***Sistematik Review :***

**Metode Penyaringan Spam pada E-mail Menggunakan Machine Learning**

Muhammad Hafidz

Mahasiswa Sarjana Teknik Informatika, Fakultas Ilmu Komputer,

Universitas Dian Nuswantoro. Email : ciazhar.id@gmail.com

ABSTRAK

**Latar Belakang** : Email merupakan media komunikasi bagi pengguna dan penyedia jasa Internet yang efektif. Perusahaan-perusahaan besar mayoritas menggunakan email sebagai media komunikasi dengan dengan para pelangganya. Namun tidak semua email yang dikirim dapat sampai ke kotak masuk email para pelanggan. Banyak faktor yang mempengaruhi hal tersebut, diantaranya karena konten yang tidak sesuai kaidah penulisan yang baik, alamat email yang tidak valid, domain pengguna yang terdaftar dalam Blacklist dan sebagainya.

**Tujuan** : Untuk mengetahui metode metode machine learning yang dapat digunakan untuk menyaring spam pada email.

**Metode** : Sebuah tinjauan sistematis melalui review jurnal mengenai metode penyaring spam pada email. Pencarian artikel diakses dari beberapa penerbit yaitu : ACM, Arxiv, IEEE, Science Direct dan Springer dengan masing masing penerbit sebanyak 10 jurnal. Dari *review* literatur ditemukan beberapa metode yang efektif untuk menyaring spam pada email. Hasil review literatur mulai dari tahun 2000 – 2018, keseluruhan jurnal berasal dari luar negeri dan dalam bahasa inggris.

**Hasil** : Pada *systematic review* ini menunjukan bahwa penyaringan spam pada email dapat dilakukan dengan beberapa metode machine learning diantaranya adalah SVM, NB, NSA dll.

**Kesimpulan** : Metode-metode Machine Learning dapat menyaring spam pada email

Kata kunci: Filter spam, Email.

**Pendahuluan**

Email merupakan media komunikasi bagi pengguna dan penyedia jasa Internet yang efektif. Perusahaan-perusahaan besar mayoritas menggunakan email sebagai media komunikasi dengan dengan para konsumenya. Namun tidak semua email yang dikirim dapat sampai ke kotak masuk email para konsumen. Banyak faktor yang mempengaruhi hal tersebut, diantaranya karena konten yang tidak sesuai kaidah penulisan yang baik, alamat email yang tidak valid, domain pengguna yang terdaftar dalam Blacklist dan sebagainya.

Spam tidak hanya mengganggu tetapi juga membahayakan konsumen. Email marketer memiliki kewajiban untuk mengirimkan email atau newsletter dari konsumen ke para subscriber. Hal ini akan sangat mengganggu karena mereka harus berjuang untuk melewati spam filter.

Domain Blacklist merupakan daftar alamat IP yang dicurigai mengirim spam dan digunakan untuk mencegah pesan email yang tidak diinginkan mencapai penerima yang tidak curiga. Satu hal yang penting untuk disebutkan adalah blacklist itu sebenarnya tidak memblokir pesan Anda, tetapi justru penyedia kotak surat yang sebenarnya. Penyedia ini menggunakan informasi ini dari berbagai layanan daftar hitam bersama dengan metrik internal untuk membuat keputusan tentang apakah atau tidak untuk memblokir pesan.

## **Surat Elektronik**

Surat elektronik (akronim: ratel, ratron, surel, atau surat-e) atau pos elektronik (akronim: pos-el.) atau imel (bahasa Inggris: *email*) adalah sarana kirim mengirim surat melalui jalur jaringan komputer (misalnya Internet).

Struktur alamat surel, sebagai contoh:

|  |
| --- |
| surelsaya@surabaya.vibriel.net.id |

Keterangan:

* surelsaya: nama kotak surat (*mailbox*) atau nama pengguna (*username*) yang ingin dituju dalam *mailserver*
* surabaya.vibriel.net.id: nama *mailserver* tempat pengguna yang dituju, rinciannya:
  + surabaya: *subdomain* (milik pemegang nama *domain*), biasanya merujuk ke suatu komputer dalam lingkungan pemilik *domain*
  + vibriel: nama *domain*, biasanya menunjukkan nama perusahaan/organisasi/perorangan (Vibriel)
  + net: *second level domain*, menunjukkan bahwa *domain* ini termasuk kategori *networking* (net)
  + id: *top level domain*, menunjukkan bahwa *domain* ini terdaftar di otoritas *domain* Indonesia (id)

Untuk mengirim surat elektronik kita memerlukan suatu program *mail-client*. Surat elektronik yang kita kirim akan melalui beberapa poin sebelum sampai di tujuan. Mulai dari surat elektronik dikirim → Internet → POP3 server penyedia surel penerima → *mail client* (di komputer si penerima) → surat elektronik dibaca si penerima

Terlihat surat elektronik yang terkirim hanya melalui 5 poin (selain komputer pengirim dan penerima). Sebenarnya lebih dari itu sebab setelah surat elektronik meninggalkan POP3 Server maka itu akan melalui banyak server-server lainnya. Tidak tertutup kemungkinan surat elektronik yang kita kirim disadap orang lain. Maka dari itu bila surat elektronik yang kita kirim mengandung isi yang sensitif sebaiknya kita melakukan tindakan pencegahan, dengan mengacak (enkripsi) data dalam surat elektronik tersebut (contohnya menggunakan PGP, sertifikat digital, dan lain-lain).

Post Office Protocol 2 (POP3) adalah sebuah mail access protocol yang digunakan oleh aplikasi klien untuk membaca pesan dari mail server. Pesan yang diterima biasanya dihapus dari server. POP mendukung fitur unduh dan hapus untuk mengakses kotak masuk.

**Struktur Email**

E-mail terdiri dari dua elemen: body dan header. Body email terdiri dari data yang tidak terstruktur seperti teks, markup HTML, objek multimedia dan lampiran. Header terdiri dari informasi penelusuran dan bidang terstruktur yang merupakan bagian dari konten pesan. Simple Mail Transfer Protocol (SMTP) mendefinisikan sesi header email berisi bidang seperti - subjek, nama pengirim, ID email, tanggal pengiriman, informasi perutean, cap waktu, dll. untuk informasi penerima dan pengiriman yang berhasil. Setiap atribut (bidang) di header memiliki nama dan makna tertentu,

* Received : Berisi informasi transit terkait dari server email, alamat IP, tanggal, dll.
* From : Nama pengirim; ID email. "Nama" e-mail@example.com
* To : Nama Penerima; ID email. "Nama" e-mail@example.com
* Return Path : Alamat opsional yang akan digunakan jika terjadi kesalahan (bounce).
* Message ID : Pengidentifikasi pesan tunggal unik yang ditunjuk oleh sistem surat.
* X-mailer : Perangkat lunak email yang digunakan untuk membuat / mengirim pesan.
* Subjek : String mengidentifikasi tema pesan yang ditempatkan oleh pengirim.
* Jenis konten: Format konten (rangkaian karakter, dll.), Ditentukan oleh MIME (Multipurpose Internet Mail Extension).

Setiap pesan email terdiri dari amplop penanganan transit yang disembunyikan dari pengguna email. Pertama alamat pengirim amplop dikirim, diikuti oleh satu atau lebih alamat penerima amplop, dan akhirnya pesan yang sebenarnya dikirim. Server e-mail sebenarnya menggunakan alamat amplop (bukan alamat tajuk pesan) untuk mengirimkan e-mail ke penerima yang benar. Penerima akhir hanya melihat header dan badan email. Alamat amplop adalah salah satu fitur email yang sangat sering disalahgunakan oleh spammer.

**Email spam**

Spam email, juga dikenal sebagai email sampah, adalah pesan yang tidak diminta yang dikirim secara massal melalui email (spamming). Spam email terus berkembang sejak awal 1990-an, dan pada 2014 diperkirakan sekitar 90% pesan email terkirim. Karena biaya spam sebagian besar ditanggung oleh penerima, maka secara efektif ongkos kirim karena iklan. Ini menjadikannya contoh luar biasa dari eksternalitas negatif. Definisi hukum dan status spam bervariasi dari satu yurisdiksi ke yurisdiksi lain, tetapi undang-undang dan tuntutan hukum tidak pernah berhasil membendung spam.

## Sebagian besar pesan spam email bersifat komersial. Baik komersial atau tidak, banyak yang tidak hanya menjengkelkan, tetapi juga berbahaya karena mengandung tautan yang mengarah ke situs web phishing atau situs yang meng-hosting malware - atau memasukkan malware sebagai lampiran file.

## Spammer mengumpulkan alamat email dari ruang obrolan, situs web, daftar pelanggan, newsgroup, dan virus yang memanen buku alamat pengguna. Alamat email yang dikumpulkan ini terkadang juga dijual kepada pengirim spam lainnya.

## **Bounce Message**

Bounce Message adalah email otomatis yang memberi tahu pengirim pesan sebelumnya bahwa pesan itu belum terkirim (atau masalah pengiriman lainnya terjadi). Pengirim terkadang menerima bounce message dari server emailnya sendiri, melaporkan bahwa ia tidak dapat mengirim pesan, atau meskipun telah menerima pesan tersebut, sekarang pesan itu tidak dapat dikirim, juga menerima tanggung jawab untuk mengirimkan DSN jika pengiriman gagal. Karena berbagai alasan, terutama spam dan email virus, pengguna mungkin menerima pesan pentalan yang salah dikirim sebagai tanggapan terhadap pesan yang sebenarnya tidak pernah mereka kirim.[4]

**Machine Learning**

Machine Learning (ML) adalah studi ilmiah tentang algoritma dan model statistik yang digunakan sistem komputer untuk secara progresif meningkatkan kinerjanya pada tugas tertentu. Algoritma Machine Learning membangun model matematika dari data sampel, yang dikenal sebagai "data pelatihan", untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit untuk melakukan tugas tersebut. Algoritma Machine Learning digunakan dalam aplikasi penyaringan email, deteksi penyusup jaringan, dan visi komputer, di mana tidak mungkin untuk mengembangkan algoritma instruksi khusus untuk melakukan tugas. Machine Learning terkait erat dengan statistik komputasi, yang berfokus pada membuat prediksi menggunakan komputer. Studi tentang optimasi matematika memberikan metode, teori dan domain aplikasi ke bidang Machine Learning. Penambangan data adalah bidang studi dalam Machine Learning, dan berfokus pada analisis data eksplorasi melalui pembelajaran tanpa pengawasan. Dalam penerapannya di seluruh masalah bisnis, Machine Learning juga disebut sebagai analytics prediktif.

**Metode**

Sebuah tinjauan sistematis melalui review jurnal mengenai metode penyaring spam pada email. Pencarian artikel diakses dari beberapa penerbit yaitu : ACM, Arxiv, IEEE, Science Direct dan Springer dengan masing masing penerbit sebanyak 10 jurnal. Hasil review literatur mulai dari tahun 2000 – 2018, keseluruhan jurnal berasal dari luar negeri dan dalam bahasa inggris. Dari *review* literatur ditemukan beberapa metode yang efektif untuk menyaring spam pada email. Metode – metode tersebut dikelompokkan kedalam beberapa basis yaitu Reputation-Based dan Content-based. Berikut gambaran Basis – basis beserta metode – metodenya :

****

**Hasil dan pembahasan**

Berdasarkan hasil pencarian, didapatkan sejumlah 50 jurnal yang dianggap sesuai dengan tujuan penelitian. Adapun jurnal - jurnal tersebut dapat dilihat pada tabel yang ada pada lampiran di file excel. Dari jurnal-jurnal tersebut kemudian dikumpulkan untuk dilakukan screening berdasarkan Research Question (RQ). Adapun RQ tersebut dapat dilihat pada tabel berikut :

**Tabel 2. RQ (Research Question)**

|  |  |  |
| --- | --- | --- |
| No | ID | Tujuan |
| 1 | RQ 1 | Teknik atau metode seperti apa sajakah yang dapat menyaring spam ? |
| 2 | RQ 2 | Bagian Email mana saja yang dapat dijadikan patokan untuk menyaring spam ? |
| 3 | RQ 3 | Bagaimana cara mengurangi spam masuk ke inbox email dengan Machine Learning ? |

Selain Research Question (RQ), dilakukan juga screening berdasarkan Search String. Adapun Search String tersebut dapat dilihat pada tabel berikut :

**Tabel 3. Search String**

|  |  |  |
| --- | --- | --- |
| No | Populasi | Golongan |
| 1 | (Metode) AND (Penyaringan) AND (Spam) AND (pada) AND (email) AND (menggunakan) AND (Machine Learning) | (Cara OR desain OR gaya OR jalan OR kaidah OR modus operandi OR organisasi OR pola OR program OR prosedur OR proses OR saluran OR system OR struktur OR tata cara OR teknik OR trik) AND (Filtrasi OR pembersihan OR pemilahan OR pemilihan OR penapisan OR penjernihan OR penyortiran OR purifikasi OR seleksi) AND (Spam) AND (Ala OR atas OR cukup OR di OR lega OR lumayan OR plong OR puas OR sedang OR tenang) AND (Email) AND (memakai OR memanfaatkan OR membonceng OR memerlukan OR menghabiskan OR mengonsumsi OR menyedot OR nunggangi) AND (Machine Learning)   |  |  | | --- | --- | |  |  | |

Setelah dilakukan proses screening didapatkan sejumlah jurnal. Adapun jurnal - jurnal tersebut dapat dilihat pada tabel berikut :

**Tabel 3. Jurnal Terseleksi**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | Penulis/Tahun | Judul | Penerbit | Tujuan | Solusi/Metode |
| 1 | Mengjun Xie 2006 | An Effective Defense Against Email Spam Laundering | ACM | Mendeteksi dan menghentikan aktivitas spam di dalam jaringan pelanggan. | 1. Recipient-oriented Techniques  * Content Based Techniques, example : Email address filters, Heuristic filters, Machine learning based filters * Non-content-based Techniques, example : DNSBLs , MARID, Challenge-Response (C-R), Tempfailing, Delaying, Sender Behavior Analysis  1. Sender-oriented Techniques   Example : Usage Regulation, Cost-based approaches   1. HoneySpam |
| 2 | Gilad Mishne  2005 | Blocking Blog Spam with Language Model Disagreement | ACM | Mendeteksi spam tautan pada komentar blog. | Membandingkan model bahasa yang digunakan dalam posting blog, komentar, dan halaman yang ditautkan oleh komentar menggunakan In- terpolated Aggregate Smoothing |
| 3 | José María Gómez Hidalgo  2016 | Content Based SMS Spam Filtering | ACM | Menganalisis sejauh mana teknik penyaringan Bayesian yang digunakan untuk memblokir email spam, dapat diterapkan pada masalah mendeteksi dan menghentikan spam mobile. | Spam Email Filtering   * White and black listing * Postage * Address management * Collaborative filtering * Digital signatures * Content-based filtering   Bayesian Filtering Techniques   * Preprocessing * Tokenization * Representation * Selection * Learning   Machine Learning Algorithms   * Naive Bayes (NB) * C4.5 * PART * Support Vector Machines (SVM) |
| 4 | Tiago A. Almeida  2011 | Contributions to the Study of SMS Spam Filtering: New Collection and Results | ACM | Membandingkan kinerja yang dicapai oleh beberapa metode pembelajaran mesin yang sudah mapan | * Basic Na ̈ıve Bayes (NB) – Basic NB * Multinomial term frequency NB – MN TF NB * Multinomial Boolean NB – MN Bool NB * Multivariate Bernoulli NB – Bern NB * Boolean NB – Bool NB * Multivariate Gauss NB – Gauss NB * Flexible Bayes – Flex NB * Boosted NB * Linear Support Vector Machine – SVM * Minimum Description Length – MDL * K-Nearest Neighbors – KNN (K = 1, 3 or 5) * C4.5 * Boosted C4.5 * PART |
| 5 | Gordon V. Cormack  2007 | Feature Engineering for Mobile (SMS) Spam Filtering | ACM | Membuktikan bahwa teknik penyaringan email memerlukan beberapa adaptasi untuk mencapai tingkat kinerja yang baik pada spam SMS, terutama mengenai representasi pesan. | * Bogofilter * DMC – Dynamic Markov Compression * LR – TR-IRLS * OSBF-Lua * SVM – SVMlight |
| 6 | Wuying Liu  2010 | Multi-field Learning for Email Spam Filtering | ACM | Memecah masalah Multi-field Document Text Classification (TC) menjadi beberapa masalah sub-dokumen, dan membuat prediksi kategori akhir dengan kombinasi linear tertimbang dari beberapa hasil sub-dokumen TC. | Multi-field learning (MFL) framework |
| 8 | D. Sculley  2007 | Relaxed Online svms for Spam Filtering | ACM | Pertama, kami menunjukkan bahwa SVM online memang memberikan kinerja klasifikasi canggih pada penyaringan spam online pada kumpulan data benchmark besar. Kedua, kami menunjukkan bahwa kinerja yang hampir setara dapat dicapai oleh SVM Santai Online (ROSVM) dengan biaya yang jauh berkurang. | * Online SVM * Relaxed Online SVM (ROSVM) |
| 10 | Gordon V. Cormack  2007 | Spam Filtering for Short Messages | ACM | Menggunakan tiga korpora pesan pendek dan bidang pesan yang berasal dari SMS, blog, dan pesan spam asli, kemudian mengevaluasi filter spam berbasis fitur dan kompresi-model. | * Bogofilter * OSBF-Lua * DMC (Dynamic Markov Compression) * SVM (Support Vector Machine) * LR (Logistic Regression) |
| 11 | David J. Miller  2018 | A Mixture Model Based Defense for Data Poisoning Attacks Against Naive Bayes Spam Filters | Arvix | Mengusulkan pertahanan berdasarkan penggunaan campuran modul NB. | * Retrained Naïve Bayes |
| 13 | Ismaila Idris  2012 | An Improved AIS Based E-mail Classification Technique for Spam Detection | Arvix | Mengembangkan sistem berbasis kekebalan dengan menggunakan pembelajaran kekebalan, memori kekebalan dalam memecahkan masalah kompleks dalam deteksi spam. | Artificial immune system |
| 14 | Xavier Carreras  2001 | Boosting Trees for Anti-Spam Email Filtering | Arvix | Menjelaskan serangkaian percobaan perbandingan untuk masalah secara otomatis memfilter pesan surat elektronik yang tidak diinginkan. | * Naive Bayes * Decision Trees |
| 15 | Eitan Menahem  2012 | Detecting Spammers via Aggregated Historical Data Set | Arvix | Mengusulkan mekanisme reputasi pengirim baru yang didasarkan pada kumpulan data historis teragregasi yang mengkodekan perilaku agen transfer surat dari waktu ke waktu. | * Content-based filtering (CBF) * Real-time blacklists (RBL) * Sender reputation mechanisms (SRM) * Machine Learning |
| 16 | Daniel Etzold  2018 | Improving spam filtering by combining Naive Bayes with simple k-nearest neighbor searches | Arvix | Menyajikan hasil empiris klasifikasi email menggunakan kombinasi Bayes naif dan pencarian tetangga k-terdekat. | * Naïve Bayes * K Nearest Neighbour |
| 17 | Ion Androutsopoulos  2000 | Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach | Arvix | Menyelidiki kinerja dua algoritma pembelajaran mesin dalam konteks penyaringan anti-spam. | * Naive Bayesian * Memory-Based Approach |
| 18 | Alexy Bhowmick  2016 | Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends | Arvix | Kami menyajikan tinjauan komprehensif dari teknik penyaringan spam berbasis konten yang paling efektif. | * Extracting Features * Feature Selection * E-mail Header Analysis * Filters Based on Non-content Features |
| 20 | P. Oscar Boykin  2004 | Personal Email Networks: An Effective Anti-Spam Tool | Arvix | Kami menyediakan metode teoretis grafik otomatis untuk mengidentifikasi jaringan teman yang tepercaya pengguna individu di dunia maya. | * Content-based filters. |
| 22 | De Wang  2013 | A Study on Evolution of Email Spam Over Fifteen | IEEE | kami menganalisis tren spam email pada dataset yang dikumpulkan oleh Arsip Spam, yang berisi 5,1 juta email spam yang tersebar selama 15 tahun (1998-2013). | * Content Analysis * Topic Modeling * Network Analysis |
| 23 | S. Dhanaraj  2013 | A Study on E-mail Image Spam Filtering Techniques | IEEE | Mereview metode penyaringan spam pada email dengan gambar | * Local Changes in Transmissiion * Language- Based Filters * Content Based Classification * Hybrid Filters |
| 24 | Morteza Zi Hayat | Content-Based Concept Drift Detection for Email Spam Filtering | IEEE | Dalam makalah ini diusulkan sistem penyaringan spam adaptif berdasarkan model bahasa yang dapat mendeteksi penyimpangan konsep berdasarkan komputasi penyimpangan dalam distribusi konten email. | * Naïve Bayes |
| 25 | El-Sayed M. El-Alfy | Discovering Classification Rules for Email Spam Filtering with an Ant Colony Optimization Algorithm | IEEE | Pekerjaan dalam makalah ini dimotivasi oleh peningkatan dramatis dalam volume lalu lintas spam dalam beberapa tahun terakhir dan kemampuan yang menjanjikan dari optimasi koloni semut dalam penambangan data. | * Naïve Bayes * Ripper |
| 26 | Seongwook Youn | Efficient Spam Email Filtering using Adaptive Ontology | IEEE | Makalah ini mengusulkan untuk menemukan metode penyaringan email spam yang efisien menggunakan ontologi adaptif | * Adaptive ontology |
| 28 | Ali Ahmed A.Abdelrahim  2013 | Feature Selection and Similarity Coefficient Based Method for Email Spam Filtering | IEEE | Hasil penelitian berdasarkan set data spam email menunjukkan bahwa pendekatan kami yang diusulkan meningkatkan tingkat deteksi, tingkat alarm palsu dan akurasi. | * Feature Selection * Similarity Coefficient Based Method |
| 29 | Dongwook Shin  2006 | Progressive Multi Gray-Leveling: A Voice Spam Protection Algorithm | IEEE | Artikel ini mengusulkan algoritma kontrol spam suara yang disebut Progressive Multi Grey-Leveling (PMG) yang sesuai dengan pengaturan VoIP. PMG secara progresif menghitung “tingkat abu-abu” dari penelepon dan menentukan apakah panggilan akan terhubung atau diblokir berdasarkan pola panggilan sebelumnya. | Progressive Multi Gray-Leveling (PMG) |
| 30 | Ching-Tung Wu  2005 | Using visual features for anti-spam filtering | IEEE | Dalam makalah ini, kami mengusulkan sistem anti-spam baru yang menggunakan petunjuk visual, selain informasi teks dalam badan email, untuk menentukan apakah suatu pesan adalah spam. | Support Vector Machines (SVM) |
| 31 | Tu Ouyang  2013 | A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise | Science Direct | Kami menyusun sistem penentuan kebenaran tanah yang baru untuk memungkinkan kami memberi label pada dataset besar ini secara akurat. | * DNS blacklists * Filters based on SYN packet features * Filters based on traffic characteristics * Filters based on message content. |
| 32 | Ismaila Idris  2014 | A combined negative selection algorithm–particle swarm optimization for an email spam detection system | Science Direct | Membuktikan Akurasi dari model NSA (negative selection algorithm) dengan PSO (particle swarm optimization) lebih baik dari akurasi model standar NSA. | * NSA (negative selection algorithm) * NSA-PSO (negative selection algorithm -particle swarm optimization) |
| 33 | Amany A. Naem  2018 | Antlion optimization and boosting classifier for spam email detection | Science Direct | Dalam karya ini, metode prediksi baru diajukan berdasarkan antlion optimization (ALO) dan meningkatkan disebut sebagai ALO-Boosting untuk menyelesaikan masalah email spam. | Antlion optimization (ALO)  Boosting classifier  Feature extraction  Feature selection |
| 34 | ENRIQUE PUERTAS SANZ  2008 | Email Spam Filtering | Science Direct | Kami menjelaskan struktur dan proses berbagai metode Pembelajaran Mesin yang digunakan untuk tugas ini, dan bagaimana kami dapat membuatnya menjadi sensitif biaya melalui beberapa metode seperti pengoptimalan ambang batas, pembobotan contoh, atau MetaCost. | Primitive Language Analysis or Heuristic Content Filtering  White and Black Listings  Graylisting  Digital Signatures and Reputation Control  Postage  Disposable Addresses  Collaborative Filtering  Honeypotting and Email Traps  Content-Based Filters |
| 35 | Ismaila Idris  2013 | Hybrid email spam detection model with negative selection algorithm and differential evolution | Science Direct | Makalah ini memperkenalkan teknik pembelajaran mesin yang dimodifikasi dari sistem kekebalan tubuh manusia yang disebut algoritma seleksi negatif (NSA) | Algoritme seleksi negatif asli (NSA)  Usulan peningkatan model algoritma seleksi negatif  penghitungan fungsi kebugaran dalam ruang spam dan non-spam  Tumpang tindih dalam ruang spam (detektor yang dihasilkan)  Representasi grafis dari model |
| 36 | Ismaila Idris  2014 | Improved email spam detection model with negative selection algorithm and particle swarm optimization | Science Direct | Dalam penelitian ini, model baru yang ditingkatkan yang menggabungkan algoritma seleksi negatif (NSA) dengan optimasi partikel swarm (PSO) telah diusulkan dan diimplementasikan. | NSA-PSO  PSO  NSA |
| 37 | Noemí Pérez-Díaz | Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification | Science Direct | Makalah ini mengulas dan menyatukan pendekatan sebelumnya dan alternatif baru untuk menerapkan teori set kasar (RS) ke domain penyaringan spam dengan mendefinisikan tiga skema pelaksanaan aturan yang berbeda: MFD (keputusan paling sering), LNO (jumlah objek terbesar) dan LTS ( kekuatan total terbesar). | Adaboost  Flexible Bayes  Naïve Bayes  SVM  MFD  LNO  LTS  MFD  LNO  LTS |
| 38 | Clotilde Lopes | Symbiotic filtering for spam email detection | Science Direct | Makalah ini menyajikan teknik penyaringan spam baru yang disebut Symbiotic Filtering (SF) yang mengagregasi filter lokal yang berbeda dari beberapa pengguna untuk meningkatkan kinerja keseluruhan deteksi spam. SF adalah pendekatan hybrid yang menggabungkan beberapa fitur dari Collaborative (CF) dan Content-Based Filtering (CBF). | Naive Bayes filtering  Symbiotic filtering |
| 39 | El-Sayed M. El-Alfy | Using GMDH-based networks for improved spam detection and email feature analysis | Science Direct | kami mengeksplorasi penerapan pendekatan pembelajaran induktif berbasis GMDH (Metode Grup Penanganan Data) dalam mendeteksi pesan spam dengan secara otomatis mengidentifikasi fitur konten yang secara efektif membedakan spam dari email yang sah. Kami mempelajari kinerja untuk berbagai kompleksitas model jaringan menggunakan basis spam, dataset benchmark yang tersedia untuk umum. Hasil mengungkapkan bahwa akurasi klasifikasi 91,7% dapat dicapai hanya dengan menggunakan 10 dari 57 atribut yang tersedia, dipilih melalui pembelajaran yang bersifat abduktif sebagai subset fitur yang paling efektif (mis. Pengurangan data 82,5%). | * AIMabductivenetworks * Networkensembles(committees) * GMDH-based feature ranking and selection |
| 41 | Enrico Blanzieri | A survey of learning-based techniques of email spam filtering | Springer | Dalam makalah ini kami memberikan ikhtisar tentang keadaan seni aplikasi pembelajaran mesin untuk penyaringan spam, dan cara evaluasi dan perbandingan berbagai metode penyaringan. Kami juga memberikan uraian singkat tentang cabang-cabang lain dari perlindungan anti-spam dan membahas penggunaan berbagai pendekatan dalam solusi perangkat lunak anti-spam komersial dan non-komersial. | Methods based on Bag-of-Words model  Language-based filters  Filters based on non-content features  Collaborative spam filtering  Hybrid approaches  Overview of the methods |
| 42 | Xin Zhang | Adaptive Email Spam Filtering Based on Information Theory | Springer | Kami mengusulkan algoritma penyaringan spam email adaptif berdasarkan teori informasi yang melonggarkan asumsi distribusi identik dan menyesuaikan pengetahuan yang dipelajari dari satu distribusi ke yang lain. Pekerjaan kami berfokus pada analisis konten yang meminimalkan hilangnya informasi timbal balik antara instance email dan fitur kata, sebelum dan sesudah klasifikasi. | AdaFilter  TSVM  SVM  NBC |
| 43 | Bing Zhou | A Three-Way Decision Approach to Email Spam Filtering | Springer | kami memperkenalkan pendekatan keputusan tiga arah untuk pemfilteran spam berdasarkan teori keputusan Bayesian, yang memberikan umpan balik yang lebih masuk akal bagi pengguna untuk tindakan pencegahan menangani email masuk mereka, sehingga mengurangi kemungkinan kesalahan klasifikasi. Keuntungan utama dari pendekatan kami adalah bahwa hal itu memungkinkan kemungkinan penolakan, yaitu, menolak untuk membuat keputusan. | The Naive Bayesian Spam Filtering |
| 44 | Xin Zhang | Adaptive Email Spam Filtering Based on Information Theory | Springer | Kami mengusulkan algoritma penyaringan spam email adaptif berdasarkan teori informasi yang melonggarkan asumsi distribusi identik dan menyesuaikan pengetahuan yang dipelajari dari satu distribusi ke yang lain. Pekerjaan kami berfokus pada analisis konten yang meminimalkan hilangnya informasi timbal balik antara instance email dan fitur kata, sebelum dan sesudah klasifikasi. | AdaFilter  TSVM  SVM  NBC |
| 45 | Kamini (Simi) Bajaj | A Multi-layer Model to Detect Spam Email at Client Side | Springer | Dalam makalah ini kami mengusulkan model multi-lapisan yang memberlakukan, di atas SpamBayes, lapisan kedua penyaringan non-tekstual yang mengeksploitasi teknik pembelajaran mesin alternatif. Model multi-layer ini meningkatkan akurasi klasifikasi dan menghilangkan email abu-abu menjadi spam dan email ham. Hasil eksperimen model ini cukup menggembirakan. | CART, SVM, k nearest neighbor and Logistic regression machine learning technique |
| 46 | Bing Zhou | Cost-sensitive three-way email spam filtering | Springer | Kami berpendapat bahwa pendekatan pengambilan keputusan tiga arah menyediakan cara yang lebih bermakna bagi pengguna untuk tindakan pencegahan menangani email masuk mereka. Tiga folder email bukannya dua diproduksi dalam sistem penyaringan spam tiga arah, folder yang dicurigai ditambahkan untuk memungkinkan pengguna melakukan pemeriksaan lebih lanjut terhadap email yang mencurigakan, sehingga mengurangi kemungkinan kesalahan klasifikasi. | Binary naive Bayesian spam filtering  Related work on ternary email spam filtering |
| 47 | Dimitris Gavrilis | Neural Recognition and Genetic Features Selection for Robust Detection of E-Mail Spam | Springer | Dalam makalah ini disajikan metode untuk pemilihan fitur dan klasifikasi pesan spam email. Pemilihan fitur dilakukan dalam dua langkah: Seleksi dilakukan dengan mengukur entropi dan seleksi fine-tuning diimplementasikan menggunakan algoritma genetika. | STOP, LEMM and LEMM-STOP corpuses. |
| 48 | Wuying Liu | Online active multi-field learning for efficient email spam filtering | Springer | Makalah ini membahas masalah penyaringan spam email dan mengusulkan pendekatan pembelajaran multi-bidang aktif online, yang didasarkan pada ide-ide berikut: (1) Penyaringan spam email adalah aplikasi online, yang menyarankan ide pembelajaran online; (2) Dokumen email memiliki struktur teks multi-bidang, yang menunjukkan ide pembelajaran multi-bidang; dan (3) Mahal untuk mendapatkan label untuk filter spam email dunia nyata, yang menyarankan ide belajar aktif. | String-frequency index  SFITC Algorithm  Space-time complexity |
| 49 | Aram Yegenian | Inexpensive Email Addresses An Email Spam-Combating System | Springer | Karya ini mengusulkan metode yang efektif untuk memerangi spam dengan mengembangkan Inexpensive Email Addresses (IEA), sistem kewarganegaraan Dismailable Email Addresses (DEA). IEA dapat secara kriptografi menghasilkan alamat email eksklusif untuk setiap pengirim, dengan kemampuan untuk membangun kembali alamat email baru setelah yang lama dikompromikan. | TMDA, Mailhide, REAP, Remailler, MIDEA, Hashcash PennyBlack, IEA |
| 50 | ZHANG Yuqiang | A New Anti-Spam Model Based on E-mail Address Concealment Technique | Springer | Makalah ini menyajikan model perlindungan privasi baru di mana alamat email pengguna diperlakukan sebagai bagian dari informasi privasi yang dikurung. | Membandingkan tiga skema yaitu : skema 1 menggunakan model perlindungan privasi anti-spam kami, skema 2 menggunakan teknologi deteksi terkait untuk menyaring spam, dan skema 3 tidak menggunakan metode anti-spam apa pun. |

**Penutup**

Machine Learning selain digunakan pada computer vision, mendeteksi penyusup pada jaringan ataupun membangun algoritma untuk mengerjakan sesuatu, juga sangat bermanfaat untuk melakukan filter spam pada email. Dengan adanya Machine Learning pengenalan spam menjadi lebih mudah dan cepat, sehingga dapat meminimalisir terkirimnya spam ke pengguna. Hal ini juga akan membantu email marketer dalam melewati spam filter.

**Daftar Pustaka**

1. Wikipedia. [Online]. Available : https://id.wikipedia.org/wiki/Surat\_ elektronik. [Accessed 20 Mei 2018].
2. Allen, David (2004). [Windows to Linux](https://books.google.com/books?id=UD0h_GqgbHgC&printsec=frontcover&dq=network%2B+guide+to+networks&hl=en&src=bmrr&ei=hMnATfmmA8j00gGMsOC2Cg&sa=X&oi=book_result&ct=result&resnum=1&ved=0CE8Q6AEwAA#v=onepage&q&f=false). Prentice Hall. p. 192. [Archived](https://web.archive.org/web/20161226051835/https://books.google.com/books?id=UD0h_GqgbHgC&printsec=frontcover&dq=network%2B+guide+to+networks&hl=en&src=bmrr&ei=hMnATfmmA8j00gGMsOC2Cg&sa=X&oi=book_result&ct=result&resnum=1&ved=0CE8Q6AEwAA#v=onepage&q&f=false) from the original on 2016-12-26.
3. Wikipedia. [Online]. https://en.wikipedia.org/wiki/Email\_marketing. [Accessed 30 Mei 2018].
4. https://en.wikipedia.org/wiki/Machine\_learning
5. Pew Internet & American Life Project, ["Tracking surveys"](http://www.pewinternet.org/trends.asp), March 2000 – March 2009
6. [How Scheduling Affects Rates](http://www.mailermailer.com/resources/metrics/2012/how-scheduling-affects-rates.rwp). Mailermailer.com (July 2012). Retrieved on July 28, 2013.
7. BtoB Magazine, ["Early Email Blasts Results in Higher Click & Open Rates"](http://www.btobonline.com/article/20110901/EMAIL13/309019997/early-morning-email-blasts-pay-off-with-strong-opens-clicks) [Archived](https://web.archive.org/web/20111122001746/http://www.btobonline.com/article/20110901/EMAIL13/309019997/early-morning-email-blasts-pay-off-with-strong-opens-clicks) 2011-11-22 at the [Wayback Machine](https://en.wikipedia.org/wiki/Wayback_Machine)., September 2011
8. Wikipedia. [Online]. https://en.wikipedia.org/wiki/Bounce\_message. [Accessed 3 Juni 2018].
9. Wikipedia. [Online]. https://id.wikipedia.org/wiki/Sistem\_Penamaan\_Domain. [Accessed 11 Juli 2018].
10. Wikipedia. [Online]. https://en.wikipedia.org/wiki/DNSBL. [Accessed 21 Juli 2018].
11. Jaeyeon Jung. 2004. An Empirical Study of Spam Traffic and the Use of DNS Black Lists.
12. Wikipedia. [Online]. https://en.wikipedia.org/wiki/Machine\_learning. [Accessed 11 Januari 2019].