IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007140 |

The Mechanism of Spam Comment Detection using Countvectorizer and Naive Bayes Machine Learning algorithms in Python

G Teja Sree, S Sreenivasulu, G Swetha Sree , P Yashwanth Sai Reddy, N Sravan Kumar Reddy, J Jhansi

Department of Electronics and Communication Engineering, Siddharth Institute of Engineering and Technology, Puttur, India

Assistant Professor, Department of Electronics and Communication Engineering, Siddharth Institute of Engineering and Technology, Puttur, India

ABSTRACT: Spamming is the process of posting unwanted and awkward comments on specific posts in any type of social sharing medium or video sharing medium. These messages are posted by bots for reducing ranking or disturbing users' viewing experience which ultimately reduces the rank of website and post. This spamming is done manually also which are mostly seen in most competitive pages. There are few methods which can remove spamming methods which use data mining techniques but in this paper we are automating process of spam comment detection using machine learning algorithmsby assuming dataset of YouTube spam messages and applying Countvectorizer and Naive Bayesalgorithm for clustering on given data set using python programming.

KEYWORDS: Machine Learning, Countvectorizer, Naive Bayes algorithm, Spam words, Python.

I. INTRODUCTION

Nowadays there is popularity and importance of social media sites that are enhanced in people's daily life. Social media allows online users to share their feelings through posting comments. However, more and more spam comments are also being posted in user's accountson social media.Spamming is the process of posting unwanted and not related comments on specific posts in any type of social sharing medium or video sharing medium. These messages are posted by bots for reducing ranking or disturbing users' viewing experience which ultimately reduces the rank of website and post.

However, more and more spam comments are also being posted in user's accountson social media. So the necessity of spam detection is increased. In traditional systems, there are different systems that are used for spam detection such as The Naive Bayes classifier and TFIDF (Term Frequency – Inverse Document Frequency). But these methods do not take the semantic information of the spam words or phrases into account, which leads to incomplete results.

Therefore, the requirement of research is to take into full account the significance of the semantic information of the words within all comments posted, including the vector expression of a word. And the vector distance between words. The skill of mining additional semantic features from words has been widely used in emotional classification and text classification, both of which have achieved good results.

In the existing system data mining techniques are used for detecting spam messages. Most of these methods work only after posting messages. There is a need for a system which can automate this process before posting a message. The main disadvantages of this existing system are Most of the existing works are based on data mining techniques which are not accurate and a time taking process. Detection process is analyzed after the review was given. Here the aim in this work is to extract the features from the given dataset and design a model by generating test and training sets. Then Naive Bayes classifier is applied for clustering and a test and training set is given as input based on this data given message is tested if it is spam or not.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007140 |

II. MATERIALS AND METHODS

The Research has been carried out as two preparations. Initially the comments of some popular YouTube videos are considered. Then they are merged into a single dataset.Then the data in the dataset are cleaned and the features are extracted in which the spam and ham comments are classified using the countvectorizer algorithm. The Naive Bayes machine learning algorithm also applied to the same dataset for clustering and classification. Then the comments are identified whether it is spam or ham and accuracy is evaluated. The accuracy analysis is done using an SPSS tool with 150 sample groups to get different mean and std. deviation values. Finally the accuracy can be calculated by comparing the test and training set.

The Spam.csv (https://www.kaggle.com/lakshmi25npathi/images), the dataset used in this research study was downloaded from Kaggle website and it has a size of 9780 with 5 attributes. The attributes used are Comment_ID, Date, Class, Author and Content. The dataset is collected for predicting the spam comments using machine learning algorithms.Generally 80:20 ratio is applied to split the dataset as training and testing sets. The data model is created using Countvectorizer on training data. Then both algorithms are applied on the testing set and the resultant accuracy is predicted.

The Proposed model has been tested in Anaconda Spyder Software (Jupyter Notebook). The specifications of the system used is with Windows 10 OS, i5 processor, 256GB SSD and 12GB Ram.

The system design (Architecture Diagram) plays a vital role in defining the working of the system. Initially the process starts with importing the required libraries. Then the Dataset is uploaded, the data is filtered and cleaned by removing extra white spaces, punctuation marks, unnecessary symbols and correcting the uppercase and lowercase text. The flow diagram (Fig 1) describes the process of spam classification and detection using Machine learning algorithms

The text classification in the dataset works with the comments of 5 different popular videos from YouTube. Thankfully, the authors who used this dataset in an article on spam collection made the data free available. (https://www.kaggle.com/lakshmi25npathi/images). This collection is composed of one csv file per dataset, where each line is composed with different attributes and the spam was coded as '1' and ham was coded as '0'.

This Research carried out using Support vector and Naive Bayes machine learning algorithms(countvectorizer as supporting algorithm) which are applied to the Dataset to find the good accuracy rate. Here, Table 1 describes the sample dataset collected from the YouTube source. This contains the comments of some popular youtube videos and is merged into a single dataset that contains a size of 9780 with 5 attributes.

III. RELATED WORKS

The data classification and the feature extraction is done using the countvectorizer algorithm. The Spam filtering system is designed to achieve a system without any spam words which can be deployed into any social network sites. A design to filter junk emails is developed using machine learning algorithms. This model is enlarged by assuming a dataset of spam words and applying machine learning algorithms to detect spam words. After designing the model, it is tested using different algorithms by comparing test and train sets.

The opposite findings of this study is a model using sequence mining algorithms to detect the spam comments. The practices of testing the train and test datasets to filter the spam comments with a comparative analysis using Clustering based support vector machine algorithm. The overall analysis of the above opposite findings and similar findings , the performance of the models are compared. The words are tested at every vectorization step. In the opposition findings the model is designed using different machine learning and sequence mining algorithms. These research may not provide accurate results. Hence, this research is more efficient than opposing research.

Some of the future enhancements that can be done to this system are As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment. In future the algorithm can be enhanced and upgraded to predict the spam comments with better accuracy rate and significant results. In this study the model of naive bayes algorithm is trained for the prediction. However, the same approach can be done with other algorithms. But the one we used was very effective with appropriate results.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007140 |

IV. EXPLANATION

Initially The comments of some popular YouTube videos are considered as datasets and they are merged into a single dataset. The required libraries of python are imported. Then the data present in the dataset is filtered and cleaned. i.e., Extra white spaces are removed, Punctuations are corrected, unnecessary symbols are removed, the uppercase and the lowercase text are corrected.

Then The Countvectorizer algorithm is applied for extracting the features in the dataset and to design a model by generating the test and training sets from the given data.

Naive Bayes Classifier is applied for clustering and testing. The training set is given as input based on this data given message is tested if it is spam or not.

Table 1: This Table describes the sample dataset collected by YouTube source. This contains the comments of some popular youtube videos and is merged into a single dataset that contains a size of 9780 with 5 attributes.

Comment ID	Author	Date	Content	Class
LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11-07	Huh, anyway check out this you[tube] channel:kobayashi 02	1
LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A	adam riyati	2013-11-07	Hey guys check out my new channel and our first vid !	1
LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11-08	just for test i have to say much murdev.com	1
z13jhobxqncu512g22wvzkasxmvvzjaz04	Elnino melendez	2013-11-09	me shaking my ***** in my channel check it !	1
z13hp0bxcuwcnbxxmvjja5	GsMega	2013-11-10	check this out .!>>?	1
z122wfnzgt304chlngpvsmgjff	Bob marley	2013-11-26	Nice song !	0

Table 2 : Mean, Std. deviation values for Accuracy and Loss [Sample size=5, mean of Naive Bayes=90] - Group Statistics

Parameters	Algorithm	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	Naive Bayes	5	90.3800	1.10544	.49437
Loss	Naive Bayes	5	9.8000	.83666	.37417

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007140 |

V. ARCHITECTURE OF DIAGRAM

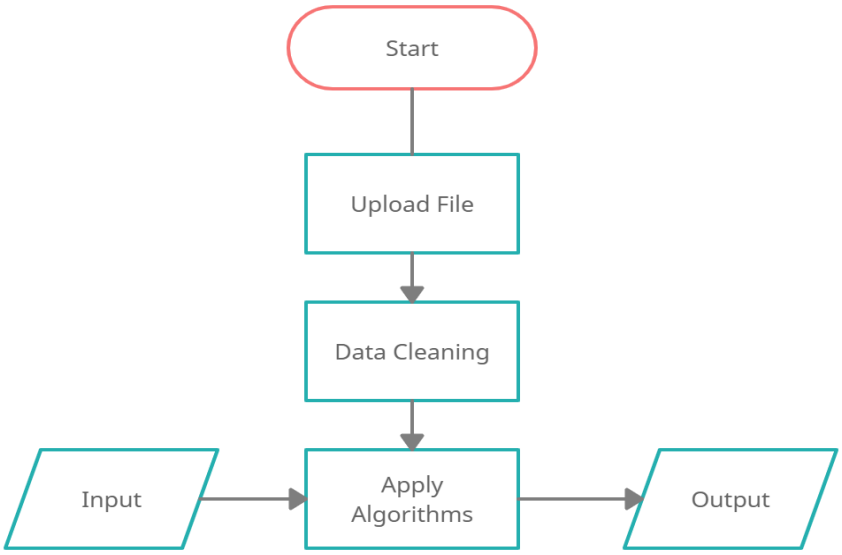

Fig.1:Architecture Diagram

The system design (Architecture Diagram) plays a vital role in defining the working of the system. Initially the process starts with importing the required libraries. Then the Dataset is uploaded, the data is filtered and cleaned by removing extra white spaces, punctuation marks, unnecessary symbols and correcting the uppercase and lowercase text.

```
Feature Extraction using Countvectorizer
        cv=CountVectorizer
        ex=cv.fit_transform(["Great song but check this out", what is this song?"])
        ex.toarray()
        cv.get_feature_names()
        #Extract Feature with CountVectorizer
        corpus=df_x
        cv=CountVectorizer()
        X=cv.fit_transform(corpus)
```

Fig.2: Feature extraction using Countvectorizer classifier

```
Naive Bayes Classifier
        from sklearn.naive_bayes import MultinomialNB
        clf=MultinomialINB()
        clf.fit(X_train,y_train)
        clf.score(X_test,y_test)
```

Fig.3: Pseudocode for Naive Bayes Classifier

STATISTICAL ANALYSIS: In this study, it is observed that this algorithm's countvectorizer and Naive Bayes has given better degree of accuracy and significant results when compared to other algorithms. But in the previous study, they use a small dataset with less attributes and values which may result in the change of accuracy. And also they've used deep learning concepts which work only after posting a message. So, there is a need for a system which automates this process before they post comments.

- The accuracy tables and graphs are built using SPSS tool

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007140 |

- Comparison of Accuracy and Loss percentage of Naive Bayes machine learning algorithm. The mean accuracy of Naive Bayes [91%]

VI. CONCLUSION

In this study, it is observed that this algorithm's countvectorizer and Naive Bayes has given better degree of accuracy and significant results when compared to other algorithms. But in the previous study, they use a small dataset with less attributes and values which may result in the change of accuracy. And also they've used deep learning concepts which work only after posting a message. So, there is a need for a system which automates this process before they post comments.

REFERENCES

[1]Peter Meland Tim Grace, "The NIST Definition of Cloud Computing", NIST, 2010.
[2]Achill Buhl, "Rising Security Challenges in Cloud Computing", in Proc. Of World Congress on Information and correspondence Technologies ,pp. 217-222, Dec.2011.
[3]Srinivasa rao D et al., "Breaking down the Superlative symmetric Cryptosystem Encryption Algorithm", Journal of Global Research in Computer Science, vol. 7, Jul. 2011
[4]TingyuanNyeandTangZhang"AninvestigationofDESandBlowfishencryption algorithm" , in Proc. IEEE Region 10 Conference, pp. 1-4 ,Jan.2009.
[5]JitendraSinghAdametal.,"ModifiedRSAPublicKeyCryptosystemUsingShort Range Natural Number Algorithm" , International Journal of Advanced Research in Computer Science and Software Engineering ,vol. 2,Aug.2012.
[6]Manikandan.G et al., "A changed cryptographic plan improving information", Journal of Theoretical and Applied Information Technology, vol. 35, no.2, Jan.2012.
[7]Niles Maintain and Subhead Bhingarkar, " The examination and Judgment of Nimbus, OpenNebula and Eucalyptus", International Journal of Computational Biology , vol. 3, issue 1, pp 44-47,2012.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details