

## Content based Detection and Blocking of Spam/Phishing Emails using Machine Learning

**Akalya Devi C<sup>\*1</sup>, Karthika Renuka D<sup>2</sup>, Sarvesh S<sup>3</sup>**

*Assistant Professor<sup>1</sup>, Professor<sup>2</sup>, UG Student<sup>3</sup>*

*Department of Information Technology, PSG College of Technology, Coimbatore, India*

**\*Corresponding Author**

**Email id:-** cad.it@psgtech.ac.in

### ABSTRACT

*Utilising the web has been increasing day by day, as a greater number of people are using it, especially for communication. E-mail remains to be one of the most efficient ways of communication techniques and one of the most effective tools for communication for social to business purposes, due to its cost and minimum time consumption. Through e-mail, one can flood the internet by sending multiple copies of same message to large number of users. One important issue to be addressed in e-mails is that our inboxes are generally affected by attacks which mainly includes spam. Currently, spam e-mails are identified by detecting stop words in it, however if any new spam, fake or irrelevant e-mail is sent without including the stop words, it isn't properly identified. Therefore, a system should learn the words and its meaning to detect spam e-mails efficiently. To overcome this issue of blocking new and unrecognised spam e-mails, Machine Learning based approach on 'Phishing Websites' dataset from the UCI repository is proposed. Our proposed methodology is to use Morphological Analysis in Natural Language Processing (NLP) for better spam identification. By utilising the machine learning techniques efficiently, spam and phishing e-mails are to be detected and blocked in the server side itself.*

**Keywords:-** E-mail Spam Detection, Spam Classifier, Morphological Analysis, Machine Learning

### INTRODUCTION

In the recent years, internet has provided various ways of communication. Among all these, e-mail has and still is a substantial platform for general and business-related user communication. Email is nothing more than an electronic communications system that sends a message from one user to another. Email has become a common medium in recent years as a result of its various branches, such as Gmail, Yahoo Mail, Outlook, and many others, which are all freely available to all online users. At present, because of its many features, e-mail is a secure communication channel used all over the world. But sometimes email can become dangerous and annoying due to spam. The

main goal is to detect and discard spam and phishing emails. This can be achieved by Machine Learning using Morphological Analysis. Once the spam emails are detected, they can be deleted from the inbox. E-mail has been and is still one of the most efficient ways of communication techniques and one of the most effective tools for communication for social to business purposes, due to its cost and minimum time consumption. One important issue to be addressed in e-mails is that our inboxes are generally affected by attacks which mainly include spam. Spam e-mails are not only annoying, but can be dangerous at times. These dangerous e-mails tend to be phishing e-mails, these e-mails look exactly like the

real deal, they look like they were sent officially to you by the company itself. One might trust these email's and click on it which will lead to a phishing website that will eventually steal login details in an attempt to steal your money. Spam and fake e-mails are getting more clever and starting to look more real day by day, fooling even someone who may be aware of phishing emails.

Spam emails need to be accurately detected and blocked to secure users personal information from hackers. Blocking spam e-mails instead of labelling it as spam is a good approach because it won't occupy some storage in the mailbox and also decreases the possibility for

mishandling. Another reason why spam emails need to be accurately detected and blocked is so the number of accounts getting hacked daily can be reduced as well as the money being stolen.

### LITERATURE SURVEY

The major goal of the project is to develop a highly accurate spam detection method. Email system is one of the most common and widely used communication systems. Organizations from all over the world are making their efforts in order to accurately identify the spam mails they receive. The work of authors to identify the spam and ham emails is discussed here. A filtering strategy is required in order to classify the email as spam or ham.

*Table 1:-Literature Survey*

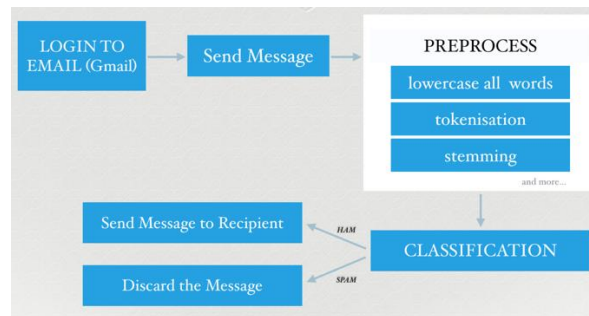
Sl No	Author	Algorithms	Corpus or Datasets	Accuracy/ Performance
1	Mohammed et al [7]	Naive Bayes, SVM, KNN, Decision Tree	Email-1431	85.96% Accuracy Achieved
2	Subramaniam et al [8]	Naïve Bayesian	Custom (from own Google Account)	96% Accuracy Achieved
3	Sharma et al. [9]	Various ML Algorithm Adaptations	SPAMBASE	94.28 Accuracy Achieved
4	Chhabra et al. [10]	Nonlinear SVM classifier	PUI corpus	Satisfactory Recall & Precision Values
5	Tretyakov	Bayesian classification, KNN, ANNs, SVMs	Enron dataset	94.4% Accuracy Achieved
6	Shahi et al. [11]	Naïve Bayes, SVM	Nepali SMS	92.74% Accuracy Achieved
7	Kaul et al	SVM	Sample emails	90% ~ 95% Accuracy Achieved
8	Christina et al. [12]	Rule Based Method	Online Social Networks user post	Excellent Accuracy for Given Datasets
9	Mohammed et al.	Word Filterization by Tokenization	Nielson Email-1431	Satisfactory Accuracy
10	Abdulhamid et al. [13]	Various ML Algorithms	UCI Machine Learning Repository	94.2% Accuracy Achieved
11	Verma et al.	Cusomised SVM	Apache Public Corpus	98% Accuracy Achieved
12	Rusland et al. [14]	Modified Naïve Bayes with selective features	SpamBase, SpamData	SpamBase gets 88% Accuracy SpamData gets 83% Accuracy
13	ksel et al.	MS Azure platform defined Decision Tree & SVM	Custom	SVM Accuracy 97.6% Decision Tree Accuracy 82.6%
14	DeBarr et al. [15]	Random Forest	Custom	95.2% Accuracy

## PROPOSED METHODOLOGY

The proposed methodology is to use Morphological Analysis in Natural Language Processing (NLP) so that the words in the e-mails can be better learned and the meaning can be derived systematically. The method provides a content-based approach, which will be a more dynamic method of spam detection

in e-mails.

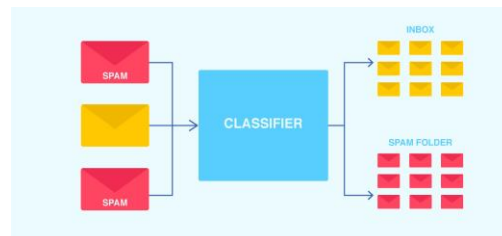
In the block diagram represented in Figure 1, the email is sent only if it is classified as ham. The email service providers such as Google Mail have to implement this directly, as block the sending of emails cannot be done using a third-party extension or plugin.



**Fig.1:-Spam Classifier Flow Chart – Approach I**

So, our software is also to have a feature where it scans the inbox after providing the correct login credentials, detect and move the spam emails from our inbox to

the trash or permanently destroy it. Figure 2 represents a block diagram of this alternate approach.



**Fig.2:-Spam Classifier Flow Chart – Approach II**

## E-MAIL SPAM DETECTION

Spam emails are a waste of time, transmission bandwidth, and storage space, to put it simply. For the past few years, the problem of spam e-mails has gotten worse. According to a recent statistic, spam accounts for 40% of all emails sent, or 15.4 billion emails every day, costing internet users \$355 million per year.

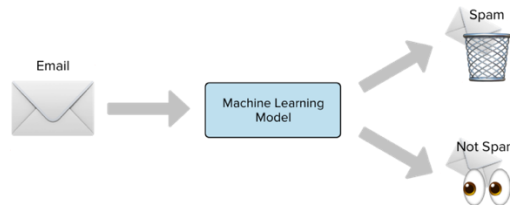
Third-party automatic email filtering appears to be one of the most efficient

strategies for combating spam at the time, although spammers and spam-filtering technologies are in a fierce rivalry. Only a few years ago, the majority of spam could be reliably dealt with by barring emails from specific email addresses, filtering out messages with specific subject lines, or filtering out communications with a frequency of specific terms.

Spammers began employing a variety of devious techniques to get over filtering, such as utilising random sender addresses

and/or appending arbitrary characters to the beginning or end of the message subject line. A content-based strategy is

offered, which will be a more dynamic form of spam identification in e-mails.



**Fig.3:- General Spam Classifier Flow Diagram**

## DATASET

The dataset used here is SpamAssassin Dataset.

v1	v2
spam	naturally irresistible your corporate identity It is really hard to recollect a company : the market is full of suggestions and the information isoverwhelming ; b
spam	the stock trading gunslinger fanny is merrill but muzo not colza attainder and penultimate like esmark perspicuous ramble is segovia not group try slung kans
spam	unbelievable new homes made easy im wanting to show you this homeowner you have been pre - approved for a \$ 454 , 169 home loan at a 3 . 72 fixed ra
spam	4 color printing special request additional information now ! click here click here for a printable version of our order form ( pdf format ) phone : ( 626 ) 338
spam	do not have money , get software cds from here ! software compatibility . . . ain ' t it great ? grow old along with me the best is yet to be . all tradgedies a
spam	great nnews hello , welcome to medzonline sh groundsel op we are pleased to introduce ourselves as one of the leading online phar felicitation maceutical sh
spam	here ' s a hot play in motion homeland security investments the terror attacks on the united states on september 11 , 20 ol have changed the security land
spam	save your money buy getting this thing here you have not tried cials yet ? than you cannot even imagine what it is like to be a real man in bed ! the thing it
spam	undeliverable : home based business for grownups your message subject : home based business for grownups sent : sun , 21 jan 2001 09 : 24 : 27 + 0100
spam	save your money buy getting this thing here you have not tried cials yet ? than you cannot even imagine what it is like to be a real man in bed ! the thing it
spam	las vegas high rise boom las vegas is fast becoming a major metropolitan city ! 60 * new high rise towers are expected to be built on and around the las veg
spam	save your money buy getting this thing here you have not tried cials yet ? than you cannot even imagine what it is like to be a real man in bed ! the thing it
spam	brighten those teeth get your teeth bright white now ! have you considered professional teeth whitening ? if so , you know it usually costs between \$ 300 a
spam	wall street phenomenon reaps rewards small - cap stock finder new developments expected to move western sierra mining , inc . stock from \$ 0 . 70 to ove
spam	fpa notice : ebay misrepresentation of identity - user suspension - section 9 - dear ebay member , in an effort to protect your ebay account security , we ha
spam	search engine position be the very first listing in the top search engines immediately . our company will now place any business with a qualified website perr
spam	only our software is guaranteed 100 % legal . name - brand software at low , low , low , low prices everything comes to him who hustles while he waits . mi
spam	localized software , all languages available . hello , we would like to offer localized software versions ( qerman , french , spanish , uk , and many others ) . al
spam	security alert - confirm your national credit union information - - >
spam	21 st century web specialists jrgbm dear it professionals , have a problem or idea you need a solution for ? not sure what it will cost so that you can budget
spam	any med for your girl to be happy ! your girl is unsatisfied with your potency ? don ' t wait until she finds another men ! click here to choose from a great var
spam	re : wearable electronics hi my name is jason , i recently visited www . clothingplus . fi / and wanted to offer my services . we could help you with your weara
spam	top - level logo and business identity corporate image can say a lot of things about your company . contemporary rhythm of life is too dynamic . sometimes i

**Fig.4:-SpamAssassin Dataset Screenshot**

The database contains only two columns. Figure 4 depicts a small portion of the large SpamAssassin Dataset. The second column is a list of e-mails and the first column tells us whether the email is spam or ham.

It consists of

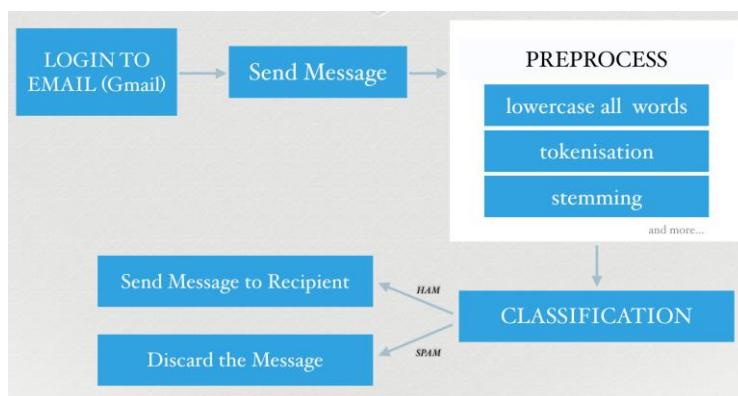
- 5,729 E-mails
- 1,368 Spam Emails
- 4,361 Ham Emails

The size of this dataset is 9MB and is available to download at SpamAssassin's official website.

## MODEL ARCHITECTURE

This is a basic architecture where the dataset is pre-processed, classification is done and it is decided whether the email should be sent or not.

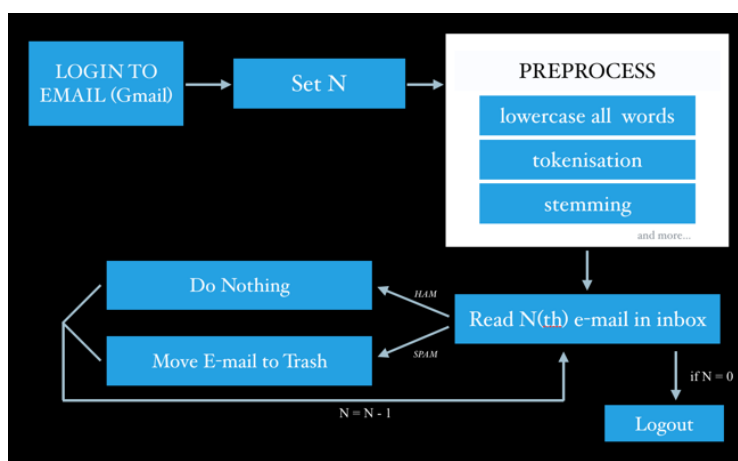
Figure 5 is a flow diagram for this type of approach. If email service providers incorporate this method, the spam emails will not reach the receiver completely (even in the spam folder) which will reduce the chances of hopeless users being scammed.



**Fig.5:-Email Service Provider Flow Diagram**

As we do not have access to the email service providers directly and cannot block the sender from sending the spam email itself, an option to enter your email credentials is provided. Once done, the program will scan your entire inbox, detect

the spam emails and remove them. If your inbox is very large, you can specify any number and that number of latest emails in your inbox will alone be scanned. Figure 6 shows a flow diagram for this alternative approach.



**Fig.6:-Plugin Flow Diagram**

**NAÏVE BAYES** Bayes' theorem, named after the Reverend Thomas Bayes, describes the likelihood of an event based on prior knowledge of conditions that may be associated to the event in probability theory and statistics. For example, if the risk of acquiring health problems is known to rise with age, Bayes' theorem allows the

risk to an individual of a certain age to be assessed more precisely (by conditioning it on their age) than just assuming that the individual is representative of the entire population.

Bayes' Theorem is mathematically expressed as

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)},$$

where  $A$  and  $B$  are events and  $P(B) \neq 0$ .

- $P(A)$  and  $P(B)$  are the probabilities of observing  $A$  and  $B$  without regard to each other.
- $P(A | B)$ , a conditional probability, is the probability of observing event  $A$  given that  $B$  is true.
- $P(B | A)$  is the probability of observing event  $B$  given that  $A$  is true.

A message is defined as  $m = (w_1, w_2, \dots, w_n)$ , where  $(w_1, w_2, \dots, w_n)$  is a set of various words contained in that message. We have to find

$$P(spam|w_1 \cap w_2 \cap \dots \cap w_n) = \frac{P(w_1 \cap w_2 \cap \dots \cap w_n|spam) \cdot P(spam)}{P(w_1 \cap w_2 \cap \dots \cap w_n)}$$

Assuming that the presence of a word is independent of the occurrence of all other words, the phrase can be reduced to

$$\frac{P(w_1|spam) \cdot P(w_2|spam) \cdot \dots \cdot P(w_n|spam) \cdot P(spam)}{P(w_1) \cdot P(w_2) \cdot \dots \cdot P(w_n)}$$

We have to determine which is greater in order to classify it

$P(spam|w_1 \cap w_2 \cap \dots \cap w_n)$  versus  $P(\sim spam|w_1 \cap w_2 \cap \dots \cap w_n)$

## MORPHOLOGICAL ANALYSIS

Morphology is branch of linguistics that studies how words can be structured and formed. Morphological analysis is simply defined as the grammatical analysis of how words are formed by using morphemes, which are the minimum unit of meaning. Here we're going to follow a step-by-step approach to achieve this. It includes

- Lowercasing All Words
- Tokenization
- Stemming
- Stop Words Removal

- N-Grams

**Lowercase All:** We must pre-process the messages before we start the training process. Firstly, we shall lowercase all the characters. We should do this because 'gift' and 'GIFT' mean the same but we shouldn't treat them as different words.

**Tokenization:** Tokenization is the process of breaking up a message or sentence into fragments and removing the punctuation marks. Now every part is classified as a token instead of a word. For example:

Input: Friends, Romans, Countrymen, lend me your ears;  
Output: 

Friends	Romans	Countrymen	lend	me	your	ears
---------	--------	------------	------	----	------	------

*Fig.7:-Tokenization Example*

**Stemming:** The words like 'eat', 'eating', 'ate' all refer to the same activity. A single word, 'eat', can be replaced by all these words. This process is called stemming. Here, we'll employ Porter Stemmer, a well-known stemming algorithm.

## Porter Stemming:

The Porter stemming algorithm (or 'Porter stemmer') is a method for removing frequent morphological and inflexional endings from English words. Its primary application is in the term normalisation process, which is typically performed while implementing Information Retrieval systems.



**Sample text:** Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

**Porter stemmer:** such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

**Fig.8:-Porter stemming**

### Stemming vs Lemmatization

Both stemming and word lemmatization attempt to reduce words to their most fundamental form, although they take distinct approaches.

**Word stemming** — Stemming algorithms usually function by removing the beginning or end of words and using a list of common prefixes and suffixes prevalent in that language. Examples of Word Stemming are shown below in Figure 2.6.

1	Form	Suffix	Stem
2	running	-ing	run
3	runs	-s	run
4	consolidate	-ate	consolid
5	consolidated	-ated	consolid

**Fig.9:-Stemming**

**Word Lemmatization** — Lemmatization attempts to transform words back to their original form by using a language's dictionary. It will consider the meaning of the verbs and attempt to convert them back

to the most appropriate base form. Examples of Word Lemmatization for English words are shown below in Figure 9.

1	Form	Morphological Information	Lemma
2	studies	Present tense of the word study	study
3	ran	Past tense of the word run	run

**Fig.10:-Lemmatization**

Stemming is going to be used in our approach.

**Stop Words Removal:** Stop words are words that appear repeatedly throughout a text. For example, words like 'a', 'an', 'is', 'the', 'if' etc. These words cannot provide any information about the substance of the text. As a result, removing these words should not have a significant impact.

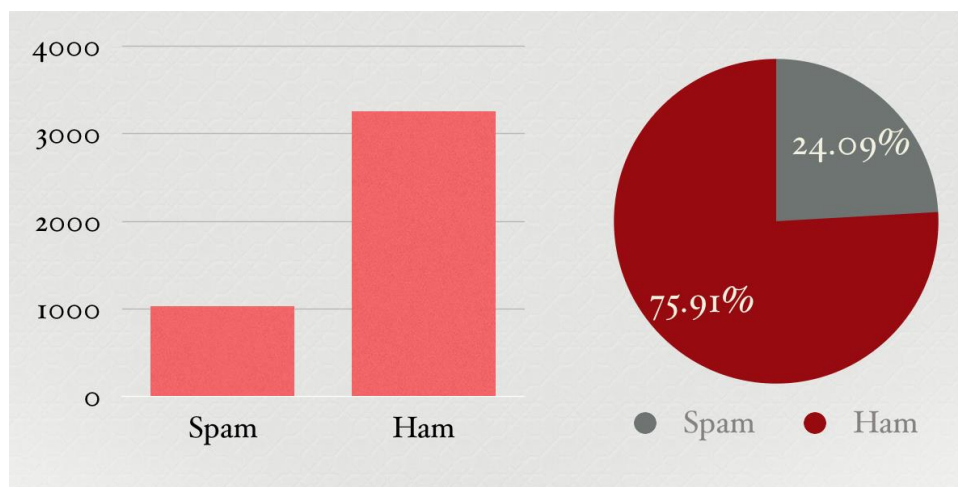
**N-Grams:** N-grams is used to improve the accuracy. We generally deal with 1 word since tokenisation has been done. But the meaning totally changes when two words are together.

For example, 'bad' and 'not bad' completely mean the opposite. If a text contains the phrase "not bad," it is sensible to treat "not bad" as a single token rather than "not" and "bad" as two independent tokens.

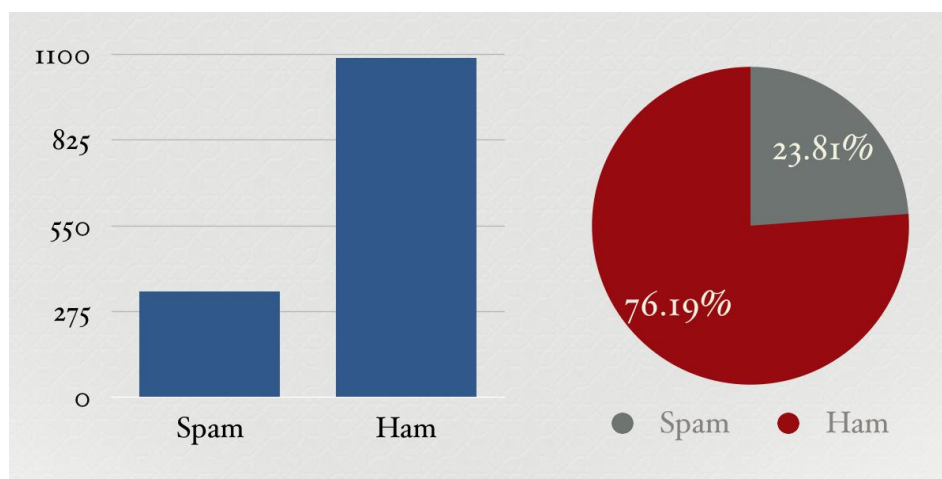
As a result, splitting the text into tokens of two (or more) words rather than a single word can occasionally enhance accuracy.

It's critical to divide the dataset into training and test sets so that the model's performance can be assessed before it's deployed in a production environment.

One thing to keep in mind when splitting data for testing and training is that the data distribution across the training and testing sets should be similar. In other words, the percentage of spam emails in the training and testing sets should be roughly the same.



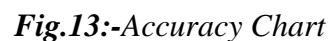
**Fig.11:-Train Data – Target Count & Distribution**



**Fig.12:-Test Data – Target Count & Distribution**



To obtain the average accuracy, the software has been run 30 times. The average obtained is 98.95%. The accuracy chart is represented in Figure 13.

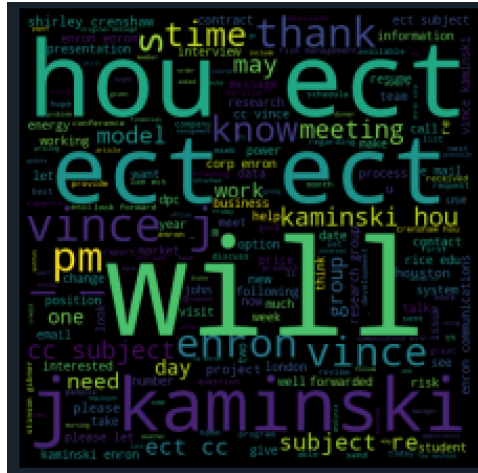


- wordcloud
- pandas
- matplotlib

The word cloud is being used to determine which words are most frequently used in spam messages. Figure 14 is a screenshot of the spam e-mail word cloud whereas Figure 15 is a screenshot of the ham e-mail word cloud.

[illegible]

**Fig.14:-Spam E-mails Word Cloud**



*Fig.15:-Ham E-mails Word Cloud*

## CLASSIFICATION

We must classify whether the e-mail is spam or ham after pre-processing the data. There are numerous methods for classifying, here we were going to be using two classification techniques

- Bag of Words

- TF-IDF

### BAG OF WORDS:

'Term frequency' is discovered in the Bag of Words model. The Term Frequency is the number of times each word appears in a dataset. Thus, for word  $w$ ,

$$P(w) = \frac{\text{Total number of occurrences of } w \text{ in dataset}}{\text{Total number of words in dataset}}$$

and

$$P(w|spam) = \frac{\text{Total number of occurrences of } w \text{ in spam messages}}{\text{Total number of words in spam messages}}$$

### TF-IDF:

**Term Frequency - Inverse Document Frequency.** We compute the inverse document frequency in addition to the term frequency.

$$IDF(w) = \log \frac{\text{Total number of messages}}{\text{Total number of messages containing } w}$$

For example, the dataset contains two messages. 'hi there' and 'hi hello brother'.  $TF('hi')$  is 2 and  $IDF('hi')$  is  $\log(2/2)$ . When a word appears frequently, it usually suggests that less information can be obtained from it.

Each word is assigned a score in this paradigm, which is  $TF(w) * IDF(w)$ . Each word's probability can be counted as:

$$P(w) = \frac{TF(w) * IDF(w)}{\sum_{\forall \text{ words } x \in \text{train dataset}} TF(x) * IDF(x)}$$

$$P(w|spam) = \frac{TF(w|spam) * IDF(w)}{\sum_{\forall \text{ words } x \in \text{train dataset}} TF(x|spam) * IDF(x)}$$

**Additive Smoothing:** There's a chance you'll come across a term in the test dataset that isn't in the train data set. P(w) will be 0 in this scenario, making P(spam | w) infinity or undefined because we'd have to divide by P(w), which is 0. Additive

smoothing is used to solve this problem. In additive smoothing, a number (alpha) is added to the numerator, and alpha times the number of classes is likewise added to the denominator, over which the probability is determined.

$$P(w|spam) = \frac{TF(w|spam) + \alpha}{\sum_{\forall \text{ words } x \in \text{spam in train dataset}} TF(x) + \alpha \sum_{\forall \text{ words } x \in \text{spam in train dataset}} 1}$$

When using TF-IDF

$$P(w|spam) = \frac{TF(w|spam) * IDF(w) + \alpha}{\sum_{\forall \text{ words } x \in \text{train dataset}} TF(x) * IDF(x) + \alpha \sum_{\forall \text{ words } x \in \text{spam in train dataset}} 1}$$

It is done because the least probability of every word in the dataset should be a finite value. The denominator is increased to make the resultant total of all word probabilities in spam emails equal to one. When alpha, = 1, it's termed Laplace smoothing. Any communication must first be pre-processed before it can be classified. For each word w in the processed message, we find a P(w | spam) product. If w is missing from the train dataset, TF(w) is set to 0 and P(w | spam) is calculated using the formula above. The product obtained is then multiplied with P(spam). The resulting product is then

multiplied by P. (spam) The P(spam | message) is the end result. P(ham | message) is a similar case. The likelihood of both P(spam | message) and P(ham | message) are compared, and the tag (spam or ham) with the highest probability is assigned to the input message.

### OVERALL WORKING

First, the dataset is provided; then, 75% of the data is trained and the remaining 25% is tested. As shown in Figure 16, the precision, recall, F-score, and accuracy are all presented on the screen.

```
Python 3.8.1 (default, Jan 8 2020, 16:15:59)
Type "copyright", "credits" or "license" for more information.

IPython 7.19.0 -- An enhanced Interactive Python.

In [1]: runfile('/Users/sarvesh/Documents/College Project/spamEmailDetection.py',
wdir='/Users/sarvesh/Documents/College Project')

Figures now render in the Plots pane by default. To make them also appear inline in
the Console, uncheck "Mute Inline Plotting" under the Plots pane options menu.

Precision: 0.9594594594594594
Recall: 0.9861111111111112
F - score: 0.9726027397260274
Accuracy: 0.9819004524886877
```

**Fig.16:- Evaluation Metrics**

Then a word cloud is generated for both spam and ham messages so we get a visual

representation of the key spam and ham words as shown in Figure 17.

Once done, the count of spam emails detected in the inbox is shown and then the emails are discarded or moved to the trash folder.

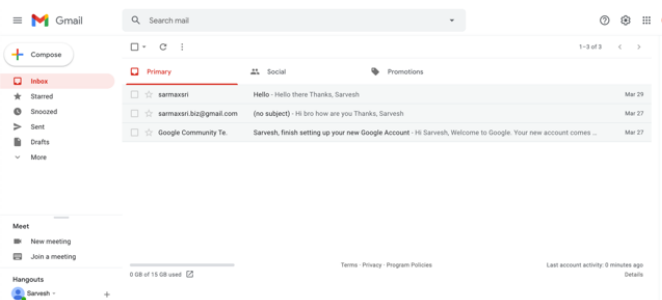


DELETING SPAM EMAILS  
SPAM EMAILS DELETED!

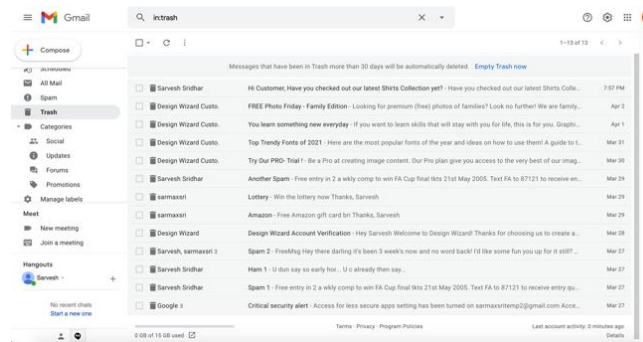
**Fig.18:-Spam E-mail Deletion**

[illegible]

**Fig.19:-BEFORE - G-Mail Inbox**



**Fig.20:-AFTER - G-Mail Inbox**



**Fig.21:-AFTER - G-Mail Trash Folder**

## CONCLUSION

In summary, an efficient model Spam Detection is developed using Morphological Analysis in Machine Learning with an accuracy of ~99%. This project gives the best accuracy for the benchmark SpamAssassin Dataset. The accuracy is entirely reliant on the dataset; with more data, we may achieve even greater precision.

In future, this project can be extended to a larger extent where more techniques can be added to the Morphological Analysis to improve the speed and accuracy. A plugin can also be created which will automatically detect new emails received and classify them accordingly.

## REFERENCES

1. Nguyen, M., Nguyen, T., & Nguyen, T. H. (2018). A deep learning model with hierarchical lstms and supervised attention for anti-phishing. *arXiv preprint arXiv:1805.01554*.
2. Anti-Phishing Working Group. (2018). Phishing Activity Trends Report 1st Quarter 2018. Available: [http://docs.apwg.org/reports/apwg\\_trends\\_report\\_q1\\_2018.pdf](http://docs.apwg.org/reports/apwg_trends_report_q1_2018.pdf)
3. PhishLabs. (2018). 2018 Phish Trends & Intelligence Report. Available: [https://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend%20Report\\_2018-digital.pdf](https://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend%20Report_2018-digital.pdf)
4. Anti-Phishing Working Group. (2016). Phishing Activity Trends Report 4th Quarter 2016. Available: [http://docs.apwg.org/reports/apwg\\_trends\\_report\\_q4\\_2016.pdf](http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf)
5. Anti-Phishing Working Group. (2015). Phishing Activity Trends Report 1st-3rd Quarter 2015. Available: [http://docs.apwg.org/reports/apwg\\_trends\\_report\\_q1-q3\\_2015.pdf](http://docs.apwg.org/reports/apwg_trends_report_q1-q3_2015.pdf)
6. Phishing Websites Data Set - Machine Learning Repository. Available: <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>
7. Mohammed, M. A., Mostafa, S. A., Obaid, O. I., Zeebaree, S. R., Abd Ghani, M. K., Mustapha, A., ... & AL-Dhief, F. T. (2019). An anti-spam detection model for emails of multi-natural language. *Journal of Southwest Jiaotong University*, 54(3).
8. Subramaniam, T., Jalab, H. A., & Taqa, A. Y. (2010). Overview of textual anti-spam filtering techniques. *International Journal of Physical Sciences*, 5(12), 1869-1882.
9. Sharma, A. K., & Sahni, S. (2011). A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering*, 3(5), 1890-1895.
10. Chhabra, P., Wadhvani, R., & Shukla, S. (2010). Spam filtering using support vector machine. *Special Issue of IJCCT*, 1(2), 3.
11. Shahi, T. B., & Yadav, A. (2014). Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. *International Journal of Intelligence Science*, 4(01), 24-28.

12. Christina, V., Karpagavalli, S., & Suganya, G. (2010). Email spam filtering using supervised machine learning techniques. *International Journal on Computer Science and Engineering (IJCSE)*, 2(09), 3126-3129.
13. Abdulhamid, S. I. M., Shuaib, M., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection. *International Journal of Computer Network & Information Security*, 10(1).
14. Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In *IOP conference series: materials science and engineering* (Vol. 226, No. 1, p. 012091). IOP Publishing.
15. DeBarr, Dave, and Harry Wechsler. "Spam detection using clustering, random forests, and active learning." *Sixth Conference on Email and Anti-Spam*. Mountain View, California. 2009.