KLASIFIKASI E-MAIL SPAM MENGGUNAKAN METODE K-NEAREST NEIGHBOR

Alan Fitriyanto, Ilham Saifudin
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
Alanfitriyanto2@gmail.com, ilhamsaifudin@unmuhjember.ac.id

ABSTRAK

Email merupakan alat komunikasi penting yang digunakan oleh individu atau perusahaan saat ini. Adanya email yang mengandung spam sangat mengganggu bagi para pengguna email. Dalam penelitian ini peneliti melakukan klasifikasi email spam menggunakan algoritma K-NN guna mengetahui tingkat akurasi dan presisi dari algoritma ini dalam melakukan klasifikasi. Data dalam penelitian ini menggunakan dataset yang berasal dari UCI Machine Learning dengan total data 4601, terbagi atas 921 data validasi dan 3680 data uji serta latih. Dalam skenario uji penelitian ini menggunakan K-Fold Cross Validation dengan nilai k adalah 2, 4, 5, 8, 10. Hasil yang diperoleh dari penelitian ini adalah model alternatif yang didapat dari skenario uji dengan nilai akurasi tertinggi adalah pada uji dengan k=8 percobaan ke 8 yaitu 85.2% dan presisi 86.6%. sedangkan setelah diuji pada data validasi model ini menghasilkan tingkat akurasi 82.7% dan presisi 86.1%. hal ini menunjukkan bahwa terjadi penurunan tingkat akurasi sebesar 2.4% dan penurunan tingkat presisi sebesar 0.4%.

Kata kunci : Klasifikasi, Email, Spam, K-NN

BAB I LATAR BELAKANG

1.1. Latar belakang

Spam adalah praktek pengiriman pesan komersial atau iklan kepada sejumlah besar newsgroup atau email yang sebetulnya tidak berkeinginan atau tidak tertarik menerima pesan tersebut (Wahid, Fatul. 2005). spam adalah pengiriman email yang disalahgunakan karena biaya yang murah dan ini dimanfaatkan oleh perusahaan-perusahaan besar untuk mempromosikan suatu produk walaupun si penerima tidak menyetujuinya. KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar. Peneliti menggunakan algoritma KNN karena memiliki konsistensi yang kuat. Ketika jumlah data mendekati tak hingga, algoritma KNN menjamin error rate yang minim. Peneliti memiliki dataset yang bersumber dari UCI Machine Learning terhadap klasifikasi email spam. Dataset tersebut memiliki perbedaan atribut dari penelitian diatas serta peneliti ingin mencoba metode lain dalam klasifikasi email spam. Dari latar belakang tersebut penulis disini memiliki ketertarikan untuk mengangkat sebuah judul yaitu "KLASIFIKASI E-MAIL SPAM MENGGUNAKAN METODE K-NEAREST NEIGHBOR".

1.2. Rumusan masalah

Dari latar belakang diatas penulis memilih beberapa pokok permasalahan sebagai berikut:

1. Berapa tingkat akurasi yang diperoleh pada klasifikasi email spam menggunakan metode K-nearest neighbor?
2. Berapa tingkat presisi yang diperoleh pada klasifikasi email spam menggunakan metode K-nearest neighbor?

1.3. Tujuan

Dari hasil penelitian ini penulis memiliki beberapa tujuan yaitu:

1. Mengetahui tingkat akurasi tertinggi dari metode K-Nearest Neghbor pada klasifikasi email spam.
2. Mengetahui tingkat presisi tertinggi dari metode K-nearest neighbor pada klasifikasi email spam.

1.4. Manfaat

Dalam penelitian ini penulis berharap adanya manfaat terhadap pihak lain atau pengembang nantinya, maka dari itu manfaat hasil penelitian ini yaitu:

1. Penelitian dalam menemukan model alternatif untuk mengklasifikasi email spam.
2. Pengembangan penelitian penggunaan data mining dalam klasifikasi email spam.

1.5. Batasan masalah

1. Dalam penelitian ini memiliki titik fokus terhadap pokok masalah. Pokok masalah tersebut yaitu:
2. Dataset yang digunakan berasal dari UCI Machine Learning dari penelitian Hewlett-Packard Labs pada tahun 1999.
3. Atribut yang digunakan antara lain, word_freq_make, word_freq_address, word_freq_all, word_freq_3d, word_freq_our,word_freq_over, word_freq_remove, word_freq_internet, word_freq_order, word_freq_mail, word_freq_receive, word_freq_will, word_freq_people, word_freq_report, word_freq_addresses, word_freq_free, word_freq_business word_freq_email, word_freq_you, word_freq_credit, word_freq_your, word_freq_font, word_freq_000, word_freq_money, word_freq_hp, word_freq_hpl, word_freq_george, word_freq_650,

word_freq_lab, word_freq_labs, word_freq_telnet, word_freq_857, word_freq_data, word_freq_415, word_freq_85, word_freq_technology, word_freq_1999, word_freq_parts, word_freq_pm, word_freq_direct, word_freq_cs, word_freq_meeting, word_freq_original, word_freq_project, word_freq_re, word_freq_edu, word_freq_table, word_freq_conference, char_freq_%3B, char_freq_%28, char_freq_%5B, char_freq_%21, char_freq_%24, char_freq_%23, capital_run_length_average, capital_run_length_longest dan capital_run_length_total.

4. Output yang dihasilkan yaitu spam dan bukan spam.
5. Skenario uji menggunakan K-Fold Cross Validation dengan nilai k= 2, 4, 5, 8 dan 10.
6. Penghitungan jarak terdekat menggunakan eucledience distance dengan nilai k tetangga terdekat = 3.

BAB II TINJAUAN PUSTAKA

2.1. Email

Eletronic mail atau lebih sering kita kenal dengan singkatan Email merupakan salah satu layanan internet yang paling banyak digunakan. Email adalah media komunikasi yang murah, cepat dan mudah penggunaannya. Format Email terdiri dari sebuah envelope, beberapa field header, sebuah blank line dan body. Email memiliki sifat data berupa teks yang semi terstruktur dan memiliki dimensi yang tinggi (Rachli, Muhammad. 2007).

2.2. Email spam

Spam adalah praktek pengiriman pesan komersial atau iklan kepada sejumlah besar newsgroup atau Email yang sebetulnya tidak berkeinginan atau tidak tertarik menerima pesan tersebut (Wahid, Fatul. 2005). Menurut Paul Graham mendefinisikan Spam sampah adalah Email yang tidak diinginkan yang dikirimkan secara otomatis. Atau dapat didefinisikan sebagai Email yang dikirimkan kepada ribuan penerima (Defiyanti, Sofi. 2008). Dari kedua pengertian di atas dapat disimpulkan bahwa Spam adalah pengiriman Email yang disalahgunakan karena biaya yang murah dan ini dimanfaatkan oleh perusahaan-perusahaan besar untuk mempromosikan suatu produk walaupun si penerima tidak menyetujuinya.

2.3. Data mining

Menurut Han dan Kamber (2011:6) menjelaskan bahwa "Data mining merupakan pemilihan atau "menggali" pengetahuan dari jumlah data yang banyak." Berbeda dengan Segall, Guha & Nonis (2008:127) menjelaskan "Data mining disebut penemuan pengetahuan atau menemukan pola yang tersembunyi dalam data. Data mining adalah proses menganalisis data dari perspektif yang berbeda dan meringkas menjadi informasi yang berguna". Bisa disimpulkan Data mining adalah Proses menganalisis data yang banyak dan membuat suatu pola untuk menjadi informasi yang berguna.

2.4. Klasifikasi

Klasifikasi memiliki dua proses yaitu membangun model klasifikasi dari sekumpulan kelas data yang sudah didefinisikan sebelumnya (training data) dan menggunakan model tersebut untuk klasifikasi data uji serta mengukur akurasi dari model. Model klasifikasi dapat disajikan dalam berbagai macam model klasifikasi seperti decision trees, bayesian classification, k-nearest neighbourhood classifier, neural network, classification (IF-THEN) rule, dan lain-lain. Klasifikasi dapat dimanfaatkan dalam berbagai aplikasi seperti diagnosa medis, selective marketing, pengajuan kredit perbankan dan Email.

2.5. K-NN

Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru berdasarkan atribut dan training sample. Classifier tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik query, akan ditemukan sejumlah k obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari k obyek.. algoritma K-Nearest Neighbor (K-NN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru

Persamaan untuk K-NN sebagai berikut :

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

Keterangan :

p_i=nilai parameter x1 pada data training

q_i=nilai parameter x1 pada data testing

2.6. Confusion matrix

Menurut Han dan Kamber (2011:365) Confusion matrix adalah alat yang berguna untuk menganalisis seberapa baik classifier mengenali tuple dari kelas yang berbeda. TP dan TN memberikan informasi ketika classifier benar, sedangkan FP dan FN memberitahu ketika classifier salah.

$$akurasi = \frac{TP+TN}{TP+TN+FN+FP}$$

$$presisi = \frac{TP}{TP+FP}$$

BAB III METODE PENELITIAN

3.1. Tahapan penelitian

1. Studi literatur

Sebelum melakukan penelitian, peneliti pokok permasalahan serta materi-materi pendukung. Diantaranya, dataset, penelitian terkait serta metode yang akan diimplementasikan serta metode pengukur untuk menguji metode yang

digunakan terhadap dataset. Metode yang digunakan dalam implementasi disini yaitu metode K Nearest Neighbor sedangkan untuk kinerja diukur nilai akurasi dan presisinya.

2. Pengumpulan data
   Data uji yang digunakan dalam penelitian ini bersumber pada database spam-mail yang diperoleh dari UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html. Database terdiri dari koleksi email dari bulan Juni sampai Juli 1999. Database terdiri dari total 4601 email, dimana 1813 (39.4%) adalah spam dan 2788 (60.6%) adalah spam. Koleksi spam-email berasal dari HP email dan spam-email individu. Koleksi spam email berasal dari email kantor dan email perseorangan. Setiap email telah dianalisa dan terdapat 58 atribut (57 atribut input dan 1 atribut target atau kelas) yang
3. Implementasi K-NN
   1. Dalam mengimplementasi K-NN ada beberapa langkah yang dapat dilakukan, yaitu:
   2. Tentukan parameter K = jumlah banyaknya tetangga terdekat
   3. Hitung jarak antara data baru dan semua data yang ada di data training.
   4. Urutkan jarak tersebut dan tentukan tetangga mana yang terdekat berdasarkan jarak minimum ke-K.
   5. Tentukan kategori dari tetangga terdekat.
   6. Gunakan kategori mayoritas yang sederhana dari tetangga yang terdekat tersebut sebagai nilai prediksi dari data yang baru.
4. Hasil
   Hasil yang ingin didapat yaitu, tingkat akurasi dan presisi dari metode K-NN terhadap klasifikasi email spam ini. selanjutnya akan menentukan apakah metode ini layak atau tidak jika dikembangkan serta jika ada penelitian lain yang ingin membandingkan dengan metode lain.

3.2. Skenario uji
Dari 4601 data terbagi atas 3 partisi yaitu data validasi dan data latih serta data uji. Data latih dan data uji menjadi satu kesatuan dalam skenario uji yang mengunakan K-Fold Cross Validation, pada penelitian ini nilai k adalah 2, 4, 5, 8 dan 10.

BAB IV IMPLEMENTASI DAN PENGUJIAN

4.1. Deskripsi data
Data penelitian ini diambil dari UCI Machine Learning, dimana data yang didapat hasil dari pengolahan email yang dihimpun perusahaan Hawlet-Packard (Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs) selaku pemilik utama sebelum dihibahkan ke pihak UCI Machine Learning, Diberiken oleh George Forman (gforman at nospam hpl.hp.com) dan data ini dihimpun pada tahun 1999 oleh perusahaan serupa. Data hasil olah yang terdapat pada UCI Machine Learning tersebut berupa angka yang didapat dari hasil olah plain text email dengan ketentuan seperti yang dijelaskan pada bab iii poin kedua sub pengumpulan data. Jumlah data keseluruhan yaitu 4601, dimana 2788 data bukan spam dan 1813 data adalah spam.

4.2. Preprocessing
Berikut aturan konversi yang dibuat oleh perusahaan Hawlet-Packard selaku penghimpun dataset.

1. 48 atribut bertipe *continous* [0,100] yang beranggotakan kata. Kata yang dimaksud antara lain :

Make	*address*	*all*	*3d*
our	*over*	*remove*	*Intenet*
order	*mail*	*receive*	*Will*
people	*report*	*addresses*	*Free*
business	*Emai*	*you*	*Credit*
your	*font*	*000*	*Money*
Hp	*Hpl*	*George*	*650*
lab	*labs*	*telnet*	*857*
data	*415*	*85*	*Technology*
1999	*parts*	*pm*	*Direct*
cs	*meeting*	*original*	*Project*
re	*edu*	*table*	*Conference*

dengan persentase:

$$\frac{jumlah\ kata\ yang\ muncul\ dalam\ e-mail}{total\ keseluruhan\ kata\ dalam\ e-mail}\ x100\%$$

2. 6 atribut bertipe *continous* [0,100] yang beranggotakan karakter berikut:

“;”	“(“	“[“	“!”	“$”	“#”

$$\frac{jumlah\ karakter\ yang\ muncul\ dalam\ e-mail}{total\ keseluruhan\ kata\ dalam\ e-mail}\ x100\%$$

3. 1 atribut bertipe *continous* real [1,…] yang berisi nilai rata-rata deret huruf kapital yang tidak bisa dipecahkan.
4. 1 atribut bertipe *continous* real [1,…] yang berisi nilai terpanjang deret huruf kapital yang tidak bisa dipecahkan.
5. 1 atribut bertipe *continous* real [1,…] yang berisi nilai jumlah deret huruf kapital yang tidak bisa dipecahkan

6. Output *class* 0 dan 1. 0 adalah email yang tergolong bukan *spam* sedang 1 adalah email yang tergolong *spam*.

4.3. Hasil K-NN

K fold ke	TP	TN	FP	FN	akurasi	presisi
2.1	921	560	183	176	80.49	83.42
2.2	948	509	200	183	79.18	82.58
K fold ke	TP	TN	FP	FN	akurasi	presisi
4.1	455	284	85	96	80.33	84.26
4.2	454	297	77	92	81.63	85.50
4.3	490	252	90	88	80.65	84.48
4.4	479	274	93	74	81.85	83.74
K fold ke	TP	TN	FP	FN	akurasi	presisi
5.1	359	230	67	80	80.03	84.27
5.2	384	223	65	64	82.47	85.52
5.3	369	240	61	66	82.74	85.81
5.4	393	201	75	67	80.71	83.97
5.5	388	221	69	58	82.74	84.90
K fold ke	TP	TN	FP	FN	akurasi	presisi
8.1	224	156	34	46	82.61	86.82
8.2	235	136	43	46	80.65	84.53
8.3	239	139	36	46	82.17	86.91
8.4	221	161	38	40	83.04	85.33
8.5	245	138	37	40	83.26	86.88
8.6	259	118	49	34	81.96	84.09
8.7	232	137	46	45	80.22	83.45
8.8	246	146	38	30	85.22	86.62
K fold ke	TP	TN	FP	FN	akurasi	presisi
10.1	181	120	27	40	81.79	87.02
10.2	178	115	35	40	79.62	83.57
10.3	198	106	37	27	82.61	84.26
10.4	189	118	27	34	83.42	87.50
10.5	179	128	30	31	83.42	85.65
10.6	195	116	27	30	84.51	87.84
10.7	210	91	31	36	81.79	87.14
10.8	186	111	43	28	80.71	81.22
10.9	192	110	34	32	82.07	84.96
10.1	197	114	32	25	84.51	86.03
K fold ke	TP	TN	FP	FN	akurasi	presisi
validasi	478	284	77	82	82.74	86.13

BAB V PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian serta analisis dari metode K-NN terhadap klasifikasi pada data *email* dengan total data 4601, penulis mendapatkan hasil sebagai beriku:

1. Dari pengujian terhadap data validasi dengan jumlah 921 data dan data latih berjumlah 3680 didapat bahwa tingkat akurasi dari metode K-NN dalam mengklasifikasi data *email* spam adalah 82.7%.
2. Dari pengujian terhadap data validasi dengan jumlah 921 data dan data latih berjumlah 3680 didapat bahwa tingkat presisi dari metode K-NN dalam mengklasifikasi data *email* spam adalah 86.1%.
3. Dari hasil *k-fold cross validation* didapatkan model alternatif dengan tingkat akurasi tertinggi pada k = 8 dan uji coba kedelapan yaitu 85.2% dan presisi 87.8%.
4. Terdapat penurunan akurasi yang dihasilkan dari data validasi dibandingkan akurasi yang dihasilkan dari skenario *k-fold cross validation* yaitu sebesar 2.4%.
5. Terdapat penurunan presisi yang dihasilkan dari data validasi dibandingkan akurasi yang dihasilkan dari skenario *k-fold cross validation* yaitu sebesar 0.4%.

5.2. Saran

Penulis sebagai penyusun dalam penelitian ini menyadari bahwa penelitian ini jauh dari kesempurnaan. Untuk itu penulis membuka lebar terhadap pembaca untuk memberikan saran. Berikut saran yang diberikan oleh penulis untuk pengembangan selanjutnya:

1. Pengembang dapat melakukan penelitian serupa dengan menggunakan metode klasifikasi lain untuk mendapatkan hasil dari perbandingan metodenya.
2. Pengembang dapat mengimplementasikan terhadap sebuah sistem penerimaan dan penyaringan *email* secara langsung.
3. Pengembang dapat mengimplementasikan terhadap *email* secara langsung dengan menerapkan implementasi dan metode *text mining*.

Daftar Pustaka

Kristiansen, S., Kimeme, J., Mbwambo, A., & Wahid, F. (2005). *Information flows and adaptation in Tanzanian cottage industries. Entrepreneurship & Regional Development*, 17(5), 365-388.

M. K. Chae. (2017). *Spam Filtering Email Classification (SFECM) using Gain and Graph Mining Algorithm*. Charles Sturt University Study Centre : Sydney, Australia.

Arie Wahyu Wijayanto. (2014). *Fighting Cyber Crime in Email Spamming: An Evaluation of Fuzzy Clustering Approach to Classify Spam Messages*. School of Electrical Engineering andr Informatics Insitut Teknologi Bandung: Bandung, Indonesia.

Aakash Atul Alurkar. (2017). *A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques*. Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University: Pune, India.

Sunil B. Rathod. (2015). *Content Based Spam Detection in Email using Bayesian Classifier*. IEEE ICCSP.

Widiasari, I. R., & Bayu, T. I. (2013). *Pembangunan Spam E-Mail Filtering System dengan Metode Naive Bayesian*.

Rachli. (2007). TARacli, http://www.certrr.or.id/~budi/course/security/2006-2007/Report-Muhamad-Rachli.doc, Diakses pada tanggal 5 Desember 2018.

Ariyus, Dony. (2009). *Pengantar Ilmu Kriptografi Teori, Analisis, dan Implementasi*. Penerbit Andi, Yogyakarta.

Defiyanti, S., & Pardede, D. L. (2010). *Perbandingan kinerja Algoritma ID3 dan C4. 5 dalam klasifikasi spam-mail*. Skripsi Program Studi Sistem Komputer.

Dyah Diwasasri, Ratnaningtyas. (2011). *Aplikasi Teorema Bayes dalam Penyaringan Email*. Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Segall, R. S., Guha, G. S., & Nonis, S. A. (2008). *Data mining of environmental stress tolerances on plants*. Kybernetes, 37(1), 127-148.