EMAIL SPAM DETECTION

Dhruv Lunagariya--20BCP153
Dhruv Patel--20BCP152

TABLE OF CONTENT

INTRODUCTION

- spam (junk mails) mails are unwanted or unsolicited messages sent over the internet.
- sometimes these spam mails become illusive.spam mail conains illusive offers which motivates consumers to provide their personal data to hackers.
- the first spam was sent out in 1978 by Gary Thuerk an employee of Now Defunct Digital equipment corp.
- he sent mail to promote a new product.
- this mail went out to about 400 of the 2600 people who had email accounts.

LITERATURE REVIEW

author	dataset size	classification approach	result
Shrivastava(2014)	corpus of 2248 emails with 1346 spam and ham texts	rule based spam detection filter	accuracy-82.7%
Lue et al.(2011)	spam corpus with 4150 spam and 1897 ham mails	rule extraction, optimization, and rule filtering models are used	accuracy-98.5%
Fuad Deb &Hossain(2004)	email corpus with 271 training and 30 test emai text	fuzzy inference system with a set of fuzzy rules	accuracy--90%
Mohammed(2013)	email-1431 rows in dataset	SVM, K-NN, NB	accuracy-85.96%

DATSET DESCRIPTION

- we have chosen our dataset from kaggle.
- dataset contains colums and 5171 rows.

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0

```
data['label_num'].value_counts()
```

```
0    3672
1    1499
Name: label_num, dtype: int64
```

- the output feature "label_num" has 2 classes. class-0 (ham mail), class-1(spam mail)
- class-0--> 3672 samples.
- class-1-->1499 samples.

DATSET DESCRIPTION

```
data.isna().sum()
```

```
Unnamed: 0    0
label         0
text          0
label_num     0
dtype: int64
```

dataset does not contains any NULL values.

	Unnamed: 0	label_num
count	5171.000000	5171.000000
mean	2585.000000	0.289886
std	1492.883452	0.453753
min	0.000000	0.000000
25%	1292.500000	0.000000
50%	2585.000000	0.000000
75%	3877.500000	1.000000
max	5170.000000	1.000000

some statistical information.

METHOD

- we have proposed Machine Learning based method.

- we are going to apply different classification algorithms like logistic regression, naive Bayes, random forest etc.

1

step-1--> data preprocessing

- this step contains data preprocessing such as removal of NULL values, filling of emply values, feature selection, feature extraction etc.

2

step-2--> model building(training)

- this step includes model building on preprocessed dataset using different classification algorithms.

3

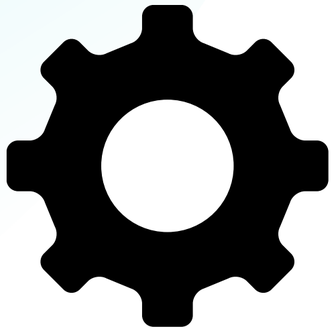
step-3--> model evaluation(testing)

- this step includes model evaluation by some random values as input.

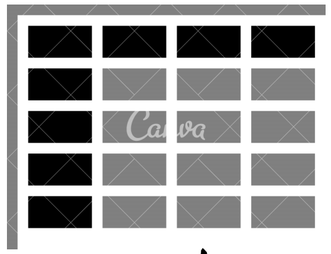
TRAINING PHASE

raw data

feature selection
NULL value removal
etc.



preprocessing



processed data

classification algorithms

training

trained model

TESTING PHASE

new data

trained model

output

ALGORITHM

Logistic regression

- Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class.
- Statistical Model
- Accuracy : 98.45%

Naive Bayes

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems.
- probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Accuracy : 91.68%

CONCLUSION

- We are able to classify the email as spam or not.
- Successfully apply Machine learning algorithm such as Logistic regression and Naive Bayes.
- Logistic Regression demonstrated exceptional performance with an accuracy rate of 98%.
- In Naive Bayes, slightly less accurate at 91%.

REFERENCES

1.https://www.researchgate.net/publication/357972513_A_systematic_literature_review_on_spam_content_detection_and_classification

2.https://www.kaggle.com/datasets/venky73/spam-mails-dataset

3.https://www.techtarget.com/searchsecurity/definition/spam

THANK YOU