Email Spam Detection

By john kutney

Objective

Automatically classify emails as spam or legitimate (ham) and

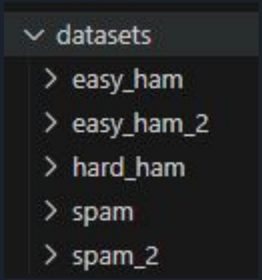
Prioritize important emails based on content

Dataset

SpamAssassin Public Corpus

https://spamassassin.apache.org/old/publiccorpus/

Emails labeled as



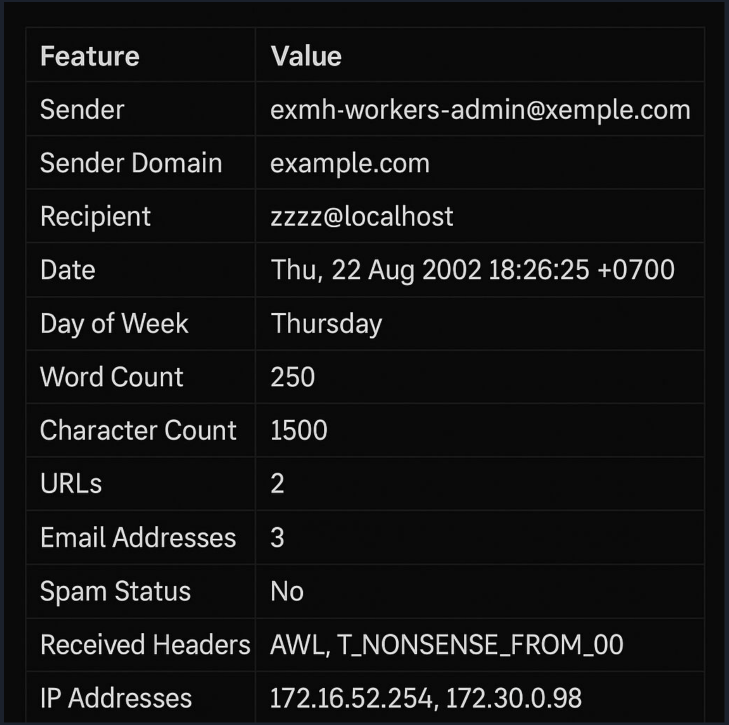
```
⌄ datasets
  › easy_ham
  › easy_ham_2
  › hard_ham
  › spam
  › spam_2
```

Real-world dataset with raw headers and bodies

Size: ~6,000+ emails

Features

EXAMPLE : in spirit not form



Feature	Value
Sender	exmh-workers-admin@xemple.com
Sender Domain	example.com
Recipient	zzzz@localhost
Date	Thu, 22 Aug 2002 18:26:25 +0700
Day of Week	Thursday
Word Count	250
Character Count	1500
URLs	2
Email Addresses	3
Spam Status	No
Received Headers	AWL, T_NONSENSE_FROM_00
IP Addresses	172.16.52.254, 172.30.0.98

Header Features

Sender Information:

- From: The sender's email address (e.g., exmh-workers-admin@example.com).
- Sender Domain: The domain of the sender's email (e.g., example.com).

Recipient Information:

- To: The recipient's email address (e.g., zzzz@localhost.netnoteinc.com).

Date and Time:

- Date: The date and time the email was sent (e.g., Thu, 22 Aug 2002 18:26:25 +0700).
- Day of Week: The day of the week the email was sent (e.g., Thursday).
- Hour of Day: The hour the email was sent (e.g., 18).

Message ID:

- Message-Id: A unique identifier for the email (e.g., <13258.1030015585@munnari.OZ.AU>).

Mailing List Information:

- List-Id: The mailing list identifier (e.g., exmh-workers.example.com).
- List-Subscribe: The subscription URL or email address (e.g., <https://listman.example.com/mailman/listinfo/exmh-workers>).

Content Features

Word Count:

- Total number of words in the email body.

Character Count:

- Total number of characters in the email body.

Average Word Length:

- Average length of words in the email body.

Special Characters:

- Count of special characters (e.g., @, #, $, %, etc.).

URLs:

- Count of URLs in the email body (e.g., https://listman.example.com/mailman/listinfo/exmh-workers).

Email Addresses:

- Count of email addresses mentioned in the body.

Capitalization:

- Percentage of words in all caps (e.g., I can't reproduce this error).

Linguistic Features

Spam Keywords:

- Presence of common spam keywords (e.g., unsubscribe, click, free, win, etc.).

Readability:

- Readability score of the email (e.g., Flesch-Kincaid readability score).

N-grams:

- Common bigrams or trigrams in the email body (e.g., pick command, sequence mercury).

Sentiment:

- Sentiment score of the email body (positive, negative, or neutral).

Structural Features

Number of Lines:

- Total number of lines in the email body.

Blank Lines:

- Count of blank lines in the email body.

Quoted Text:

- Count of lines starting with > (indicating quoted text).

Attachments:

- Presence of attachments (e.g., MIME types).

Network Features

Received Headers:

- Count of Received headers (indicating the number of hops the email took).

IP Addresses:

- Extracted IP addresses from the Received headers (e.g., 172.16.52.254).

Geolocation:

- Geolocation of the IP addresses (e.g., Thailand for 172.30.0.98).

Spam-Specific Features

Spam Status:

- X-Spam-Status: Whether the email is marked as spam (e.g., No).

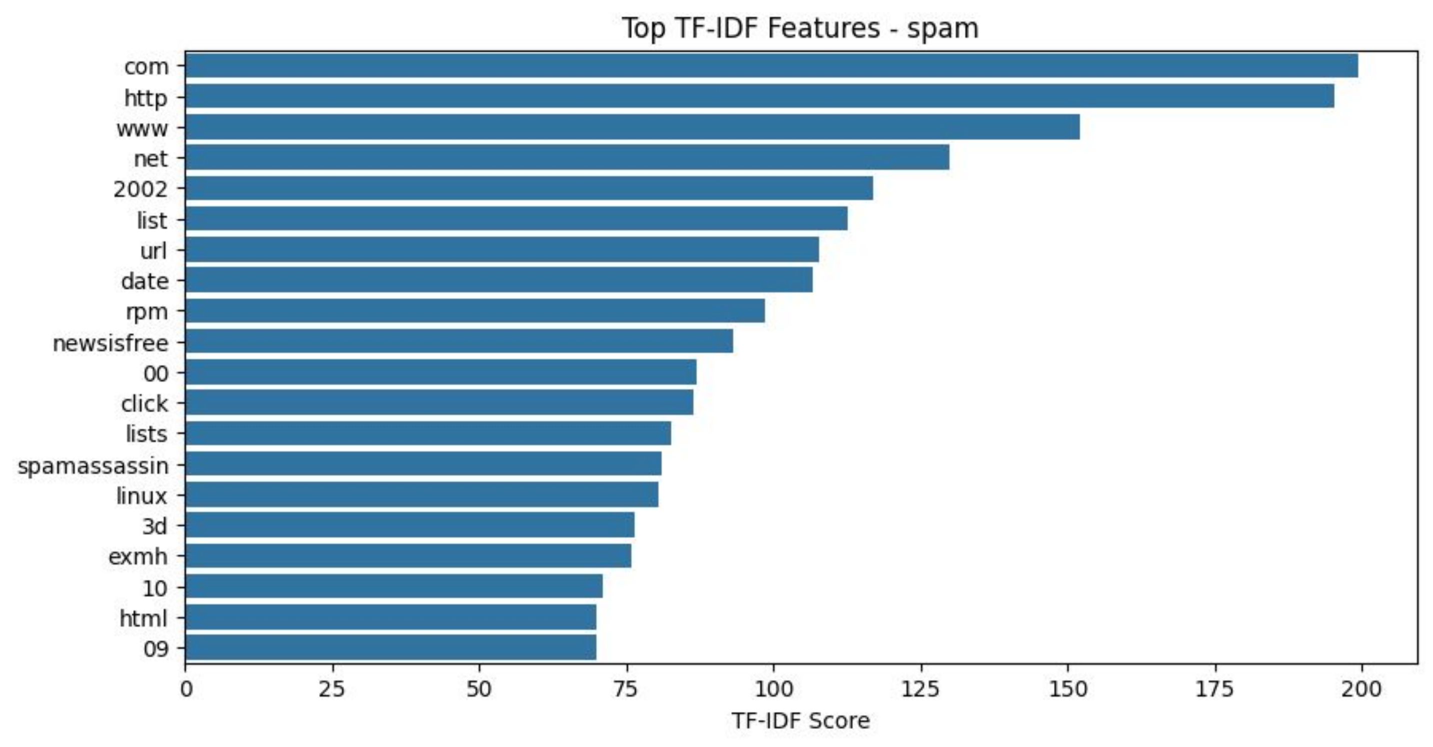
Spam Score:

- X-Spam-Level: The spam score of the email (e.g., hits=-988.3 required=5.0).

Spam Tests:

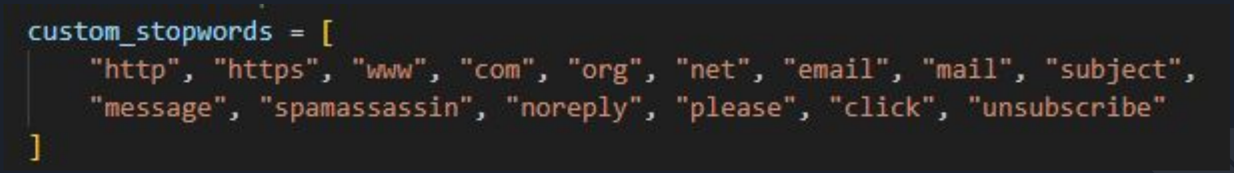
- List of spam tests triggered (e.g., AWL, T_NONSENSE_FROM_00_10).

Issues


Top TF-IDF Features - spam
com
http
www
net
2002
list
url
date
rpm
newsisfree
00
click
lists
spamassassin
linux
3d
exmh
10
html
09
0
25
50
75
100
125
150
175
200
TF-IDF Score

FIX:

filtering out the words that shouldn’t count



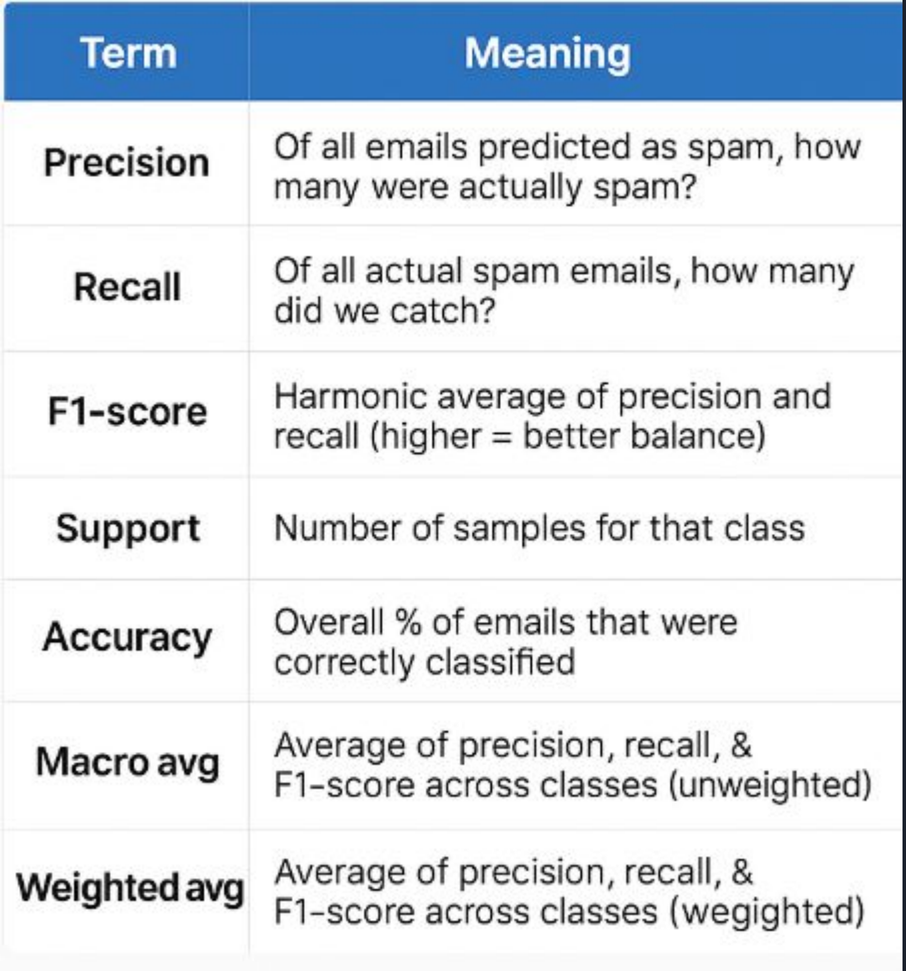
```
custom_stopwords = [
    "http", "https", "www", "com", "org", "net", "email", "mail", "subject",
    "message", "spamassassin", "noreply", "please", "click", "unsubscribe"
]
```

Outcome

A spam classifier with >90% accuracy

Insight into which terms or senders are spammy

Visualizations showing model performance and key contributors



Term	Meaning
Precision	Of all emails predicted as spam, how many were actually spam?
Recall	Of all actual spam emails, how many did we catch?
F1-score	Harmonic average of precision and recall (higher = better balance)
Support	Number of samples for that class
Accuracy	Overall % of emails that were correctly classified
Macro avg	Average of precision, recall, & F1-score across classes (unweighted)
Weighted avg	Average of precision, recall, & F1-score across classes (wegighted)

Thank you