Penyaringan Spam email menggunakan K-Means

Eko Puji Laksono[1], Ardy Wicaksono[2]
[1,2]Program Studi Teknologi Informasi, Fakultas Teknik, Univ Proklamasi 45
Jl. Proklamasi No. 1 Babarsari Yogyakarta
email : [1]ekopujilaksono@up45.ac.id

Abstrak - Di Indonesia banyak kasus penyalahgunaan email yang merugikan orang lain. Email yang dikenal sebagai email sampah yang berisikan phising, scam, malware bahkan iklan. Penelitian ini bertujuan untuk memilah email spam dan ham menggunakan K-Means Clustering sebagai upaya mengurangi jumlah spam. K-means dapat membagi berdasarkan cluster yang dibuat. Dari hasil penelitian optimasi menggunakan K-Means Clustering menghasilkan akurasi 100%. Jadi berdasarkan nilai akurasi yang diperoleh, distribusi frekuensi clustering dan K-Means dapat digunakan untuk mengoptimasi pemilahan spam pada email..
Kata Kunci : *Spam email, K-Means, Preprocessing, machine learning*.

Abstract - In Indonesia, there are many cases of misuse of email that harm others. Emails that are known as junk emails contain phishing, scams, malware and even advertisements. This study aims to sort out spam and ham emails using K-Means Clustering as an effort to reduce the amount of spam. K-means can divide based on the cluster created. From the results of optimization research using K-Means Clustering produces 100% accuracy. So based on the accuracy value obtained, the clustering frequency distribution and K-Means can be used to optimize spam sorting in emails..
Keywords: *Spam email, K-Means, Preprocessing, machine learning*.

1. PENDAHULAN

Teknologi sekarang semakin canggih dan rentan terhadap spam. Salah satunya teknologi paling sering digunakan adalah email.Email atau surat elektronik adalah contoh penerapan teknologi dari segi komunikasi yang diminati karena cepat, mudah dan biaya relatif murah. Pengguna email di seluruh dunia semakin berkembang dan banyak dengan pesat. Radicati Group menemukan bahwa akun email yang terdaftar pada tahun 2015 mencapai 3,5 miliar. Beberapa orang tidak dapat memakai email dengan baik, hal ini berakibat penyalahgunaan email untuk merugikan perusahaan maupun individu[1].

Spam atau penyalahgunaan email dikenal dengan email sampah, yang berisikan penipuan dengan kedok pemenang undian, iklan penjualan produk, virus serta malware [2]. Beberapa jenis spam diantara spam mesin pencari, informasi web, spam blog dan spam jejaring sosial [3].

Banyaknya spam email dapat menimbulkan kerugian besar antara lain kerugian ekonomis dan meningkatkan data traffic terutama bagi perusahaan. kondisi tersebut mendasaru dilakukannya penelitian clustering spam dan ham. Pentingnya penelitian ini dikarenakan penanganan email spam yang efektif tidak hanya mengurangi kerugian perusahaan tetapi meningkatkan kepuasan dari pengguna email sendiri.

Penelitian ini pendeteksian email spam berupa isi email, dengan cara menganalisis frekuensi penggunaan kata dalam email ham dan spam. Metoda K-Means berusaha mengelompokkan data ke dalam beberapa kelompok berdasarkan ciri-ciri yang sama dengan yang lainnya, dan ciri-ciri yang berbeda dengan data yang berada pada kelompok lainnya, sehingga metoda ini dapat digunakanuntuk meminimalisir variasi data yang terdapat satu cluster serta maksimalkan variasi antar data-data yang terdapat dalam cluster lainnya. Berdasarkan latar belakang tersebut jurnal ini mengangkat judul "penyaringan email spam menggunakan K-Means"

II. LANDASAN TEORI DAN METODE

A. Landasan teori

1. Spam

Spam Adalah email yang tidak diinginkan oleh penerimanya[4]. SPAM adalah kepanjangan dari Stupid Pointless Annoying Message [5]. Spam pertama kali bulan mei tahun 1978. Spam dapat berbentuk pengiriman pesan secara berulang-ulang ke server newsgroup atau milist dengan topik yang tidak sesuai. [6].

2. Email

Email (*Electronic Mail*) merupakan sebuah fasilitas komunikasi dalam internet yang berfungsi mengirimkan surat secara elektronik yang dapat dijangkau diseluruh dunia [2]. Jika dibandingkan dengan surat biasa email mempunyai keunggulan lebih aman serta tidak membedakan jarak dan waktu. Email selain bisa mengirim surat dalam bentuk teks biasa, juga dapat mengirim *file* dalam bentuk foto, video dan lain-lain. Pada beberapa aplikasi seperti facebook, twitter, Instagram, traveloka, tiket.com dan sebagainya untuk bisa login ke dalam aplikasi diperlukan satu akun email yang digunakan untuk mendaftar sebelum menggunakan aplikasi. Secara umum, ada tiga jenis email yang dikenal saat ini yaitu *Post Office Procokol* (Email Berbasis POP), *Web Based mail* dan *Email Forwading*.

3. Preprocessing

Preprocessing merupakan tahap awal dari text mining yang bertujuan untuk mengubah data sesuai proses yang dibutuhkan. Proses ini dilakukan untuk mengolah dan mengatur kalimat sebelum dilakukan ekstraksi kata kunci dan penentuan label dari data terstruktur dan data tidak terstruktur[3]

4. Pembobotan TF IDF

Metode TF-IDF merupakan cara untuk menghitung bobot kata berdasarkan frekuensi munculnya kata.

Metode ini menghitung nilai (TF) *Term Frequency* serta (IDF) *Inverse Document Frequency* pada setiap token (kata) yang terdapat pada dokumen, dimana TF adalah proses untuk menghitung jumlah kemunculan term dalam satu dokumen, proses ini akan menujukkan seberapa penting kata itu di dalam satu dokumen dan IDF digunakan untuk menghitung term yang muncul di berbagai dokumen (komentar) yang dianggap sebagai term umum[7]

5. K-means

K-means clustering adalah suatu metode data mining yang melakukan proses pemodelan *unsupervised learning* dan juga salah satu metode yang menggunakan metode pengelompokan data secara partisi. K-means berupaya mengelompokkan data kedalam beberapa kelompokberdasarkan karakteristik data yang sama antar satu dengan yang lainnya tetapi memiliki karakteristik berbeda dengan data yang ada pada kelompok lain, sehingga metode ini dapat digunakan untuk meminimalkan variasi antar data yang terdapat dalam suatu cluster serta memaksimalkan variasi dengan data-data yang terdapat dalam cluster yang lainnya

4. Distribusi Clustering

Distribusi Clustering adalah metode pengelompokkan data ke beberapa kategori dengan melihat banyaknya data pada setiap kategori. Data yang didapatkan dari suatu penelitian yang berupa data acak dapat dibagi menjadi data yang berkelompok (datayang telah disusun ke dalam kelas – kelas tertentu).

III. HASIL DAN PEMBAHASAN

A. Hasil Penelitian

1. Text Processing

Preprocessing perlu dilakukan untuk membersihkan dataset dari karakter-karakter yang tidak digunakan dalam proses clustering nantinya. Sebelum melalui proses klasifikasi, perlu dilakukan *preprocessing* pada dataset yang akan digunakan untuk klasifikasi. *Preprocessing* perlu dilakukan untuk membersihkan dataset dari karakter-karakter yang tidak digunakan dalam proses klasifikasi nantinya. Dalam penelitian ini terdapat proses perubahan dataset yang dilakukan secara manual dari default data yang didapat .txt ke dalam format .csv yang bertujuan agar data yang ada dapat diolah untuk dilanjutkan ke dalam proses klasifikasi.

Dalam weka proses *preprocessing* dalam penelitian ini diawali dengan memilih dataset .csv yang telah didapat dalam proses sebelumnya yang kemudian diberi filter pada *unsupervised* bagian *nominal to string* yang kemudian dilanjutkan dengan filter *string to word factor* untuk *preprocessing* datanya. Tahapan *preprocessing* pada penelitian ini terdiri dari *case folding*, *data cleaning*, *tokenisasi*, *stopword* dan *stemming*

2. Case Folding

Tahap awal proses preprocessing teks merupakan masukkan data yang akan diproses kemudian dilakukan proses case folding. Case folding untuk merubah semua huruf kapital yang ada dala dokumen menjadi huruf kecil atau lower case. Pada tahap ini hanya karakter a-z yang akan diterima, dan karakter lain dianggap sebagai delimiter. Gambar 5 menunjukkan contoh proses case folding

Dalam weka proses ini berada pada filter *string to word factor* bagian *lowercase* yang diberi perintah (*true*). Gambar 5 adalah potongan kode program dari proses *case folding*.

```
public void setLowerCaseTokens(boolean downCaseTokens) {
   m_dictionaryBuilder.setLowerCaseTokens(downCaseTokens);
}
```

Gambar 1 Potongan Kode Program *Case Folding*

Fungsi set Lower Case Tokens digunakan untuk mengubah karakter a-z yang kapital menjadi lower case atau huruf kecil.

3. Data Cleaning

Data cleaning berfungsi untuk menghilangkan semua simbol dan tanda baca atau delimiter pada dokumen berita. Pada tahap ini semua simbol akan dihilangkan sehingga dalam dokumen berita hanya terdapat huruf kecil dan tidak ada simbol sama sekali. Dalam weka daftar delimiter yang dimasukkan dalam *field* yang akan dihilangkan antara lainnya

`~!@#$%^&*()_+1234567890||<>?.,;: '"-/[]=|.

Gambar 2 adalah potongan kode program dari proses *data cleaning*.

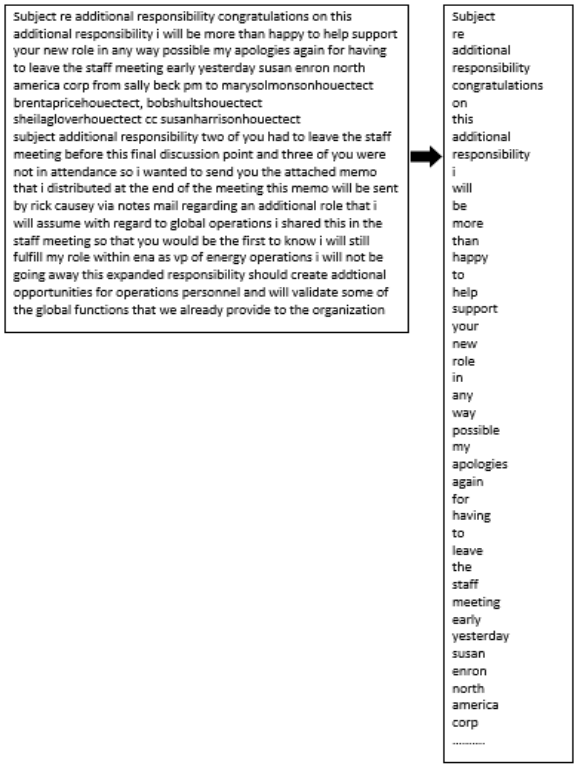
```
String tmpStr = Utils.getOption("delimiters", options);
    if (tmpStr.length() != 0) {
      setDelimiters(tmpStr);
    } else {
      setDelimiters(" \r\n\t.,;:'\"()?!");
    }
  }
```

Gambar 3 Kode Program Data Cleaning

Fungsi setDelimiters digunakan untuk menentukkan delimiter yang akan dihapus.

4. Tokenisasi

Langkah selanjutnya setelah data cleaning adalah tokenisasi. Tokenisasi merupakan tahap yang berfungsi untuk memotong setiap dokumen sehingga berubah kata berdiri sendiri. Pada tahap tokenisasi akan menghilangkan spasi (*whitespace*)



Gambar 4 Tokenisasi

Gambar 4 menggambarkan dokumen email yang di tokenisasi. Dalam tahap ini teks email yang sudah melalui tahap case folding dan data cleaning.

5. Stopword

Stopword merupakan tahap *preprocessing* untuk mengambil kata-kata yang penting dari setiap dokumen berita hasil dari tokenisasi. Pada tahap ini kata yang tidak memiliki arti penting dan kata yang tidak digunakan dalam proses klasifikasi akan dihapus untuk mengurangi jumlah kata yang disimpan oleh sistem (Manning, Raghavan, & Schutze, 2008). Potongan kode program stopword ditunjukkan pada Gambar

6. Stemming

Stemming adalah proses mengubah kata imbuhan menjadi kata dasar. Proses ini dilakukan dengan menghapus semua kata imbuhan prefixes, infixes, suffixes, confixes.

```
class PorterStemmer
{
  private char[] b;
  private int i,
    j, k, k0;
  private boolean dirty = false;
  private static final int INC = 50;
  private static final int EXTRA = 1;

  public PorterStemmer() {
    b = new char[INC];
    i = 0;
  }
```

Gambar 5 Kode Program Stemming

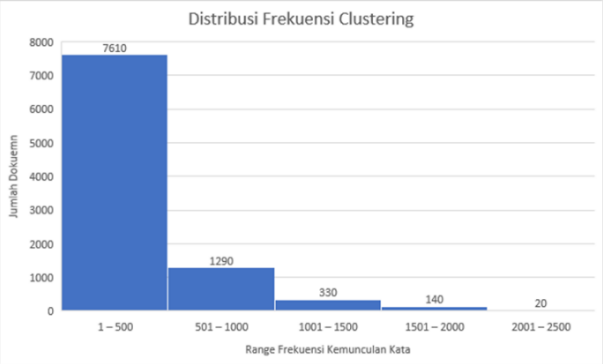
7. Distribusi Frekuensi Clustering

Optimasi parameter nilai K yang kecil pada KNN agar dapat menghasilkan nilai yang lebih besar dapat menggunakan metode distribusi frekuensi *clustering* untuk pembagian datanya. Dalam hal ini, dari proses pembobotan TF-IDF 9390 dokumen data yang ada didapatkan 1004 kata yang memiliki jumlah frekuensi kemunculan katanya sebanyak 17 sampai 2186 kali.

Tabel 1. Range Distribusi Frekuensi

Distribusi Frekuensi	Range Frekuensi Kemunculan Kata	Jumlah Dokumen
D1	1 – 500	7610
D2	501 – 1000	1290
D3	1001 – 1500	330
D4	1501 – 2000	140
D5	2001 – 2500	20
	Total	9390

Pada Tabel 1 dari hasil pembagian data menggunakan distribusi frekuensi yang ada maka terdapat 5 range frekuensi yang setiap intervalnya memiliki panjang kelas sebanyak 500. Berikut merupakan histogram hasil dari pembagian data berdasarkan frekuensi kemunculan kata yang ada dengan bantuan distribusi frekuensi *clustering* yang ada



Gambar 6. Distribusi Frekuensi Clustering

persebaran distribusi frekuensi yang ada, maka didapatkan 5 model pembagian data sebagaimana yang ada pada range frekuensi kemunculan data pada Tabel 1

3. Dari 5 model ini maka didapatkan 9 skenario penelitian yang akan digunakan untuk mengoptimalkannilai K yang kecil (K=1) pada KNN agar nilai akurasi yang dihasilkan lebih baik yaitu :

1. Skenario 1 = D 1
2. Skenario 2 = D 2
3. Skenario 3 = D 3
4. Skenario 4 = D 4
5. Skenario 5 = D 5
6. Skenario 6 = D 1 dan D 2
7. Skenario 7 = D 1, D 2 dan D 3
8. Skenario 8 = D 1, D 2, D 3 dan D 4
9. Skenario 9 = D 1, D 2, D 3, D 4 dan D 5

Berikut merupakan hasil klasifikasi menggunakan KNN dengan nilai K paling kecil (K=1) yang ada dari skenario yang dihasilkan dari distribusi frekuensi *clustering.*

Tabel 2. Hasil Evaluasi skenario 1-9

Skenario	Evaluasi			
	Akurasi (%)	*Presisi (%)*	*Recall (%)*	*Waktu (s)*
1	89.4	90.1	89.4	71.27

2	100	100	100	2.29
3	95.8	96	95.8	0.75
4	95	95.3	95	0.50
5	90	90	90	0.49
6	90.6	91	90.6	63.17
7	91.1	91.7	91.1	61.54
8	91.4	91.8	91.2	64.03
9	91.4	91.9	91.4	104.99

dapat diketahui jika nilai KNN dengan skenario 2 pada distribusi frekuensi clustering ini nilai akurasi paling tinggi sebesar 100% dan diikuti skenario 3 yang mempunyai nilai akurasi 95,8%. Dalam hal ini dibuktikan dengan nilai awal akurasi tertinggi oleh KNN saat K=1 sebesar 91,4% dapat di optimalkan dengan metode K-Means menggunakan distribusi frekuensi.

8. Klasifikasi K-Means Clustering

K-Means adalah suatu metode data mining yang melakukan proses pemodelan tanpa *supervised dan merupakan metode pengelompokkan data yang ada ke dalam beberapa kelompk*

Dalam penelitian ini data yang digunakan 9390 data dokumen dilakukan proses cluster menggunakan K-Means dengan melakukan perhitungan jarak tiap data ke centroid menggunakan Euclidean Distance.

Tabel 3. Hasil Klasifikasi K-Means Clustering

Model K-means	Jumlah Data C_0	Jumlah Data C_1	Jumlah Data C_2	Jumlah Data C_3
1	2487	6903	-	-
2	1743	4186	3461	-
3	2175	6317	691	207

Dari hasil pembagian data menggunakan K-Means yang ada pada tabel 3. Setiap model memiliki nilai k yang berbeda bedadan perbedaan tersebut didasarkan pada perubahan nilai terurut, sehingga nilai dari k yang digunakan dalam penelitian ini adalah 2, 3, dan 4 untuk mengelompokkan data awal dalam *preprocessing* sebelum masuk ke dalam proses klasifikasi. Setiap model memiliki lebih dari 1 macam skenario yang dilakukan agar mendapatkan hasil klaster yang maksimal sehingga dapat mengoptimasi nilai K optimal algoritma KNN agar akurasi yang dimiliki dapat sebanding dengan akurasi yang dimiliki algoritma Naïve Bayes

IV. KESIMPULAN

Pada penelitian ini memiliki kesimpulan adalah :

1. untuk meningkatkan nilai K pada KNN agar dapat setara dengan nilai yang dimiliki oleh KNN asli dan Naïve Bayes maka dilakukan proses pembagian data dengan metode K-means *clustering* menggunakan *euclidean distance* untuk menghitung jarak tiap data ke centroid. Dari hasil yang ada, pada K-means *clustering* skenario ke-2 dengan nilai k sebanyak 4 cluster pada model 3, memiliki nilai akurasi yang tinggi yaitu 100% dengan nilai presisi dan *recall*nya juga 100% yang hanya membutuhkan waktu 1.08 detik untuk pemrosesannya sehingga dapat dibandingkan dengan nilai akurasi yang dimiliki KNN asli dan Naïve Bayes

2. Perbandingan kinerja akurasi metode klasifikasi KNN yang dioptimasi lebih baik daripada KNN yang tidak dioptimasi untuk deteksi spam serta ham. Dalam proses dan hasil yang telah diketahui, nilai K=1 memiliki akurasi tertinggi jika dibandingkan dengan nilai K lain yang lebih besar. KNN tanpa menggunakan optimasi menghasilkan keakuratan sebesar 91.4% dengan nilai presisi dan nilai *recall* sebesar 91,9% dan 91.4%, membutuhkan durasi selama 93 detik untuk

pemrosesan. Sedangkan KNN yang dioptimasi menggunakan distribusi frekuensi dalam pembagian datanya menghasilkan akurasi sebesar 100% serta nilai presisi dan *recall*nya juga 100% dengan waktu 2 detik dan optimasi menggunakan K-means *clustering* memiliki nilai akurasi sebesar 99% dengan nilai presisi dan *recall*nya juga 99% yang hanya membutuhkan waktu 1.08 detik

REFERENSI

[1] Cranor, L.F. and LaMacchia, B.A. Spam! *Commun. ACM 41*, 8 (Aug. 1998), 74--83.

[2] Burhanudin, Y. Musa'adah, and Y. Wihardi, "Klasifikasi Komentar Spam Pada Youtube Menggunakan Metode Naïve Bayes, Support Vector Machine, dan K-Nearest Neighbors," *J. Inform. dan Komput.*, vol. 3, no. 2, pp. 54–59, 2018.

[3] Irfa, Adiwijaya, and M. S. Mubarok, "Klasifikasi Topik Berita Berbahasa Indonesia Menggunakan k-Nearest Neighbor," *e-Proceeding Eng.*, vol. 5, no. 2, p. 3631, 2018.

[4] Implementasi Spam Filter Untuk Mail Server Menggunakan Tools Spamassassin,e-Proceeding of Applied Science : Vol.3, No.3 Desember 2017 ,ISSN : 2442-5826.p.1925-1933

[5] Darma Juang,Analisis Spam Dengan Menggunakan Naïve Bayes,Jurnal Teknovasi, Volume 03, Nomor 2, 2016,ISSN : 2355-701X,pp. 51 – 57

[6] Ratih Yulia Hayuningtyas,Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes,IJCIT (Indonesian Journal on Computer and Information Technology) Vol.2 No.1, Mei 2017,E-ISSN: 2549-7421, pp. 53-60

[7] Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval Introduction. In *Computational Linguistics* (Vol. 35). https://doi.org/10.1162/coli.2009.35.2.307

.