauds we are proposing the system called as Intelligent Spam Detection System. This ISDS classifies the incoming emails as spam or ham (normal mails) and will also enable the user to categorize them in one or more categories such as personal, official, social, promotions and many more user defined categories. To classify these incoming mails, Intelligent Spam Detection System uses different machine learning algorithm. Many machine learning algorithms such as Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine etc. are studied in-order to classify the emails and the most accurate one is selected for classification. The proposed system differs from Google or Yahoo as it allows the user to categorize the mails as per their need, for example he/she can categorize the mails as personal or official separately. This system provides webpage as well as android application which are accessible anywhere and anytime. ISDS is a micro-service so can access any mail from any service provider for example Google, Yahoo. Rediff etc. The dataset from University of California is used in order to train the machine learning model. Instead of considering mail word by word it considers whole message and classifies the mail as spam or ham. In our proposed system we are using Amazon Web Services for computation purpose.

II. Literature Survey

A. Tokenization

The term tokenization comes under the Preprocessing of the data. The divination of text into some meaningful pieces is nothing but tokenization. It is nothing but act of breaking a sentence into pieces such as meaningful symbols phrases etc. These set of pieces is called as tokens. These set of tokens are considered as a set of input for the next processing of the data. The next processing is nothing but parsing and text mining.

Tokenization is dependent on simple heuristics in order to separate tokens using following steps.

1. In tokenization words are separated by blank spaces, or punctuation marks or line breaks.
2. Blank Spaces or punctuation marks are optional. It depends on the on the need of the system whether to use them or not.
3. Each and every character in a string is a part of a token. The tokens are nothing but alpha characters, alpha numeric characters or numeric characters only.

B. Lemmatization

Lemmatization is nothing but the grouping together of different forms of the same words. Lemmatization is one of the main part of the natural language processing and natural language understanding. Lemmatization is related to stemming. Basically the goal of both the process is same that is reducing inflectional forms of each word to its base. The main difference between stemming and lemmatization is that stemming cuts the beginning or end of the word and taking back to its root form, while lemmatization considers the morphological analysis of the words.

C. Removal of Stop Words

In data pre-processing it is necessary to convert the data to something which can be understood by the computer. In NLP the useless words are called as stop words. The stop words are such as “the”, “a”, “an”, “in”. The process of removing these types of words is called as removal of stop words.

III. Proposed System

The proposed system consists of webpage or android application in which the user can login with different email

service providers. The incoming mail will be first classified as spam or ham mail. Spam mail is nothing but the fraud mail which is consists of viruses, unwanted messages or phishing links. Spamming is one of the major cyber-attack which can fool the people by sending fake emails and can try to access the confidential information from user or target. And Ham mail is a normal mail. This system also provides the option to user to categorize the incoming mails in different categories as per their need such as official, personal and many more user defined categories. The machine leaning plays important role to classify all the emails. Many machine learning algorithms are developed to classify the mails more accurately. The algorithms like Naive Bayes, Logistic regression, Support vector machine, Random Forest, Neural Network are trained to get more accurate results and algorithm is selected by comparing their accuracy with one another.

A. *Architecture to classify emails automatically*

For the purpose of classification of email, it is needed to have the dataset which contains spam and ham contents. In our system we are using the dataset designed by University of California, which contains two attributes where first is label in which whether the message is spam or ham is mentioned and another field contains actual message. The architecture is consist of three levels where first level is Data Pre-processing, second level is Learning level and third level is Data Classification.

1) Data Pre-processing

After collecting data for classification, the next task is cleansing data which is also called as Data Pre-processing. In this level data cleansing is done in which data is converted into tokens and unwanted words or stop words get eliminated. It results in reducing the amount of data which need to be analyzed. After removal of stop words Stemming and lemmatization takes place on tokens to convert them into their original form.

2) Learning level

In the second level ie the learning level first feature sets are created and extracted. Features are nothing but the signs that represent a measurement of some aspect of given user's behavior. To classify the emails more efficiently extraction of features is essential to make the task of learning more exact. Selection of features is done to enhance the accuracy and efficiency of the classifier.

3) Classifier level

In this level the classifier is developed. This developed classifier is saved for further classification. Finally, at the classification level this developed classifier classifies all the incoming mail into specific classes such as legitimate or spam.

4) Naive bayes classifier

Naive Bayes classifier is a classifier with strong assumption between independent features which works using Bayes theorem. It is a method of text categorization. In order to identify spam and for categorization it uses bag of word feature. This algorithm uses Bayes theorem for determining probabilities. While using Bayes theorem for spam classification is gives probability of certain message is spam or ham based on words in the message or its titles. The main aim of learning is to reduce false positive.

5) Bayes theorem

Let's suppose suspected message contains "replica" word.

The formula can be given as:

$P(S|W) = P(W|S)*P(S) / P(W|S)*P(S) + P(W|H)$

Where,

$P(S|W)$ = the probability that message is spam, and replica word is present in it.

$P(S)$ = It is total probability that the message is spam.

$P(W|S)$ = It is probability that replica appears in message.

$P(H)$ = It is total probability that message is Ham.

$P(W|H)$ = It is probability that the word replica is in ham message.

For getting output, process is divided into three phases:

1. Preprocessing
2. Feature selection
3. Naive Bayes classifier
4. Head and shoulders shots of authors that appear at the end of our papers.

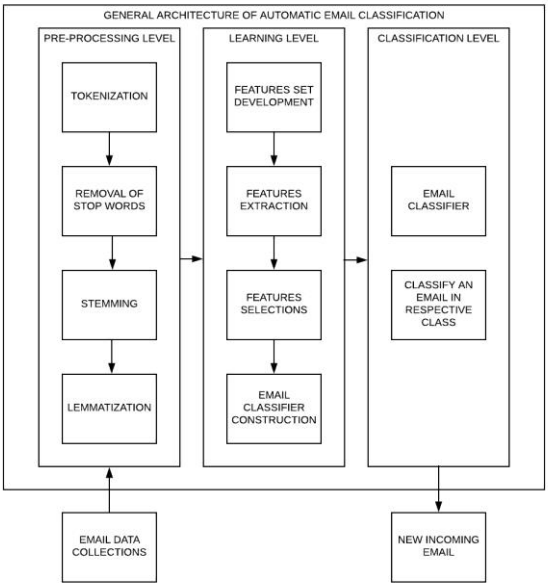

Fig. 1. General architecture of email classification

IV. Advantages

1) No maintenance

As service provider is going to take care of it.

2) More generic

The scope of system is not limited to Google or yahoo, it is more general.

3) Light weight micro service

The system is lightweight and can be use anywhere as it is platform independent.

4) More accurate

As many different algorithms are trained and more accurate is chosen so it gives more accurate classification result.

V. Future Scope of the System

1) *Real time learning*
   Most of the researches focus on classifying historical or past dataset which does not include real time data. So real time learning is a challenge for experts. However it is very difficult to work on real-time data.
2) *Dynamic updating*
   Designing a system which can add or remove the features without redesigning or rebuilding the whole model to work with advance features in spam classification is also a challenge.
3) *Image and text based classification*
   In current system the classification is done only on text message however one can update the system by adding new feature of image based classification so as to improve the performance of classification system.
4) *Language based classification*
   In this system the classification is taking place for mails only in English language but there is no classifier which can classify mails in different languages so one can add feature of classification which can work on multiple languages.

VI. Conclusion

Five major application areas of email classification, namely, spam, phishing, spam and phishing, multi-folder categorization, and other related application areas, were analytically summarized.

VII. Acknowledgement

We would like to acknowledge Zeal college of Engineering and Research, Pune for supporting and encouraging us to work on "Intelligent Spam Detection Micro service using Server less Computing" and we also thank Dr. Ubale (HoD IT dept) and all guides for their helpful guidance and support in improving paper.

References

[1] R. Team, ‘‘Email statistics report, 2015-2019,’’ The Radicati Group, Inc. Palo Alto, CA, USA, Mar. 2015
[2] K. Xu, C. Wen, Q. Yuan, X. He, and J. Tie, ‘‘A mapreduce based parallel SVM for email classification,’’ J. Netw., vol. 9, pp. 1640–1647, Sep. 2014.
[3] E. Blanzieri and A. Bryl, ‘‘A survey of learning-based techniques of email spam filtering,’’ Artif. Intell. Rev., vol. 29, pp. 63–92, Sep. 2008.
[4] T. Ichimura, A. Hara, Y. Kurosawa, T. Ichimura, A. Hara, and Y. Kurosawa, ‘‘A classification method for spam e-mail by self-organizing map and automatically defined groups,’’ in Proc. IEEE Int. Conf. Syst.,Man Cybern., vols. 1–8. Oct. 2007, pp. 310–315.
[5] S. Misina, ‘‘Incremental learning for e-mail classification,’’ in Computational Intelligence, Theory and Application, B. Ruesch, Ed. Berlin,Germany: Springer-Verlag, 2006, pp. 545–553.