Chap4#5#1: Machine Learning for Anomaly-based Spam Detection

April 17, 2023

भारतीय प्रौद्योगिकी
संस्थान जम्मू
INDIAN INSTITUTE OF
TECHNOLOGY JAMMU

Devesh C Jinwala,
Professor, SVNIT and Adjunct Prof., CSE, IIT Jammu

Department of Computer Science and Engineering,
Sardar Vallabhhai National Institute of Technology, SURAT

Topics to study in Chapter 4

- Machine learning for Anomaly Detection: Definition of an anomaly. Types of Anomalies or outliers in machine learning. Motivation for machine learning for anomaly detection.
  Data Visualization. Supervised, Unsupervised and Semi-supervised Learning methods for Anomaly Detection.
  Applications of Anomaly Detection: Intrusion detection, Fraud detection, Health monitoring, Defect detection, and lastly Spam detection. Intrusion Detection with Heuristics.Goodness-of-fit. Host Intrusion Detection. Network Intrusion Detection. Web Application Intrusion Detection.
  Overview of Machine learning Approaches for Anomaly Detection: Distance-based, Clustering-based and Model-based Approaches. Algorithms for Distance and Density-based approaches, Rank-based approaches, Ensemble Methods Algorithms for Time Series Data. Deep Learning for Anomaly Detection. Behavioural-based Anomaly Detection [8 hours]

Topics in Handouts#1, #2, #3

1. Email Basics, Spam Detection Basics, and ML-based Spam Detection Basics
   - Background & Basics: Email architecture
   - Basic Spam Detection Mechanism
   - Basic Categories of Spam Detection Mechanisms
   - Spam Detection: Historical aspects and impact of Spam
   - Spam Detection: How Gmail, Yahoo and Outlook ... work
   - Basic ML-based Spam Detection Architecture, Building Blocks
   - Categories Spam Detection Mechanisms
   - Basic ML-based Spam Filtering process
   - Performance evaluation measures
2. Classical ML-based Spam Filtering
3. Deep Learning-based Spam Filtering

Background & Basics: Email architecture

Background: Typical Email architecture

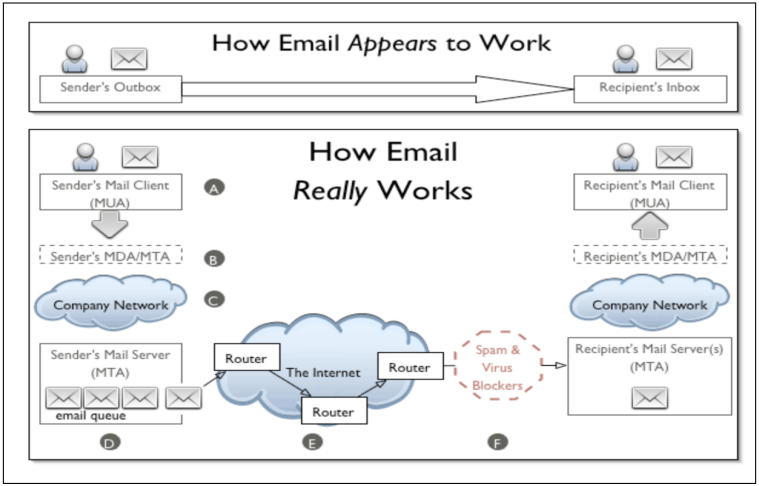


Figure: Basic Email Architecture[1]

[1]:Source: https://www.oasis-open.org/khelp/kmlm/user_help/html/how_email_works.html

Background: Typical Email architecture...

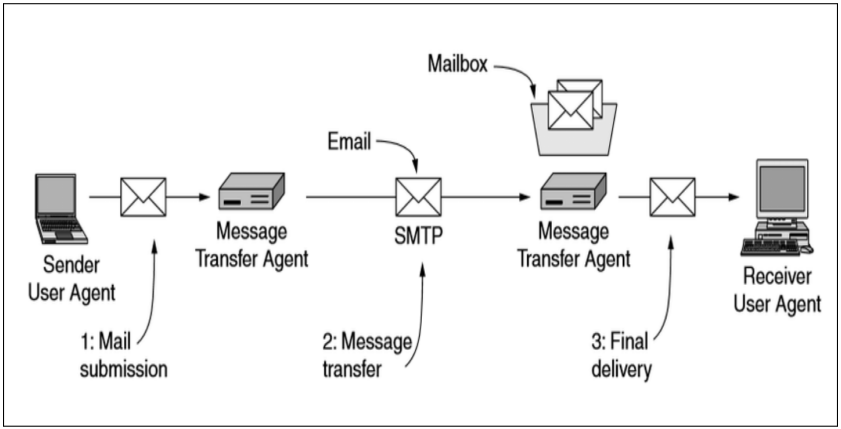


Figure: Basic Email Architecture[1]

[1]:Source: M. Jameel, Noor et al (2017). An Online Content Based Email Attachments Retrieval System. Kurdistan Journal for Applied Research

Background: Typical Email architecture...

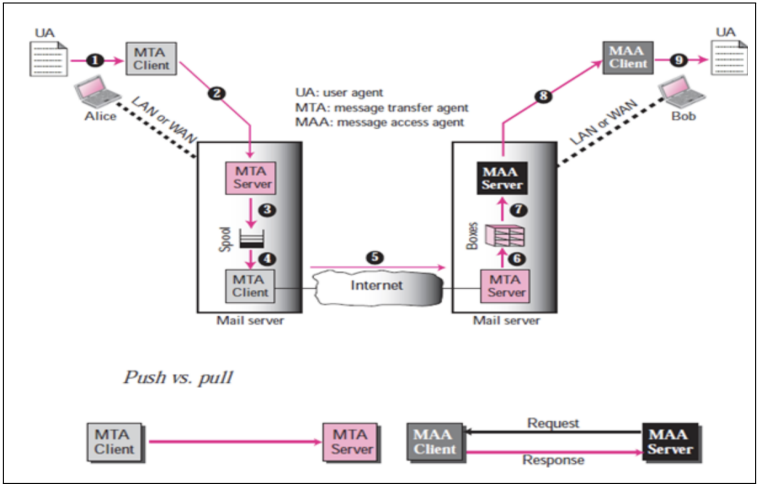


Figure: Basic Email Architecture[1]

[1]:Source: Forouzan

Basic Spam Detection Mechanism

Fundamental Spam Detection approaches

Fundamental Spam Detection may work at either of the two levels

- an individual level

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Fundamental Spam Detection approaches

Fundamental Spam Detection may work at either of the two levels

- an individual level
- an enterprise level

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Fundamental Spam Detection approaches

Fundamental Spam Detection may work at either of the two levels

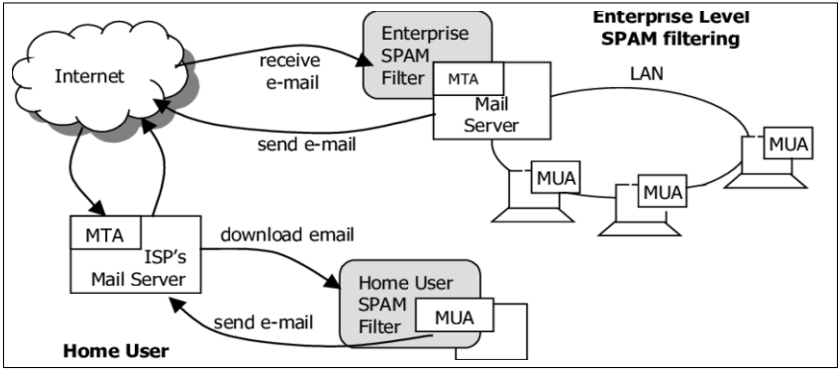
- an individual level
- an enterprise level

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Fundamental Spam Detection approaches

Fundamental Spam Detection may work at either of the two levels

- an individual level
- an enterprise level



[1]

Figure: Spam Filtering Mechanism

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Basic Categories of Spam Detection Mechanisms

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.
  - obviously, the main drawback is that a client or some other entity, such as a software vendor, must maintain and update the set of rules on an ongoing basis.

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.
  - obviously, the main drawback is that a client or some other entity, such as a software vendor, must maintain and update the set of rules on an ongoing basis.
- Machine learning-based approach.

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.
  - obviously, the main drawback is that a client or some other entity, such as a software vendor, must maintain and update the set of rules on an ongoing basis.
- Machine learning-based approach.
  - does not require pre-defined rules, but instead, it requires messages which have been pre-classified successfully.

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.
  - obviously, the main drawback is that a client or some other entity, such as a software vendor, must maintain and update the set of rules on an ongoing basis.
- Machine learning-based approach.
  - does not require pre-defined rules, but instead, it requires messages which have been pre-classified successfully.

Spam Detection Fundamental Approaches

Filtering of email Spam has three primary approaches viz.

- Knowledge engineering-based approach
  - objective is to set up a knowledge-based system using predefined rules to decide if an incoming message is valid or not.
  - obviously, the main drawback is that a client or some other entity, such as a software vendor, must maintain and update the set of rules on an ongoing basis.
- Machine learning-based approach.
  - does not require pre-defined rules, but instead, it requires messages which have been pre-classified successfully.

Our focus shall be on the latter, but vital to learn about what are different Spam detection approaches of interest - to understand their basic characteristics.

Machine Learning-based Spam Detection

Conventional Machine Learning-based approaches

- do not require pre-defined rules - as in case of conventional Knowledge-based approach - but instead, it requires messages which have been pre-classified successfully.

Machine Learning-based Spam Detection

Conventional Machine Learning-based approaches

- do not require pre-defined rules - as in case of conventional Knowledge-based approach - but instead, it requires messages which have been pre-classified successfully.
- such messages allow sample messages to construct the training dataset used to fit the model's unique learning algorithm.

Machine Learning-based Spam Detection

Conventional Machine Learning-based approaches

- do not require pre-defined rules - as in case of conventional Knowledge-based approach - but instead, it requires messages which have been pre-classified successfully.
- such messages allow sample messages to construct the training dataset used to fit the model's unique learning algorithm.
- hence, the classification of Spam emails can adopt an ML approach for classification for learning from the input data and a program uses the learning to classify new observations.

Machine Learning-based Spam Detection

Conventional Machine Learning-based approaches

- do not require pre-defined rules - as in case of conventional Knowledge-based approach - but instead, it requires messages which have been pre-classified successfully.
- such messages allow sample messages to construct the training dataset used to fit the model's unique learning algorithm.
- hence, the classification of Spam emails can adopt an ML approach for classification for learning from the input data and a program uses the learning to classify new observations.
- algorithms such as Support Vector Machines and Naïve Bayes have been investigated on their effectiveness to successfully detect and filter Spam emails

Machine Learning-based Spam Detection

Conventional Machine Learning-based approaches

- do not require pre-defined rules - as in case of conventional Knowledge-based approach - but instead, it requires messages which have been pre-classified successfully.
- such messages allow sample messages to construct the training dataset used to fit the model's unique learning algorithm.
- hence, the classification of Spam emails can adopt an ML approach for classification for learning from the input data and a program uses the learning to classify new observations.
- algorithms such as Support Vector Machines and Naïve Bayes have been investigated on their effectiveness to successfully detect and filter Spam emails
- drawback is it requires feature extraction and the associated processes.

Self-reading & preparation topics start...

Spam Detection: Historical aspects and impact of Spam

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia
- Kaspersky Mail Anti-Virus blocked 166,187,118 malicious email attachments

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia
- Kaspersky Mail Anti-Virus blocked 166,187,118 malicious email attachments
- Anti-Phishing system thwarted 507,851,735 attempts to follow phishing links

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia
- Kaspersky Mail Anti-Virus blocked 166,187,118 malicious email attachments
- Anti-Phishing system thwarted 507,851,735 attempts to follow phishing links
- 378,496 attempts to follow phishing links were associated with Telegram a/c hijacking

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia
- Kaspersky Mail Anti-Virus blocked 166,187,118 malicious email attachments
- Anti-Phishing system thwarted 507,851,735 attempts to follow phishing links
- 378,496 attempts to follow phishing links were associated with Telegram a/c hijacking

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam

Kaspersky lab's survey on 2022. [1]

- 48.63% of all emails around the world were Spam, whereas 52.78% of all emails in the Russian segment of the internet were Spam
- As much as 29.82% of all Spam emails originated in Russia
- Kaspersky Mail Anti-Virus blocked 166,187,118 malicious email attachments
- Anti-Phishing system thwarted 507,851,735 attempts to follow phishing links
- 378,496 attempts to follow phishing links were associated with Telegram a/c hijacking

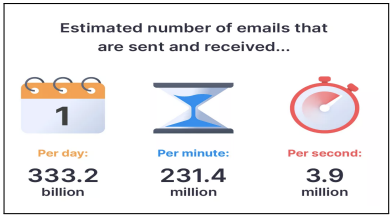


Figure: Src: www.mackeeper.com

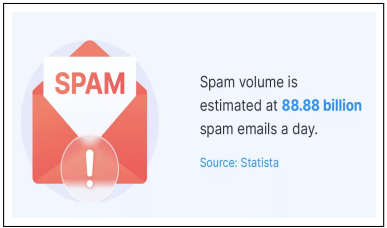


Figure: Src: statistica report

[1] https://securelist.com/Spam-phishing-scam-report-2022/108692/

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters
  - do more than just checking junk emails using pre-existing rules

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters
  - do more than just checking junk emails using pre-existing rules
  - the Spam filters employed generate new rules themselves based on what they have learnt for effective filtering.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters
  - do more than just checking junk emails using pre-existing rules
  - the Spam filters employed generate new rules themselves based on what they have learnt for effective filtering.
- The Google ML model is claimed to detect and filter out Spam and phishing emails with about 99.9 percent accuracy.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters
  - do more than just checking junk emails using pre-existing rules
  - the Spam filters employed generate new rules themselves based on what they have learnt for effective filtering.
- The Google ML model is claimed to detect and filter out Spam and phishing emails with about 99.9 percent accuracy.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- Leading email providers such as Gmail, Yahoo mail and Outlook use the combination of different machine learning (ML) techniques in its Spam filters.
  - these learn and identify Spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers.
- by adapting to varying conditions, the Spam filters
  - do more than just checking junk emails using pre-existing rules
  - the Spam filters employed generate new rules themselves based on what they have learnt for effective filtering.
- The Google ML model is claimed to detect and filter out Spam and phishing emails with about 99.9 percent accuracy.

1

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.

[1]:SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.
  - delaying the delivery of some of these suspicious emails allows ML filtering engine conduct a deeper examination while more messages arrives in due course of time

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.
  - delaying the delivery of some of these suspicious emails allows ML filtering engine conduct a deeper examination while more messages arrives in due course of time
  - allows the the algorithms to be updated in real time.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.
  - delaying the delivery of some of these suspicious emails allows ML filtering engine conduct a deeper examination while more messages arrives in due course of time
  - allows the the algorithms to be updated in real time.
  - it is observed that only about 0.05 percent of emails are affected by this deliberate delay.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.
  - delaying the delivery of some of these suspicious emails allows ML filtering engine conduct a deeper examination while more messages arrives in due course of time
  - allows the the algorithms to be updated in real time.
  - it is observed that only about 0.05 percent of emails are affected by this deliberate delay.

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Historical aspects and impact of Spam...

- As per the statistics from Google between 50-70 percent of emails that Gmail receives are unsolicited mail.
- Google's detection models also use, tools viz. Google Safe Browsing for identifying websites that have malicious URLs, delay the delivery of some Gmail messages
  - to carry out additional comprehensive scrutiny of the phishing messages since they are easier to detect when they are analyzed collectively.
  - delaying the delivery of some of these suspicious emails allows ML filtering engine conduct a deeper examination while more messages arrives in due course of time
  - allows the the algorithms to be updated in real time.
  - it is observed that only about 0.05 percent of emails are affected by this deliberate delay.

1

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Spam Detection: How Gmail, Yahoo and Outlook emails Spam filters work

How Gmail, Yahoo and Outlook emails Spam filters work ?

- X

How Gmail, Yahoo and Outlook emails Spam filters work ?...

- X

How Gmail, Yahoo and Outlook emails Spam filters work ?...

- X

Self-reading & preparation topics end...

Basic ML-based Spam Detection Architecture, Building Blocks

Email Spam filtering architecture

- Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails.

Email Spam filtering architecture

- Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails.
- Mail filters are generally used to manage incoming mails, filter Spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware.

Email Spam filtering architecture

- Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails.
- Mail filters are generally used to manage incoming mails, filter Spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware.
- Spam filters are deployed by many ISPs at every layer of the network, in front of email server or at mail relay where there is the presence of firewall.

Email Spam filtering architecture

- Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails.
- Mail filters are generally used to manage incoming mails, filter Spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware.
- Spam filters are deployed by many ISPs at every layer of the network, in front of email server or at mail relay where there is the presence of firewall.
- Filters block unsolicited or suspicious emails that are a threat to the security of network.

Email Spam filtering architecture

- Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails.
- Mail filters are generally used to manage incoming mails, filter Spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware.
- Spam filters are deployed by many ISPs at every layer of the network, in front of email server or at mail relay where there is the presence of firewall.
- Filters block unsolicited or suspicious emails that are a threat to the security of network.
- In order to identify an email as a Spam, various ML-based approaches could be employed - a broad schematic of which is shown in the next slide.....

Spam Detection background: Conventional ML-based approach

Generally consists of four major steps:

- Dataset collection,
- Pre-processing the dataset,
- Training and Testing models and
- Comparing and Analyzing results.

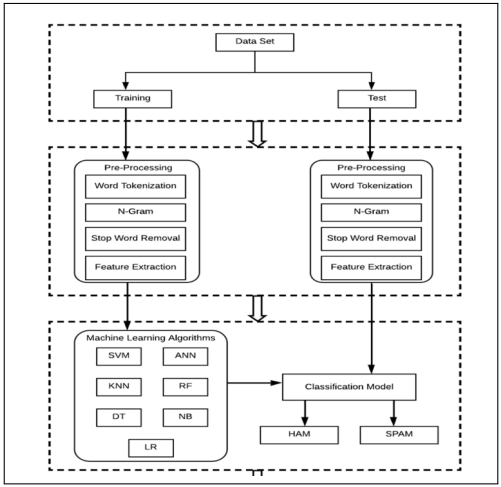


Figure: ML-based Spam filtering

[1]

[1] :SourceEmmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Categories Spam Detection Mechanisms: Another View

Filtering of email Spam used to follow one of the following approaches viz.

- Content Based Filtering Technique

Spam Detection Approaches Classification/Alternative Mechanisms

Filtering of email Spam used to follow one of the following approaches viz.

- Content Based Filtering Technique
- Case Base Spam Filtering Method

Spam Detection Approaches Classification/Alternative Mechanisms

Filtering of email Spam used to follow one of the following approaches viz.

- Content Based Filtering Technique
- Case Base Spam Filtering Method
- Heuristic OR Rule Based OR Knowledge engineering-based approach

Spam Detection Approaches Classification/Alternative Mechanisms

Filtering of email Spam used to follow one of the following approaches viz.

- Content Based Filtering Technique
- Case Base Spam Filtering Method
- Heuristic OR Rule Based OR Knowledge engineering-based approach
- Previous Likeness Based Spam Filtering Technique

Spam Detection Approaches Classification/Alternative Mechanisms

Filtering of email Spam used to follow one of the following approaches viz.

- Content Based Filtering Technique
- Case Base Spam Filtering Method
- Heuristic OR Rule Based OR Knowledge engineering-based approach
- Previous Likeness Based Spam Filtering Technique
- Adaptive Spam Filtering Technique

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

[1]

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on

[1] [1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on
  - analyzing words, the occurrence, and distributions of words and phrases in the content of emails

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on
  - analyzing words, the occurrence, and distributions of words and phrases in the content of emails
  - identifying certain features (normally keywords frequently utilized in Spam emails) and generating rules to filter the incoming email Spams

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on
  - analyzing words, the occurrence, and distributions of words and phrases in the content of emails
  - identifying certain features (normally keywords frequently utilized in Spam emails) and generating rules to filter the incoming email Spams
- the rate at which these features appear in emails ascertain the probabilities for each characteristic in the email,

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on
  - analyzing words, the occurrence, and distributions of words and phrases in the content of emails
  - identifying certain features (normally keywords frequently utilized in Spam emails) and generating rules to filter the incoming email Spams
- the rate at which these features appear in emails ascertain the probabilities for each characteristic in the email,
- the probability ascertained is measured against a threshold value pre-computed and

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Content-based filtering

- used to create automatic filtering rules and to classify emails using ML approaches, using classifiers such as

Naïve Bayesian classification	Neural Networks,
K Nearest Neighbor	Decision-tree
Artificial immune systemm	Support Vector Machine,
or other classifiers	-

- the method is based on
  - analyzing words, the occurrence, and distributions of words and phrases in the content of emails
  - identifying certain features (normally keywords frequently utilized in Spam emails) and generating rules to filter the incoming email Spams
- the rate at which these features appear in emails ascertain the probabilities for each characteristic in the email,
- the probability ascertained is measured against a threshold value pre-computed and
- email messages that exceed the threshold value are classified as Spam

[1]

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.
- two key issues

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.
- two key issues
  - an effective signature mechanism needs to be devised and

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.
- two key issues
  - an effective signature mechanism needs to be devised and
  - a process for sharing these signatures needs to be developed.

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.
- two key issues
  - an effective signature mechanism needs to be devised and
  - a process for sharing these signatures needs to be developed.
- Spammers insert random characters into messages to foil hash-based signatures

Spam Detection Approaches: Collaborative Spam filtering

Collaborative Spam filtering is a variation of the signature based filtering wherein,

- in addition to the content-based approach to Spam filtering there is also some work on a collaborative approach, to do so.
- this approach does not consider the content of the email but depends on the collaboration of groups of users who share information about Spam.
- when a new Spam message appears, an early receiver of the Spam shares a signature for that Spam, with the rest of the group.
- the signature is typically one or more hash codes.
- if the other users also receive this message their filters can identify it as Spam based on the shared signature.
- two key issues
  - an effective signature mechanism needs to be devised and
  - a process for sharing these signatures needs to be developed.
- Spammers insert random characters into messages to foil hash-based signatures
- hence, flexible and adaptive signatures are needed.

Spam Detection Approaches: Case base or sample base filtering

Case base or Sample based filtering

- is one of the popular Spam filtering methods.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Case base or sample base filtering

Case base or Sample based filtering

- is one of the popular Spam filtering methods.
- firstly, all emails both non-Spam and Spam emails are extracted from each user's email using collection model.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Case base or sample base filtering

Case base or Sample based filtering

- is one of the popular Spam filtering methods.
- firstly, all emails both non-Spam and Spam emails are extracted from each user's email using collection model.
- subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Case base or sample base filtering

Case base or Sample based filtering

- is one of the popular Spam filtering methods.
- firstly, all emails both non-Spam and Spam emails are extracted from each user's email using collection model.
- subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process.
- the data is then classified into two vector sets.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Spam Detection Approaches: Case base or sample base filtering

Case base or Sample based filtering

- is one of the popular Spam filtering methods.
- firstly, all emails both non-Spam and Spam emails are extracted from each user's email using collection model.
- subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process.
- the data is then classified into two vector sets.
- lastly, the machine learning algorithm is used to train datasets and test them to decide whether the incoming mails are Spam/non-Spam.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

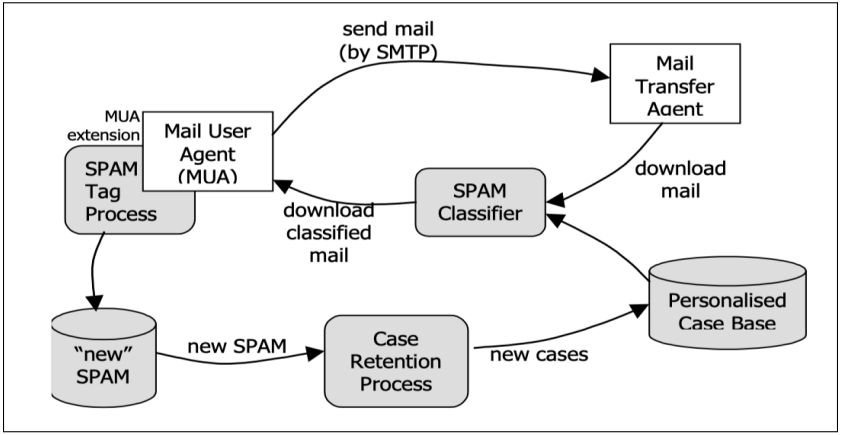
Spam Detection Approaches: Case base or sample base filtering...

Case base or sample base filtering works at the enterprise level

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Spam Detection Approaches: Case base or sample base filtering...

Case base or sample base filtering works at the enterprise level



[1]

Figure: Case base or sample base filtering

[1] Src: https://www.analyticsvidhya.com/blog/2020/10/what-is-the-convolutional-neural-network-architecture/

Heuristic/Rule Based Spam Filtering

- approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Heuristic/Rule Based Spam Filtering

- approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.
- several similar patterns increase the score of a message, whereas if any of the patterns did not correspond, it deducts from the score.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Heuristic/Rule Based Spam Filtering

- approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.
- several similar patterns increase the score of a message, whereas if any of the patterns did not correspond, it deducts from the score.
- any message's score that surpasses a specific threshold is filtered as a Spam; else it is counted as valid.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Heuristic/Rule Based Spam Filtering

- approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.
- several similar patterns increase the score of a message, whereas if any of the patterns did not correspond, it deducts from the score.
- any message's score that surpasses a specific threshold is filtered as a Spam; else it is counted as valid.
- with the new Spam messages, ranking rules are constantly updated to be able to cope with the introduction of new Spam messages.

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Heuristic/Rule Based Spam Filtering

- approach uses already created rules or heuristics to assess a huge number of patterns which are usually regular expressions against a chosen message.
- several similar patterns increase the score of a message, whereas if any of the patterns did not correspond, it deducts from the score.
- any message's score that surpasses a specific threshold is filtered as a Spam; else it is counted as valid.
- with the new Spam messages, ranking rules are constantly updated to be able to cope with the introduction of new Spam messages.
- A good example of a rule based Spam filter is SpamAssassin - tagged as the #1 Enterprise Open-Source Spam Filter (Ref: https://Spamassassin.apache.org/)

1

[1] V. Christina, S. Karpagavalli, G. Suganya, Email Spam filtering using supervised machine learning techniques, Int. J. Comput. Sci. Eng. 02 (09) (2010) 3126–3129.

Previous Likeness Based Spam Filtering

- approach uses memory-based, or instance-based, ML methods to classify incoming emails based on their resemblance to stored examples (e.g. training emails).

1

[1] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, Stacking classifiers for anti-Spam filtering of E-mail, in: Empirical Methods in Natural Language Processing, 2001, pp. 44–50.

Previous Likeness Based Spam Filtering

- approach uses memory-based, or instance-based, ML methods to classify incoming emails based on their resemblance to stored examples (e.g. training emails).
- the attributes of the email are used to create a multi-dimensional space vector, which is used to plot new instances as points.

1

[1] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, Stacking classifiers for anti-Spam filtering of E-mail, in: Empirical Methods in Natural Language Processing, 2001, pp. 44–50.

Previous Likeness Based Spam Filtering

- approach uses memory-based, or instance-based, ML methods to classify incoming emails based on their resemblance to stored examples (e.g. training emails).
- the attributes of the email are used to create a multi-dimensional space vector, which is used to plot new instances as points.
- the new instances are afterwards allocated to the most popular class of its K-closest training instances.

1

[1] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, Stacking classifiers for anti-Spam filtering of E-mail, in: Empirical Methods in Natural Language Processing, 2001, pp. 44–50.

Previous Likeness Based Spam Filtering

- approach uses memory-based, or instance-based, ML methods to classify incoming emails based on their resemblance to stored examples (e.g. training emails).
- the attributes of the email are used to create a multi-dimensional space vector, which is used to plot new instances as points.
- the new instances are afterwards allocated to the most popular class of its K-closest training instances.
- This approach uses the k-nearest neighbor (kNN) for filtering Spam emails.

1

[1] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, Stacking classifiers for anti-Spam filtering of E-mail, in: Empirical Methods in Natural Language Processing, 2001, pp. 44–50.

Adaptive Spam Filtering Technique

- the method detects and filters Spam by grouping them into different classes.

1

[1] L. Pelletier, J. Almhana, V. Choulakian, Adaptive filtering of Spam, in: Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004.

Adaptive Spam Filtering Technique

- the method detects and filters Spam by grouping them into different classes.
- it divides an email corpus into various groups, each group has an emblematic text.

1

[1] L. Pelletier, J. Almhana, V. Choulakian, Adaptive filtering of Spam, in: Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004.

Adaptive Spam Filtering Technique

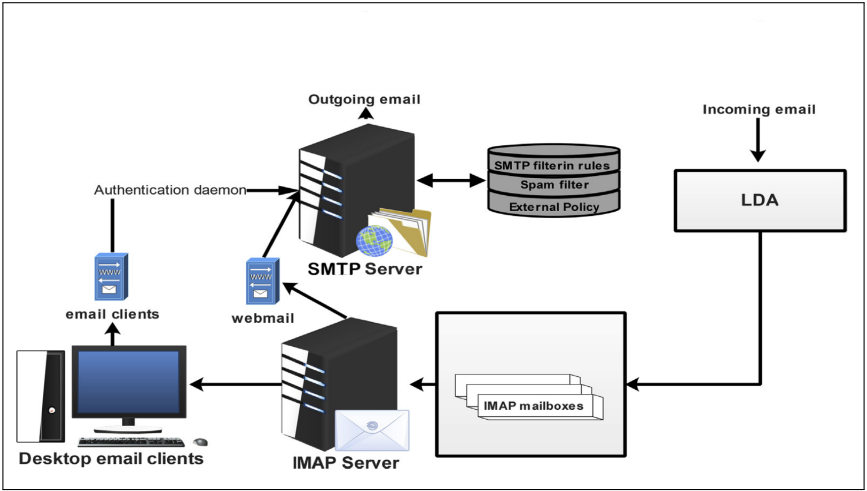
- the method detects and filters Spam by grouping them into different classes.
- it divides an email corpus into various groups, each group has an emblematic text.
- a comparison is made between each incoming email and each group, and a percentage of similarity is produced to decide the probable group the email belongs to.

1

[1] L. Pelletier, J. Almhana, V. Choulakian, Adaptive filtering of Spam, in: Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004.

Basic ML-based Spam Filtering process

Fundamental Spam Detection: Basic Architecture



1

Figure: Spam Filtering Mechanism

1_1

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body

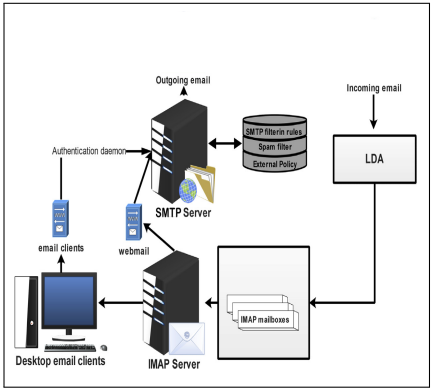


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body
- the header contains information about the content of the email e.g.

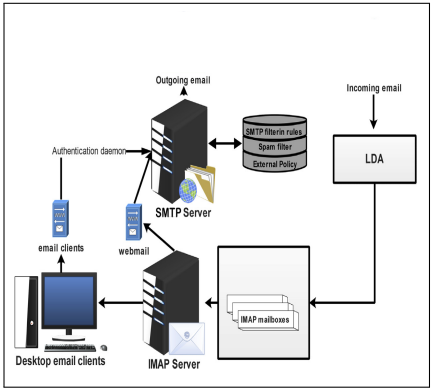


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body
- the header contains information about the content of the email e.g.
  - the subject, sender's email ID, the receiver email ID, the time-stamp (i.e. when the message was sent by intermediary servers to the MTAs)

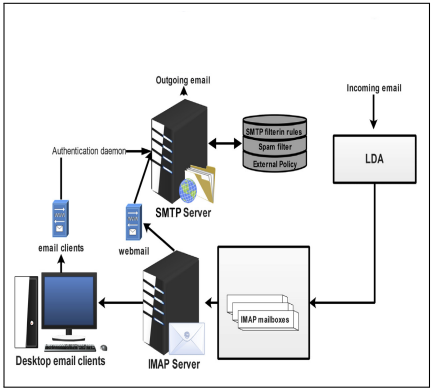


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body
- the header contains information about the content of the email e.g.
  - the subject, sender's email ID, the receiver email ID, the time-stamp (i.e. when the message was sent by intermediary servers to the MTAs)
- the header goes through modification whenever it moves from one server to another through an in-between server.

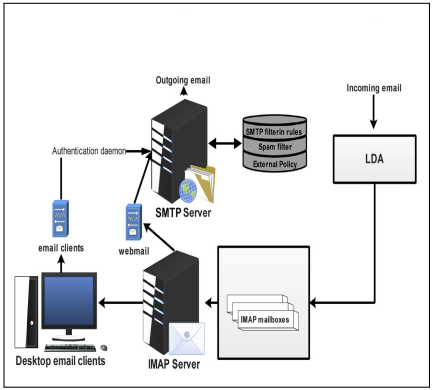


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body
- the header contains information about the content of the email e.g.
  - the subject, sender's email ID, the receiver email ID, the time-stamp (i.e. when the message was sent by intermediary servers to the MTAs)
- the header goes through modification whenever it moves from one server to another through an in-between server.
- allows the user to view the route the email passes through, and the time taken by each server to treat the mail.

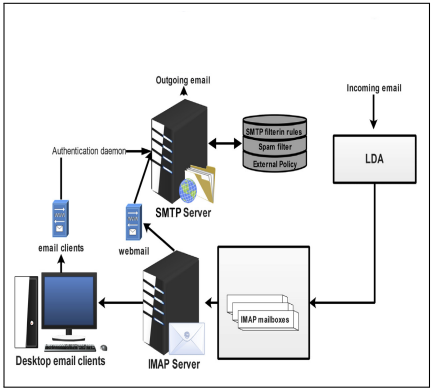


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process

- As already known, an email message is made up of the header and the body
- the header contains information about the content of the email e.g.
  - the subject, sender's email ID, the receiver email ID, the time-stamp (i.e. when the message was sent by intermediary servers to the MTAs)
- the header goes through modification whenever it moves from one server to another through an in-between server.
- allows the user to view the route the email passes through, and the time taken by each server to treat the mail.
- this information can be used by an ML classifier for detecting Spam and filtering mails.

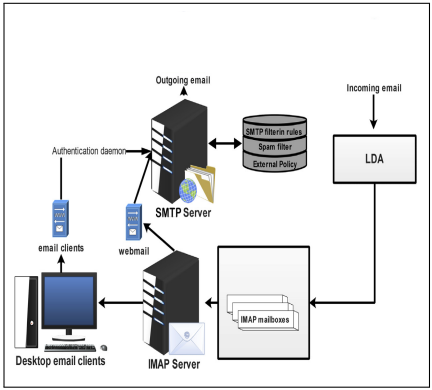


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages that must be observed in the mining of data from an email message can be:

- Pre-processing: the first stage that is executed whenever an incoming mail is received.

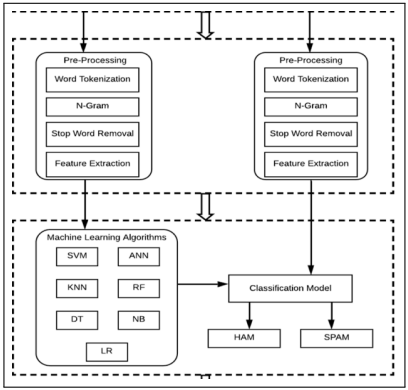


Figure: Spam Filtering Mechanism

1

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages that must be observed in the mining of data from an email message can be:

- Pre-processing: the first stage that is executed whenever an incoming mail is received.
- Tokenization: a process that removes the words in the body of an email and also transforms a message to its meaningful parts.

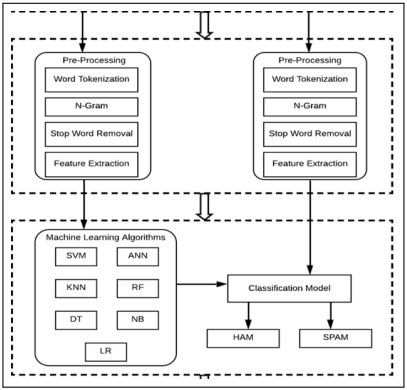


Figure: Spam Filtering Mechanism

1

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages that must be observed in the mining of data from an email message can be:

- Pre-processing: the first stage that is executed whenever an incoming mail is received.
- Tokenization: a process that removes the words in the body of an email and also transforms a message to its meaningful parts.
  - divides the email into a sequence of representative symbols called tokens.

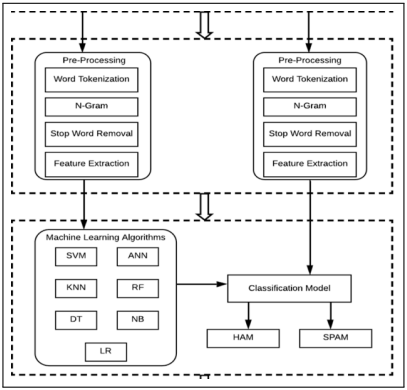


Figure: Spam Filtering Mechanism

1

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages that must be observed in the mining of data from an email message can be:

- Pre-processing: the first stage that is executed whenever an incoming mail is received.
- Tokenization: a process that removes the words in the body of an email and also transforms a message to its meaningful parts.
  - divides the email into a sequence of representative symbols called tokens.
  - the representative symbols are extracted from the body of the email, the header and subject.

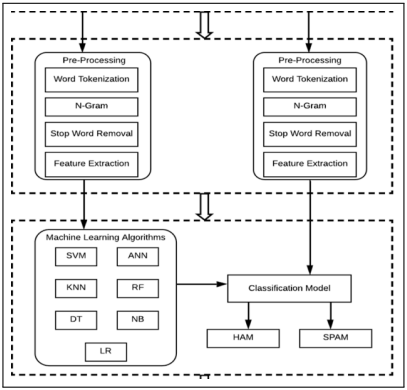


Figure: Spam Filtering Mechanism

1

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages that must be observed in the mining of data from an email message can be:

- Pre-processing: the first stage that is executed whenever an incoming mail is received.
- Tokenization: a process that removes the words in the body of an email and also transforms a message to its meaningful parts.
  - divides the email into a sequence of representative symbols called tokens.
  - the representative symbols are extracted from the body of the email, the header and subject.
  - at times, the symbols are replaced with distinctive identification symbols to extricate all the characteristics and words from the email exclusive of taking into account the meaning.

[1]

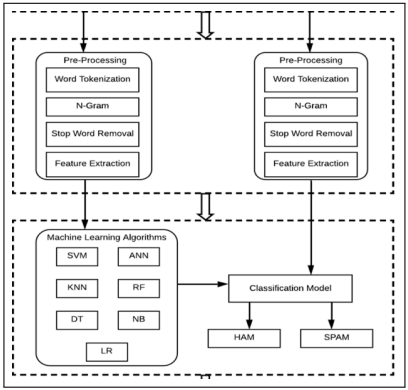


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages in in the mining of data from an email message can be:

- Feature selection: a kind of reduction in the measure of spatial coverage that reduces fragments of email message as a compressed feature vector.

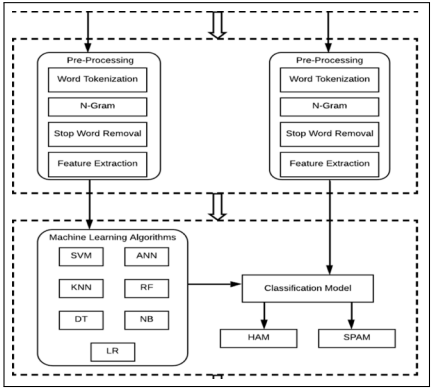


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages in in the mining of data from an email message can be:

- Feature selection: a kind of reduction in the measure of spatial coverage that reduces fragments of email message as a compressed feature vector.
- essential when the size of the message is large and a condensed feature representation is needed to make the task of text or image matching snappy.

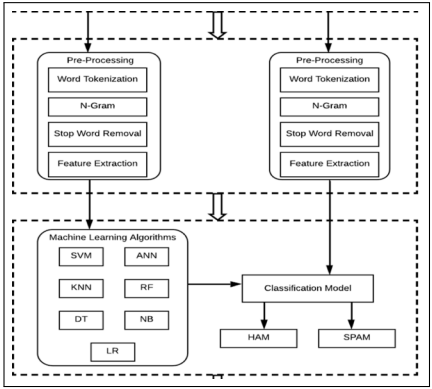


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages in in the mining of data from an email message can be:

- Feature selection: a kind of reduction in the measure of spatial coverage that reduces fragments of email message as a compressed feature vector.
- essential when the size of the message is large and a condensed feature representation is needed to make the task of text or image matching snappy.
- the recognition of Spam e-mails with minimum number of features is important in view of computational complexity.

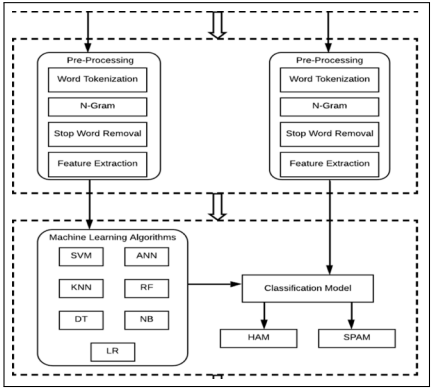


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages in in the mining of data from an email message can be:

- Feature selection: a kind of reduction in the measure of spatial coverage that reduces fragments of email message as a compressed feature vector.
- essential when the size of the message is large and a condensed feature representation is needed to make the task of text or image matching snappy.
- the recognition of Spam e-mails with minimum number of features is important in view of computational complexity.
- involves processes like stemming, noise removal and stop word removal steps.

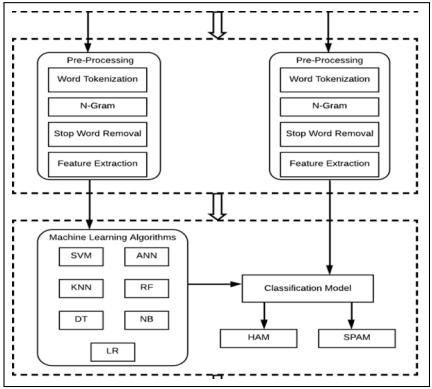


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

The necessary stages in in the mining of data from an email message can be:

- Feature selection: a kind of reduction in the measure of spatial coverage that reduces fragments of email message as a compressed feature vector.
- essential when the size of the message is large and a condensed feature representation is needed to make the task of text or image matching snappy.
- the recognition of Spam e-mails with minimum number of features is important in view of computational complexity.
- involves processes like stemming, noise removal and stop word removal steps.
- the indicate types of messages/features shown on the next slide.

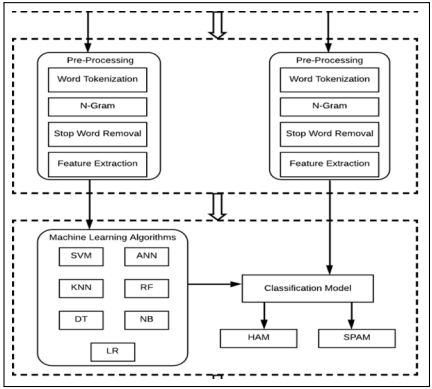


Figure: Spam Filtering Mechanism

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Basic ML-based Spam Filtering process...

Typical messages that can be classified as Spam

Advance fee fraud, including inheritance, lottery, visa and customs-clearance scams

Romance scams, including marketing sex enhancement drugs to cure erection dysfunctional, online dating,

Military scams, Ads for porn sites, Ads for miscellaneous external sites, earning big money through "work-from-home" jobs,

Online shopping, pleading and gift requests, business proposals and others.

Basic ML-based Spam Filtering process...

Typical messages that can be classified as Spam

Advance fee fraud, including inheritance, lottery, visa and customs-clearance scams
Romance scams, including marketing sex enhancement drugs to cure erection dysfunctional, online dating,
Military scams, Ads for porn sites, Ads for miscellaneous external sites, earning big money through "work-from-home" jobs,
Online shopping, pleading and gift requests, business proposals and others.

Some of the most important features for Spam filtering include:

1. Message body and subject,
2. Adult content
3. Recipient age, Sex and country
4. Volume of the message, Occurrence count of words
5. Recipient replied (indicates whether the recipient replied to the message),
6. Circadian patterns of the message (Spam messages usually have many semantic discrepancies),
7. Bag of words from the message content

Basic ML-based Spam Filtering process...

Some of the Sender Account Features used for Spam filtering are: include:

1. Sender Country (from IPA),
2. Sender IP address,
3. Username of the sender,
4. Sender & Recipient Age,
5. Sender Reputation.
6. Sender's date of birth,
7. Sender Email,
8. Account lifespan,
9. Sex of sender
10. Age of recipient.
11. Geographical distance between sender & receiver,

1

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Publicly available email Spam corpus

A comprehensive list of the corpora made available to the public in the different techniques are as follows:

Table 2
Publicly available email spam corpus.

Dataset name	Number of messages		Rate of spam	Year of creation	References
	Spam	Non-spam			
Spam archive	15090	0	100%	1998	Almeida and yamakami [32]
Spambase	1813	2788	39%	1999	Sakkis et al [33]
Lingspam	481	2412	17%	2000	Sakkis et al [33]
PU1	481	618	44%	2000	Attar et al [34]
Spamassassin	1897	4150	31%	2002	Apache spamassassin [35]
PU2	142	579	20%	2003	Zhang et al [36]
PU3	1826	2313	44%	2003	Zhang et al [36]
PUA	571	571	50%	2003	Zhang et al [36]
Zh1	1205	428	74%	2004	Zhang et al [36]
Gen spam	31,196	9212	78%	2005	Cormack and lynam [37]
Trec 2005	52,790	39,399	57%	2005	Androutsopoulos et al [38]
Biggio	8549	0	100	2005	Biggio et al [39]
Phishing corpus	415	0	100	2005	Abu-nimeh et al [40]
Enron-spam	20170	16545	55%	2006	Koprinska et al [41]
Trec 2006	24,912	12,910	66%	2006	Androutsopoulos et al [42]
Trec 2007	50,199	25,220	67%	2007	Debarr and wechsler [43]
Princeton spam image Benchmark	1071	0	100%	2007	Wang et al [44]
Dredze image spam Dataset	3297	2021	62%	2007	Dredze, gevaryahu and elias-bachrach [45]
Hunter	928	810	53%	2008	Gao et al [46]
Spamemail	1378	2949	32%	2010	Csminninggroup [47]

Figure: Publicly available email Spam corpus

[1]

[1] Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

Performance Evaluation Measures

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve
- Weighted Accuracy (W_{Acc}),

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve
- Weighted Accuracy (W_{Acc}),
- Weighted Error Rate (W_{Err})

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve
- Weighted Accuracy (W_{Acc}),
- Weighted Error Rate (W_{Err})
- Total Cost Ratio (TCR)

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve
- Weighted Accuracy (W_{Acc}),
- Weighted Error Rate (W_{Err})
- Total Cost Ratio (TCR)
- F-measure i.e. (F1-score or F-score)

Performance evaluation measures

Typical performance measures that used in the research papers on Spam detection using ML/DL approaches are as follows:

- Classification Accuracy(Acc)
- Classification Error (Err)
- From the field of information retrieval viz. Recall, Precision & other derived measures
- From the field of decision theory viz. false positives and false negatives
- True positive event and true negative event
- False Positive Rate (FPR) & False Negative Rate
- True Positive Rate (TPR) & True Negative Rate
- Receiver Operating Characteristics (ROC) curve
- Weighted Accuracy (W_{Acc}),
- Weighted Error Rate (W_{Err})
- Total Cost Ratio (TCR)
- F-measure i.e. (F1-score or F-score)
- λ - a factor that depicts how risky it is to wrongly classify a mail as a Spam

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...
- When a Spam message is wrongly classified as ham, what does a user loose ?

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...
- When a Spam message is wrongly classified as ham, what does a user loose ?
- When a non-Spam message is wrongly labeled as Spam, it indicates the possibility of losing valuable information as a result of the filter's classification error.

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...
- When a Spam message is wrongly classified as ham, what does a user loose ?
- When a non-Spam message is wrongly labeled as Spam, it indicates the possibility of losing valuable information as a result of the filter's classification error.
  - This is very imperative especially where Spam messages are automatically deleted.

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...
- When a Spam message is wrongly classified as ham, what does a user loose ?
- When a non-Spam message is wrongly labeled as Spam, it indicates the possibility of losing valuable information as a result of the filter's classification error.
  - This is very imperative especially where Spam messages are automatically deleted.
  - What would be the classifier's accuracy value if a setting is lopsided, that is, one where the number of Spam messages utilized for testing the performance of the filter is very much higher than that of ham messages ?

Performance evaluation measures: Acc and Err

Typical performance measures...

- Classification Accuracy(Acc):
  - is the comparative number of messages rightly classified as the Spam
  - the percentage of messages rightly classified is used as a measure for evaluating performance of the filter.
  - But the issue is: Is using Acc as the only performance index sufficient?
  - Is Acc the real indicator of the costs attached to misclassification ?
- In addition, let us analyze the following...
- When a Spam message is wrongly classified as ham, what does a user loose ?
- When a non-Spam message is wrongly labeled as Spam, it indicates the possibility of losing valuable information as a result of the filter's classification error.
  - This is very imperative especially where Spam messages are automatically deleted.
  - What would be the classifier's accuracy value if a setting is lopsided, that is, one where the number of Spam messages utilized for testing the performance of the filter is very much higher than that of ham messages ?
- Can one have a zero probability of wrongly categorizing a ham message in the real world ? Therefore the other metrices.....

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:
  - the number of Spam messages correctly classified as the Spam messages, denoted by $|S \rightarrow S|$. Note that the event $S \rightarrow S$ is a true negative event.

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:
  - the number of Spam messages correctly classified as the Spam messages, denoted by $|S \rightarrow S|$. Note that the event $S \rightarrow S$ is a true negative event.
- True Negatives:

Performance evaluation measures: Acc and Err

Let use now use the following notations:

- NH: the Number of non-Spam messages to be classified
- NS: the Number of Spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the Spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:
  - the number of Spam messages correctly classified as the Spam messages, denoted by $|S \rightarrow S|$. Note that the event $S \rightarrow S$ is a true negative event.
- True Negatives:
  - the number of HAM messages correctly classified as the HAM messages, denoted by $|H \rightarrow H|$. Note that the event $H \rightarrow H$ is a true positive event.

Performance evaluation...: False/True Positives/Negatives

True Positive: Cases when we **predicted spam** and our **prediction is true** that means **actual value is also spam.**

True Negatives: Cases when we **did not predict spam** and our **prediction is true** that means **actual value is not spam.**

False Positives: Cases when we **predicted spam** and our **prediction is false** that means **actual value is not spam.**

False Negatives: Cases when we **did not predict spam** and our **prediction is false** that means **actual value is spam.**

Figure: False/True Positives/Negatives

1

[1] https://thatascience.com/learn-machine-learning/Precision-and-Recall/

Performance evaluation measures: Acc and Err...

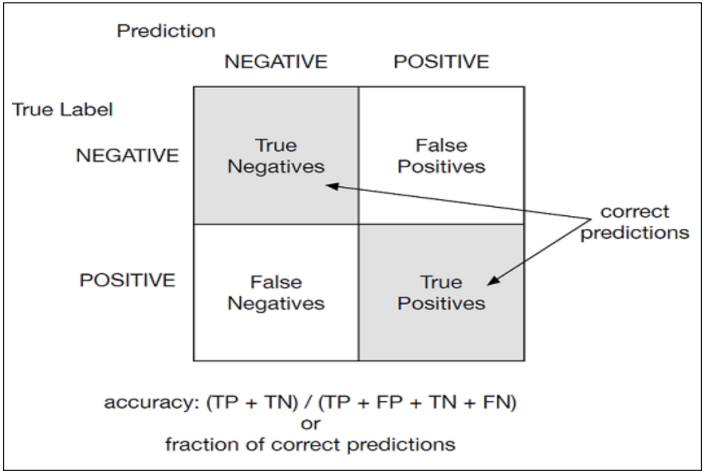


Figure: Accuracy

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

Performance evaluation measures: Acc and Err...

Now, we can formally define the *Acc* and *Err*, as follows:

Performance evaluation measures: Acc and Err...

Now, we can formally define the *Acc* and *Err*, as follows:

$$A_{cc} = \frac{|H \rightarrow H| \ + \ |S \rightarrow S|}{N_H \ + \ N_S} \qquad (1)$$

Performance evaluation measures: Acc and Err...

Now, we can formally define the *Acc* and *Err*, as follows:

$$A_{cc} = \frac{|H \rightarrow H| \ + \ |S \rightarrow S|}{N_H \ + \ N_S} \tag{1}$$

$$E_{rr} = 1 - A_{cc} = \frac{|H \rightarrow S| \ + \ |S \rightarrow H|}{N_H \ + \ N_S} \tag{2}$$

Let use now use the following notations:

- NH: the Number of non-spam messages to be classified
- NS: the Number of spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:
  - the number of Spam messages correctly classified as the spam messages, denoted by $|S \rightarrow S|$. Note that the event $S \rightarrow S$ is a true negative event.
- False Negatives:
  - the number of HAM messages correctly classified as the HAM messages, denoted by $|H \rightarrow H|$. Note that the event $H \rightarrow H$ is a true positive event.

Performance evaluation measures: Acc and Err...

Now, we can formally define the Acc and Err, as follows:

$$A_{cc} = \frac{|H \rightarrow H| \ + \ |S \rightarrow S|}{N_H \ + \ N_S} \quad (1)$$

$$E_{rr} = 1 - A_{cc} = \frac{|H \rightarrow S| \ + \ |S \rightarrow H|}{N_H \ + \ N_S} \quad (2)$$

Let use now use the following notations:

- NH: the Number of non-spam messages to be classified
- NS: the Number of spam messages to be classified
- H - a message actually classified as HAM message,
- S - a message actually classified as SPAM message,

Then,

- False Positives:
  - the number of HAM messages wrongly classified as the spam messages, denoted by $|H \rightarrow S|$
- False Negatives:
  - the number of Spam messages wrongly classified as the HAM messages, denoted by $|S \rightarrow H|$
- True Positives:
  - the number of Spam messages correctly classified as the spam messages, denoted by $|S \rightarrow S|$. Note that the event $S \rightarrow S$ is a true negative event.
- False Negatives:
  - the number of HAM messages correctly classified as the HAM messages, denoted by $|H \rightarrow H|$. Note that the event $H \rightarrow H$ is a true positive event.

- Here, do A_{cc} and E_{rr} mutually **assume** equal cost of a False Positive $|H \rightarrow S|$ and a False Negative $|H \rightarrow S|$ events OR assume different costs ?
- Is it a realistic assumption ?
- this is the proof of our earlier hypothesis that A_{cc} and E_{rr} are necessary but not sufficient metrices to measure the performance of a classifier ?
- hence, now we shall introduce one more expression viz. the false positive rate (FPR) - defined as the ratio of ham or valid e-mails that are classified as Spam....

next slide.

Performance evaluation measures:FPR, FNR

The False Positive Rate (FPR) is expressed as follows:

$$FPR = \frac{Number\ of\ false\ positives}{Number\ of\ false\ positives\ +\ Number\ of\ true\ negatives}$$

i.e.

Performance evaluation measures:FPR, FNR

The False Positive Rate (FPR) is expressed as follows:

$$FPR = \frac{Number\ of\ false\ positives}{Number\ of\ false\ positives\ +\ Number\ of\ true\ negatives}$$

i.e.

$$FPR = \frac{|H \rightarrow S|}{|H \rightarrow S|\ +\ |H \rightarrow H|}$$

Performance evaluation measures:FPR, FNR

The False Positive Rate (FPR) is expressed as follows:

$$FPR = \frac{Number\ of\ false\ positives}{Number\ of\ false\ positives\ +\ Number\ of\ true\ negatives}$$

i.e.

$$FPR = \frac{|H \rightarrow S|}{|H \rightarrow S|\ +\ |H \rightarrow H|}$$

Similarly, The False Negative Rate (NPR) is expressed as follows:

$$NPR = \frac{Number\ of\ false\ negatives}{Number\ of\ true\ positives\ +\ Number\ of\ false\ negatives}$$

i.e.

Performance evaluation measures:FPR, FNR

The False Positive Rate (FPR) is expressed as follows:

$$FPR = \frac{Number\ of\ false\ positives}{Number\ of\ false\ positives\ +\ Number\ of\ true\ negatives}$$

i.e.

$$FPR = \frac{|H \rightarrow S|}{|H \rightarrow S|\ +\ |H \rightarrow H|}$$

Similarly, The False Negative Rate (NPR) is expressed as follows:

$$NPR = \frac{Number\ of\ false\ negatives}{Number\ of\ true\ positives\ +\ Number\ of\ false\ negatives}$$

i.e.

$$FPR = \frac{|S \rightarrow H|}{|S \rightarrow S|\ +\ |S \rightarrow H|}$$

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- therefore, is described as the worth or the reliability of the filter.

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- therefore, is described as the worth or the reliability of the filter.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$) and are actually Spam, by the total number of email messages detected as the Spam i.e.

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- therefore, is described as the worth or the reliability of the filter.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$) and are actually Spam, by the total number of email messages detected as the Spam i.e.

Performance evaluation measures: Precision

Precision (P_s) is

- in general, is used to measure the accuracy of positive predictions, i.e. in Spam detection is used for obtaining the characteristic of Spam filters.
- therefore, is often in general, referred to as is the ratio between the True Positives and all the Positives. That is,

$$Precision = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Positives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- therefore, is described as the worth or the reliability of the filter.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$) and are actually Spam, by the total number of email messages detected as the Spam i.e.

$$FPR = \frac{|S \rightarrow S|}{|H \rightarrow S|\ +\ |S \rightarrow S|}$$

e.g. if it is 0.843, or when it predicts that an email is a Spam, the classifier is correct around 84% of the time.

Performance evaluation measures: Precision

Thus, graphically the Precision is....

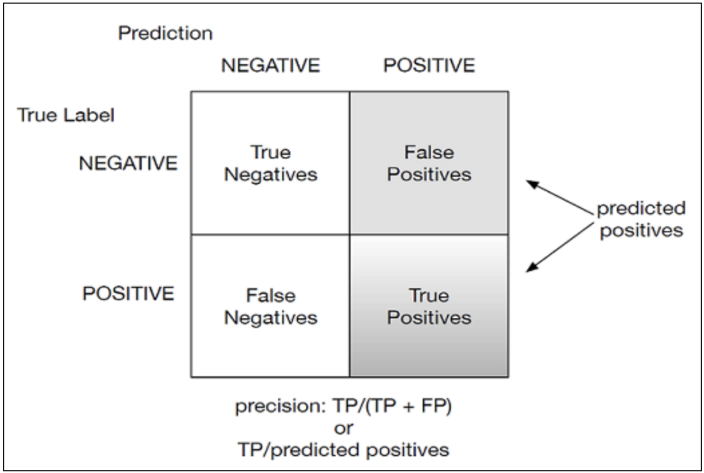


Figure: Precision

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

Performance evaluation measures: Recall

Recall (R_s) is

- in general, is also used to measure the accuracy of positive predictions, but in an opposite manner as compared to Precision.

Performance evaluation measures: Recall

Recall (R_s) is

- in general, is also used to measure the accuracy of positive predictions, but in an opposite manner as compared to Precision.
- is the measure of the model correctly identifying messages as True Positives. Thus, for all the mails that may have been Spam, Recall tells us how many were correctly identified as Spam.

Performance evaluation measures: Recall

Recall (R_s) is

- in general, is also used to measure the accuracy of positive predictions, but in an opposite manner as compared to Precision.
- is the measure of the model correctly identifying messages as True Positives. Thus, for all the mails that may have been Spam, Recall tells us how many were correctly identified as Spam.
- thus, in other words, it is the comparative number of Spam messages that the filter succeeded in preventing from entering email inbox i.e. detected as Spam, even when some messages may have been wrongly classified as HAM.

Performance evaluation measures: Recall

Recall (R_s) is

- in general, is also used to measure the accuracy of positive predictions, but in an opposite manner as compared to Precision.
- is the measure of the model correctly identifying messages as True Positives. Thus, for all the mails that may have been Spam, Recall tells us how many were correctly identified as Spam.
- thus, in other words, it is the comparative number of Spam messages that the filter succeeded in preventing from entering email inbox i.e. detected as Spam, even when some messages may have been wrongly classified as HAM.
- it is also known as fraction of all positive results which were detected.

Performance evaluation measures: Recall

Recall (R_s) is

- in general, is also used to measure the accuracy of positive predictions, but in an opposite manner as compared to Precision.
- is the measure of the model correctly identifying messages as True Positives. Thus, for all the mails that may have been Spam, Recall tells us how many were correctly identified as Spam.
- thus, in other words, it is the comparative number of Spam messages that the filter succeeded in preventing from entering email inbox i.e. detected as Spam, even when some messages may have been wrongly classified as HAM.
- it is also known as fraction of all positive results which were detected.
- that is, Recall is the percentage of correctly ruled positives (as Spam) out of all of the actual positives - the term actual meaning, even if a mail is termed as false negative, it IS actually a Spam and hence is an actual positive.

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- to control false negatives, one would look for a higher Recall score - that is possible only of the False Negatives i.e. $|S \rightarrow H|$ is reduced.

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- to control false negatives, one would look for a higher Recall score - that is possible only of the False Negatives i.e. $|S \rightarrow H|$ is reduced.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$), by the sum of the total number of email messages detected as the Spam and those that are actually Spam but detected as HAM i.e.

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- to control false negatives, one would look for a higher Recall score - that is possible only of the False Negatives i.e. $|S \rightarrow H|$ is reduced.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$), by the sum of the total number of email messages detected as the Spam and those that are actually Spam but detected as HAM i.e.

Performance evaluation measures: Recall...

Recall (R_s)

- therefore, is often in general, referred to as is the ratio between the True Positives and all the actual Positives. That is,

$$Recall = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

- thus, it is the fraction of results classified as Positive i.e. as the Spam , out of all the Positives, i.e. the Spam..
- to control false negatives, one would look for a higher Recall score - that is possible only of the False Negatives i.e. $|S \rightarrow H|$ is reduced.
- thus, is calculated by dividing the number of messages categorised by the filter as Spam (i.e. $|S \rightarrow S|$), by the sum of the total number of email messages detected as the Spam and those that are actually Spam but detected as HAM i.e.

$$Recall = \frac{|S \rightarrow S|}{|S \rightarrow S|\ +\ |S \rightarrow H|}$$

Performance evaluation measures: Recall...

- Recall is also envisaged to indicate sensitivity of a model

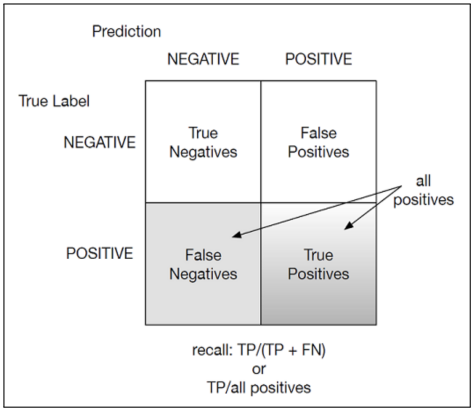


Figure: Recall, graphically

1

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

Performance evaluation measures: Recall...

- Recall is also envisaged to indicate sensitivity of a model
- i.e. say in a medical application, what if a patient has heart disease, but no treatment is given to him/her because our model predicted a false negative report ?

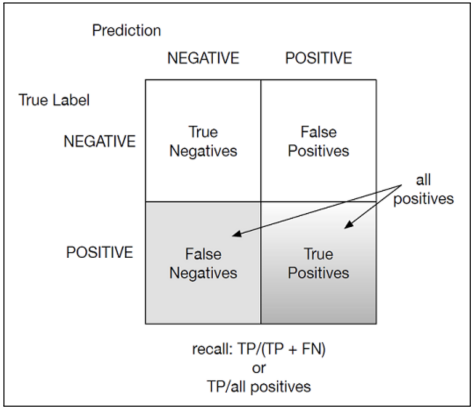


Figure: Recall, graphically

1

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

Performance evaluation measures: Recall...

- Recall is also envisaged to indicate sensitivity of a model
- i.e. say in a medical application, what if a patient has heart disease, but no treatment is given to him/her because our model predicted a false negative report ?
- That is a situation we would like to avoid!

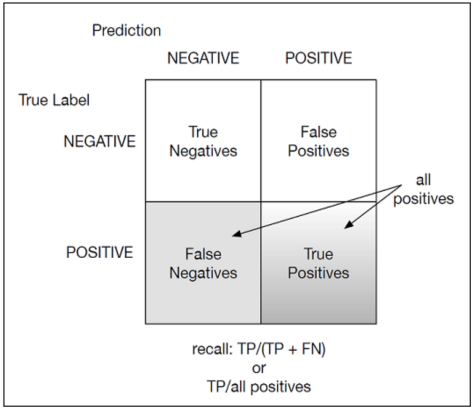


Figure: Recall, graphically

1

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

Performance evaluation measures: Recall...

- Recall is also envisaged to indicate sensitivity of a model
- i.e. say in a medical application, what if a patient has heart disease, but no treatment is given to him/her because our model predicted a false negative report ?
- That is a situation we would like to avoid!
- And, the occurrence of such situation would be conveyed with a low Recall because the denominator in above expression would be high !

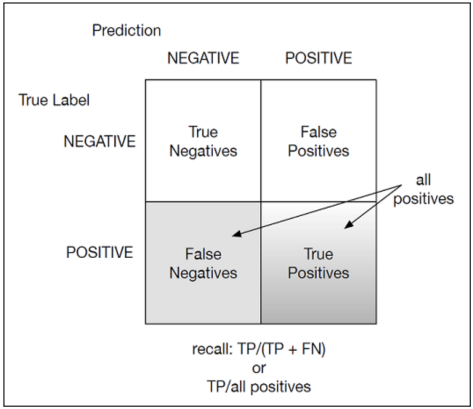


Figure: Recall, graphically

1

[1] https://freecontent.manning.com/evaluating-a-classification-model-with-a-Spam-filter/

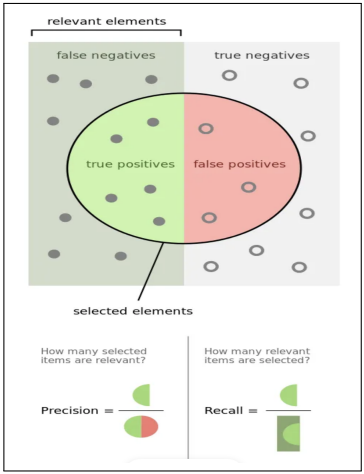


Figure: Recall & Precision

[1] https://commons.wikimedia.org/w/index.php?curid=36926283

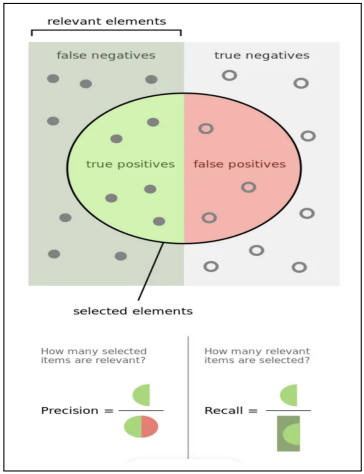


Figure: Recall & Precision

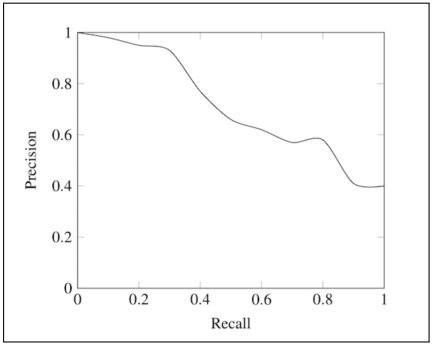


Figure: Recall & Precision: inversely proportional to each other

[1] https://commons.wikimedia.org/w/index.php?curid=36926283

Performance evaluation measures: An Example

- Consider that out of 100 emails, only 2 are Spams and 98 are HAMs. IClassifier predicts all as HAM. Then Acc = ?. But is that acceptable?
- Now look at other data as follows:
  - True Positives = $|S \rightarrow S| = 221$
  - True Negatives = $|H \rightarrow H| = 1414$
  - False Positives = $|H \rightarrow S| = 20$
  - False Negatives = = $|S \rightarrow H| = 17$
- Acc = ?
- Err = ?
- Pr = ?
- Rr= ?

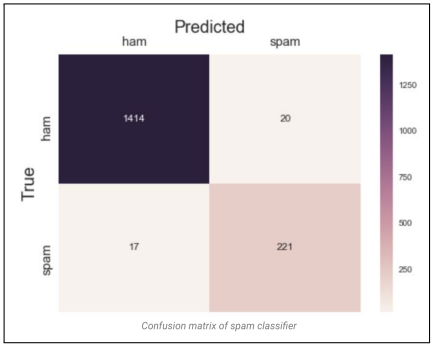


Figure: Confusion Matrix of Spam Classifier

Performance evaluation measures: Relating Pr and Rr

- At the lowest point, i.e., at (0, 0)- the threshold is set at 1.0. That is, the model makes no distinctions between the Spam and the HAM emails.

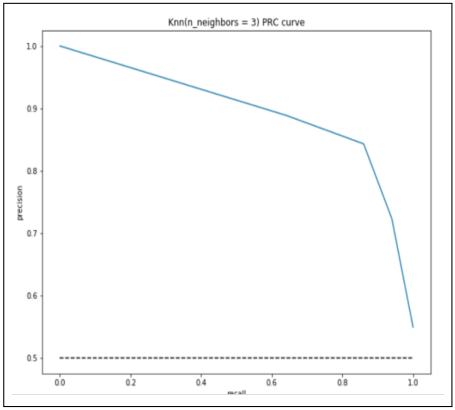


Figure: Relating Pr and Rr

1

[1] https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Relating Pr and Rr

- At the lowest point, i.e., at (0, 0)- the threshold is set at 1.0. That is, the model makes no distinctions between the Spam and the HAM emails.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. That is, the Precision and Recall are high, and the model makes distinctions perfectly.

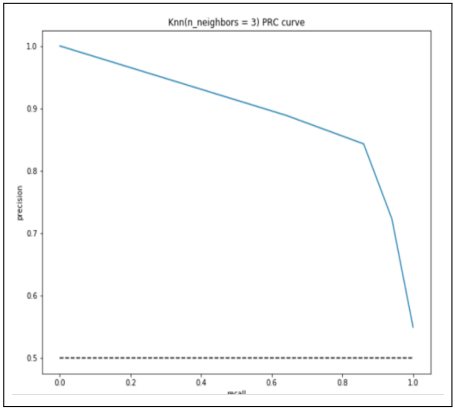


Figure: Relating Pr and Rr

1

[1] https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Relating Pr and Rr

- At the lowest point, i.e., at (0, 0)- the threshold is set at 1.0. That is, the model makes no distinctions between the Spam and the HAM emails.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. That is, the Precision and Recall are high, and the model makes distinctions perfectly.
- The rest of the curve shows the values of Precision and Recall for the threshold values between 0 and 1.

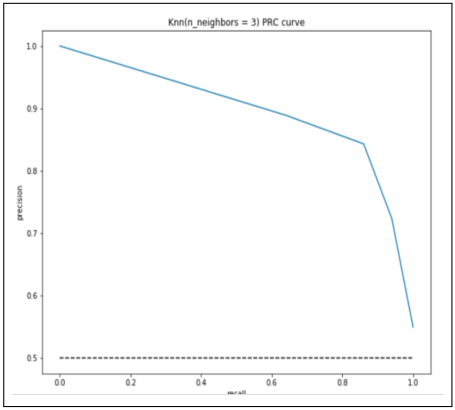


Figure: Relating Pr and Rr

1

[1] https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Relating Pr and Rr

- At the lowest point, i.e., at (0, 0)- the threshold is set at 1.0. That is, the model makes no distinctions between the Spam and the HAM emails.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. That is, the Precision and Recall are high, and the model makes distinctions perfectly.
- The rest of the curve shows the values of Precision and Recall for the threshold values between 0 and 1.
- The aim must be to make the curve as close to (1, 1) as possible - meaning good Precision and Recall.

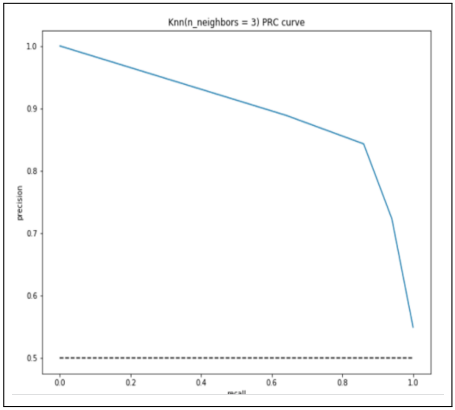


Figure: Relating Pr and Rr

1

[1] https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.
- On the other hand, the efficiency of filters that explicitly estimate the group conditional probabilities and then

Performance evaluation measures: Receiver Operating Characteristics curve

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.
- On the other hand, the efficiency of filters that explicitly estimate the group conditional probabilities and then
  - execute classification based on estimated probabilities, can be represented by a curve called ROC (Receiver Operating Characteristics) curve.

Performance evaluation measures: Receiver Operating Characteristics curve

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.
- On the other hand, the efficiency of filters that explicitly estimate the group conditional probabilities and then
  - execute classification based on estimated probabilities, can be represented by a curve called ROC (Receiver Operating Characteristics) curve.
- ROC curve, is a graphical plot that demonstrates the analytical capability of a Spam filter as its bias level is modified.

Performance evaluation measures: Receiver Operating Characteristics curve

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.
- On the other hand, the efficiency of filters that explicitly estimate the group conditional probabilities and then
  - execute classification based on estimated probabilities, can be represented by a curve called ROC (Receiver Operating Characteristics) curve.
- ROC curve, is a graphical plot that demonstrates the analytical capability of a Spam filter as its bias level is modified.
- The ROC curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.

Performance evaluation measures: Receiver Operating Characteristics curve

- Spam filters with a drastically reduced FPR and FNR are said to have a better performance.
- FNR & FPR represent the efficiency of filters that directly aim at the classification decision borderline devoid of generating the probability estimate.
- On the other hand, the efficiency of filters that explicitly estimate the group conditional probabilities and then
  - execute classification based on estimated probabilities, can be represented by a curve called ROC (Receiver Operating Characteristics) curve.
- ROC curve, is a graphical plot that demonstrates the analytical capability of a Spam filter as its bias level is modified.
- The ROC curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.
- When the ROC curve of a Spam filter closely sits on top of another, such filter can be classified a filter with superior performance in all implementation setups.

- As must be clear, RoC curve is the plot between the TPR(y-axis) and FPR(x-axis).

- As must be clear, RoC curve is the plot between the TPR(y-axis) and FPR(x-axis).
- Since the model must classify the emails as Spam/HAM based on the probabilities generated for each class, one can decide the threshold of the probabilities as well.

- As must be clear, RoC curve is the plot between the TPR(y-axis) and FPR(x-axis).
- Since the model must classify the emails as Spam/HAM based on the probabilities generated for each class, one can decide the threshold of the probabilities as well.
- For example, a threshold value of 0.4 implies that the model will classify the data point/email as a Spam if the probability of an email being a Spam is greater than 0.4.

- As must be clear, RoC curve is the plot between the TPR(y-axis) and FPR(x-axis).
- Since the model must classify the emails as Spam/HAM based on the probabilities generated for each class, one can decide the threshold of the probabilities as well.
- For example, a threshold value of 0.4 implies that the model will classify the data point/email as a Spam if the probability of an email being a Spam is greater than 0.4.
- This will obviously give a high Recall value and reduce the number of False Positives.

- As must be clear, RoC curve is the plot between the TPR(y-axis) and FPR(x-axis).
- Since the model must classify the emails as Spam/HAM based on the probabilities generated for each class, one can decide the threshold of the probabilities as well.
- For example, a threshold value of 0.4 implies that the model will classify the data point/email as a Spam if the probability of an email being a Spam is greater than 0.4.
- This will obviously give a high Recall value and reduce the number of False Positives.
- RoC curve makes these aspects clearer.....see next slide...

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.

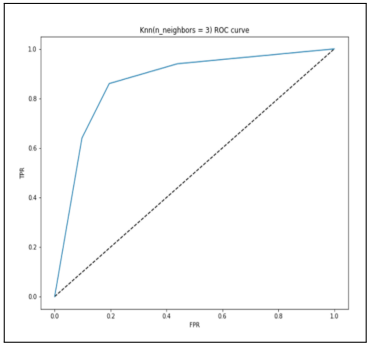


Figure: Relating TPR and FPR

Source:
https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.

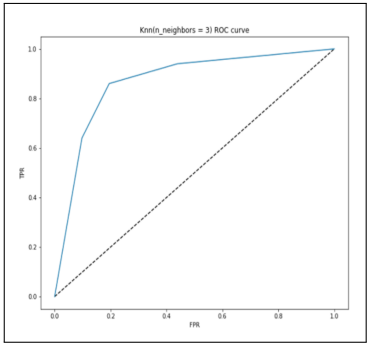


Figure: Relating TPR and FPR

Source:
https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.
- The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1.

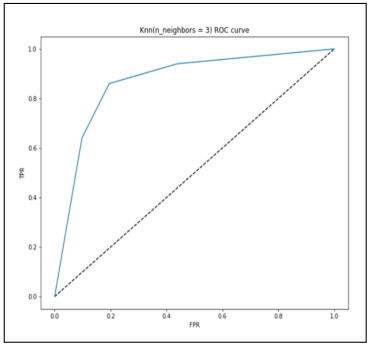


Figure: Relating TPR and FPR

Source:
https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.
- The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1.
- At some threshold values, see that for FPR close to 0, TPR is close to 1. This is when the model will predict the emails as Spam almost perfectly.

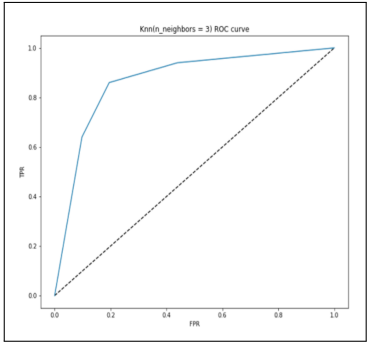


Figure: Relating TPR and FPR

Source:
https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.
- The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1.
- At some threshold values, see that for FPR close to 0, TPR is close to 1. This is when the model will predict the emails as Spam almost perfectly.
- The area with the curve and the axes as the boundaries is called the Area Under Curve(AUC).

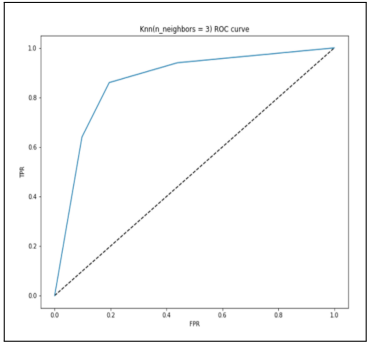


Figure: Relating TPR and FPR

Source: https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.
- The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1.
- At some threshold values, see that for FPR close to 0, TPR is close to 1. This is when the model will predict the emails as Spam almost perfectly.
- The area with the curve and the axes as the boundaries is called the Area Under Curve(AUC).
- AUC is considered as a metric of a good model. With this metric ranging from 0 to 1, one must aim for a high value of AUC.

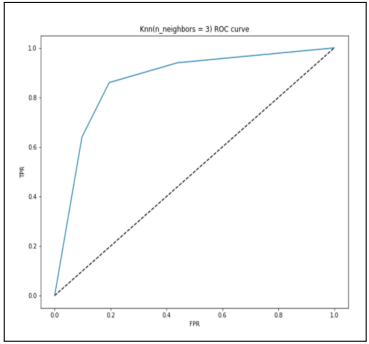


Figure: Relating TPR and FPR

Source: https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: Interpreting RoC

- At the lowest point, i.e., at (0, 0) - the threshold is set at 1.0. Implies that the model classifies all emails as HAM.
- At the highest point, i.e., at (1, 1), the threshold is set at 0.0. Implies that the model classifies all emails as Spam.
- The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1.
- At some threshold values, see that for FPR close to 0, TPR is close to 1. This is when the model will predict the emails as Spam almost perfectly.
- The area with the curve and the axes as the boundaries is called the Area Under Curve(AUC).
- AUC is considered as a metric of a good model. With this metric ranging from 0 to 1, one must aim for a high value of AUC.

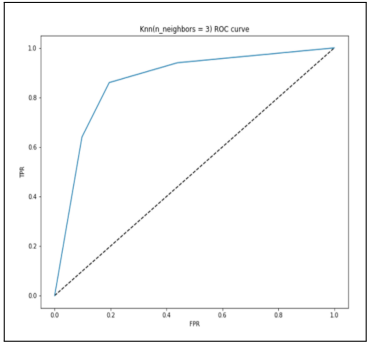


Figure: Relating TPR and FPR

Source: https://www.analyticsvidhya.com/blog/2020/09/Precision-Recall-machine-learning/

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives = (λ) * Cost of False Negatives

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives = (λ) * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives $= (\lambda)$ * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives $= (\lambda)$ * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives = (λ) * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as
  - Weighted Accuracy (WAcc),

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives = (λ) * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as
  - Weighted Accuracy (WAcc),
  - Weighted Error Rate (WErr) and

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives $= (\lambda)$ * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as
  - Weighted Accuracy (WAcc),
  - Weighted Error Rate (WErr) and
  - Total Cost Ratio (TCR)

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives $= (\lambda)$ * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as
  - Weighted Accuracy (WAcc),
  - Weighted Error Rate (WErr) and
  - Total Cost Ratio (TCR)
  - $F - measure\ ORF_1$ Score

Performance evaluation measures: λ

- When the cost of false positives is much more than that of false negatives, the measure of λ is useful.
- Cost of False Positives $= (\lambda)$ * Cost of False Negatives
- This, (λ) is a numerical factor that stipulates how 'risky or 'harmful' it is to wrongly classify a valid e-mail as Spam.
- It also indicates how difficult it can be for the user to recuperate from such abysmal and unacceptable performance of the Spam filter.
- Therefore, Cost sensitivity measures such as
  - Weighted Accuracy (WAcc),
  - Weighted Error Rate (WErr) and
  - Total Cost Ratio (TCR)
  - $F - measure\ ORF_1$ Score
- have been introduced by some authors.

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.
  - Rr=88%, $\implies$ that with $\mathbb{A}$, 12% of the Spam email still makes it into inbox i.e. they are not detected as the Spam. Low Recall is to be avoided but it is not a serious issue with Spam filters.

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.
  - Rr=88%, $\implies$ that with $\mathbb{A}$, 12% of the Spam email still makes it into inbox i.e. they are not detected as the Spam. Low Recall is to be avoided but it is not a serious issue with Spam filters.
- Consider now Akismet (comment Spam filtering service/plugin for WordPress) email Spam filter. It is known as a good Spam filter. Its various values are as follows, for the sake of comparison.

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.
  - Rr=88%, $\implies$ that with $\mathbb{A}$, 12% of the Spam email still makes it into inbox i.e. they are not detected as the Spam. Low Recall is to be avoided but it is not a serious issue with Spam filters.
- Consider now Akismet (comment Spam filtering service/plugin for WordPress) email Spam filter. It is known as a good Spam filter. Its various values are as follows, for the sake of comparison.
  - Acc = 99.88%, Pr = 99.99%, Rr = 99.87%, what does this now imply?

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.
  - Rr=88%, $\implies$ that with $\mathbb{A}$, 12% of the Spam email still makes it into inbox i.e. they are not detected as the Spam. Low Recall is to be avoided but it is not a serious issue with Spam filters.
- Consider now Akismet (comment Spam filtering service/plugin for WordPress) email Spam filter. It is known as a good Spam filter. Its various values are as follows, for the sake of comparison.
  - Acc = 99.88%, Pr = 99.99%, Rr = 99.87%, what does this now imply?
- Note that In both cases most Spam is tagged as Spam.

Performance evaluation measures: An Example

- Consider an email Spam filter $\mathbb{A}$, Acc = 92%, Pr=92% and Rr=88%. What does this imply ?
  - Acc=92% $\implies$ $\mathbb{A}$ makes wrong predictions 8% of the times (not acceptable, normally),
  - Pr=92% $\implies$ that with $\mathbb{A}$, 8% of what was flagged as Spam wasn't actually Spam i.e. an indication of how precise $\mathbb{A}$, was in predicting positive result. 8% imprecision is an unacceptable rate for losing possibly important messages.
  - Rr=88%, $\implies$ that with $\mathbb{A}$, 12% of the Spam email still makes it into inbox i.e. they are not detected as the Spam. Low Recall is to be avoided but it is not a serious issue with Spam filters.
- Consider now Akismet (comment Spam filtering service/plugin for WordPress) email Spam filter. It is known as a good Spam filter. Its various values are as follows, for the sake of comparison.
  - Acc = 99.88%, Pr = 99.99%, Rr = 99.87%, what does this now imply?
- Note that In both cases most Spam is tagged as Spam.
- In Spam filtering, Precision is emphasized over Recall. This is appropriate, because it's more important to not lose non-Spam email than it is to filter every single piece of Spam out of inbox.

Performance evaluation measures: Motivation for F1 Score

Thus, important to remember

Precision is a measure of confirmation (when the classifier indicates positive, how often it's correct), and Recall is a measure of utility (how much the classifier finds of what there is to find).

Performance evaluation measures: Motivation for F1 Score

Thus, important to remember

Precision is a measure of confirmation (when the classifier indicates positive, how often it's correct), and Recall is a measure of utility (how much the classifier finds of what there is to find).

Deciding over two Spam filters

Suppose one had multiple Spam filters to choose from, each with different values of Precision and Recall. How to pick the Spam filter to use?

Performance evaluation measures: Motivation for F1 Score

Thus, important to remember

Precision is a measure of confirmation (when the classifier indicates positive, how often it's correct), and Recall is a measure of utility (how much the classifier finds of what there is to find).

Deciding over two Spam filters

Suppose one had multiple Spam filters to choose from, each with different values of Precision and Recall. How to pick the Spam filter to use?

In situations like this, it is preferable to have one number to compare all the different choices by. Such score is the F1 score.
The F1 score measures a tradeoff between Precision and Recall.
It's defined as the harmonic mean of the Precision and Recall. This is most easily shown with an explicit calculation.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters
  - a HAM should not be tagged as a Spam even when a Spam is predicted as a Ham.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters
  - a HAM should not be tagged as a Spam even when a Spam is predicted as a Ham.
- In other situations, Precision and Recall are equally important.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters
  - a HAM should not be tagged as a Spam even when a Spam is predicted as a Ham.
- In other situations, Precision and Recall are equally important.
- In such cases, F1-score is useful. F1-score is the Harmonic mean of the Precision and Recall.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters
  - a HAM should not be tagged as a Spam even when a Spam is predicted as a Ham.
- In other situations, Precision and Recall are equally important.
- In such cases, F1-score is useful. F1-score is the Harmonic mean of the Precision and Recall.
- In such cases, F1-score is useful. F1-score is the Harmonic mean of the Precision and Recall.

Performance evaluation measures: F-measure OR F_1 Score

- For applications like Medical diagnosis: achieving a high Recall is more important than getting a high Precision - as one would like to detect as many heart patients as possible.
- For applications like Banking & Spam:
  - classifying whether or not a bank customer is a loan defaulter, it is desirable to have high Precision since the bank wouldn't want to lose customers who were denied a loan based on the model's prediction that they would be defaulters
  - a HAM should not be tagged as a Spam even when a Spam is predicted as a Ham.
- In other situations, Precision and Recall are equally important.
- In such cases, F1-score is useful. F1-score is the Harmonic mean of the Precision and Recall.
- In such cases, F1-score is useful. F1-score is the Harmonic mean of the Precision and Recall.
- is also considered to be a measure of the accurateness of a test and is described as the weighted harmonic mean of the Precision (Ps) and Recall (Rs) of the test in a single equation.

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has
  - the value of one when a classifier has perfect Precision and Recall

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has
  - the value of one when a classifier has perfect Precision and Recall
  - the value of zero when classifiers have either low Precision or Recall (or both).

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has
  - the value of one when a classifier has perfect Precision and Recall
  - the value of zero when classifiers have either low Precision or Recall (or both).
- As a designer, if one thinks Spam filter is losing too much real email, and wants to make it *pickier* about marking email as Spam; the way out is to increase its Precision.

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has
  - the value of one when a classifier has perfect Precision and Recall
  - the value of zero when classifiers have either low Precision or Recall (or both).
- As a designer, if one thinks Spam filter is losing too much real email, and wants to make it *pickier* about marking email as Spam; the way out is to increase its Precision.
- Quite often, increasing the Precision of a classifier also lowers its Recall. Then, a *pickier* Spam filter may also mark fewer real Spam emails as Spam i.e. mark it as a HAM.

Performance evaluation measures: F-measure OR F_1 Score

- F-measure makes use of a parameter that enables a compromise to be reached concerning Recall and Precision.

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

- instead of balancing Precision and Recall, we can just aim for a good F1-score, which would also indicate good Precision and a good Recall value.
- F1 has
  - the value of one when a classifier has perfect Precision and Recall
  - the value of zero when classifiers have either low Precision or Recall (or both).
- As a designer, if one thinks Spam filter is losing too much real email, and wants to make it *pickier* about marking email as Spam; the way out is to increase its Precision.
- Quite often, increasing the Precision of a classifier also lowers its Recall. Then, a *pickier* Spam filter may also mark fewer real Spam emails as Spam i.e. mark it as a HAM.
- If the filter's Recall falls too low as its Precision increases, this results in a lower F1. IT signifies that one has traded off too much Recall for better Precision.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].
- this measure allows the performance of an anti-Spam filter to be compared easily to that of the baseline

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].
- this measure allows the performance of an anti-Spam filter to be compared easily to that of the baseline
- higher the value implies better performance. When the value of TCR ¡ 1, the filter performance is said to be very poor, to the extent its use is not advisable i.e. the baseline (not using the filter) is better.

$$TCR = \frac{N_s}{\lambda(|H \rightarrow S|) + |S \rightarrow H|}$$

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].
- this measure allows the performance of an anti-Spam filter to be compared easily to that of the baseline
- higher the value implies better performance. When the value of TCR ¡ 1, the filter performance is said to be very poor, to the extent its use is not advisable i.e. the baseline (not using the filter) is better.

$$TCR = \frac{N_s}{\lambda(|H \rightarrow S|) + |S \rightarrow H|}$$

- Interestingly the authors also state that *if TCR is proportional to wasted time, an intuitive meaning for TCR is the following:*

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].
- this measure allows the performance of an anti-Spam filter to be compared easily to that of the baseline
- higher the value implies better performance. When the value of TCR ¡ 1, the filter performance is said to be very poor, to the extent its use is not advisable i.e. the baseline (not using the filter) is better.

$$TCR = \frac{N_s}{\lambda(|H \rightarrow S|) + |S \rightarrow H|}$$

- Interestingly the authors also state that *if TCR is proportional to wasted time, an intuitive meaning for TCR is the following:*
  - *TCR measures how much time is wasted to delete manually all Spam messages when no filter is used, as compared to the time wasted to delete manually any Spam messages that passed the filter, plus the time needed to recover from mistakenly blocked legitimate messages.*

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: Total Cost Ratio (TCR)

Total Cost Ratio (TCR)

- is used to measure the accuracy of filters, proposed in [1].
- this measure allows the performance of an anti-Spam filter to be compared easily to that of the baseline
- higher the value implies better performance. When the value of TCR ¡ 1, the filter performance is said to be very poor, to the extent its use is not advisable i.e. the baseline (not using the filter) is better.

$$TCR = \frac{N_s}{\lambda(|H \rightarrow S|) + |S \rightarrow H|}$$

- Interestingly the authors also state that *if TCR is proportional to wasted time, an intuitive meaning for TCR is the following:*
  - *TCR measures how much time is wasted to delete manually all Spam messages when no filter is used, as compared to the time wasted to delete manually any Spam messages that passed the filter, plus the time needed to recover from mistakenly blocked legitimate messages.*
- the only apparent strength is that : it is a single-figure measurement, while majority of the other cost sensitive measures require a minimum of two figures.

[1] I. Androutsopoulos et al, "An experimental comparison of naïve Bayesian and keyword-based anti-Spam filtering with personal e-mail messages", In: Proc of the Ann Int ACM SIGIR Conf on Res and Devel in Inform Retrieval, 2000.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

- These cost-sensitive formulae use the λ parameter to adjust the weight of a false positive.

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

- These cost-sensitive formulae use the λ parameter to adjust the weight of a false positive.
- three values of λ used commonly in the literature: 1, 9 and 999. With the cost of a false positive C_{FP} and the cost of false negative C_{FN}

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

- These cost-sensitive formulae use the λ parameter to adjust the weight of a false positive.
- three values of λ used commonly in the literature: 1, 9 and 999. With the cost of a false positive C_{FP} and the cost of false negative C_{FN}
  - $\lambda = 1$ implies that $C_{FP} = C_{FN}$ ...then the Spam messages are simply marked as Spam.

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

- These cost-sensitive formulae use the λ parameter to adjust the weight of a false positive.
- three values of λ used commonly in the literature: 1, 9 and 999. With the cost of a false positive C_{FP} and the cost of false negative C_{FN}
  - $\lambda = 1$ implies that $C_{FP} = C_{FN}$ ...then the Spam messages are simply marked as Spam.
  - $\lambda = 9$ implies that $C_{FP} = 9 * C_{FN}$ . .....the Spam messages are returned to their senders.

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: W_{Acc}, W_{Err}

These two measures were introduced in the [1]

- Weighted Accuracy (W_{Acc}) is given by

$$W_{Acc} = \frac{\lambda(|H \rightarrow H|) + |S \rightarrow S|}{\lambda N_H + N_S}$$

- Weighted Error Rate (W_{Err}) is given by

$$W_{Err} = \frac{\lambda(|H \rightarrow S|) + |S \rightarrow H|}{\lambda N_H + N_S}$$

- These cost-sensitive formulae use the λ parameter to adjust the weight of a false positive.
- three values of λ used commonly in the literature: 1, 9 and 999. With the cost of a false positive C_{FP} and the cost of false negative C_{FN}
  - $\lambda = 1$ implies that $C_{FP} = C_{FN}$ ...then the Spam messages are simply marked as Spam.
  - $\lambda = 9$ implies that $C_{FP} = 9 * C_{FN}$ . .....the Spam messages are returned to their senders.
  - $\lambda = 999$ implies that $C_{FP} = 999 * C_{FN}$..... tthe Spam messages are deleted immediately.

[1] Clark, K.P., 2008. A survey of content-based Spam classifiers. Google Scholar, pp.1-19. Citeseer.

Performance evaluation measures: Sensitivity with λ values

- In computing cost sensitivity of filters, λ decides the strictness of penalty for wrongly classifying a non-Spam email as Spam.
- Table shows different strata of cost sensitivity of model that have been taken into consideration, seen as follows:
  - the efficiency of a filter for a given λ is compared with a baseline case by means of total cost ratio as explained in.
  - λ is used for fine-tuning the weight of false positive and its performance is assessed by the cost sensitivity.
  - the three values used for λ and their significance is as discussed before, shown again here.

Table 3
Levels of cost sensitivity of model.

λ	Maximum Tolerance Level $T = \lambda/(1+\lambda)$	Significance of having such cost sensitivity?
999	0.999	Filtered messages are thrown away and no additional processing is carried out.
99	0.9	Filtering a non-spam message is slightly penalized above allowing a spam message to go through. It is used to demonstrate that it is cumbersome re-sending a filtered spam message than deleting it manually.
1	0.5	If the receiver is not concerned as regards missing a non-spam message.

Source: Emmanuel Gbenga Dada et al, Machine learning for email Spam filtering: review, approaches and open research problems, Heliyon, 5(6), 2019

ML-based Methods used for Email Spam filtering

Blank Slide