Deteksi Surel Spam dan Non Spam Bahasa Indonesia Menggunakan Metode Naïve Bayes

Fayruz Rahma
Program Studi Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman,
Yogyakarta, Indonesia
E-mail: fayruz.rahma@uii.ac.id

Azmiardhy Zulkifli Farmadiansyah
Program Studi Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman,
Yogyakarta, Indonesia
E-mail: 17523225@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman,
Yogyakarta, Indonesia
E-mail: fathan@uii.ac.id

Abstrak — Penggunaan surel yang mudah saat ini banyak sekali dimanfaatkan banyak orang sehingga menimbulkan dampak positif maupun negatif. Surel negatif biasa kita sebut dengan surel spam yang berisi berupa iklan, penipuan, virus dan *malware* yang berpotensi untuk merugikan orang lain. Masalah tersebut memerlukan penanganan untuk mengatasinya. Penelitian ini bertujuan untuk membuat sebuah model klasifikasi surel spam dan non spam berbahasa Indonesia menggunakan algoritma *Naïve Bayes*. Berdasarkan hasil penelitian yang telah dilakukan, ditemukan bahwa *algoritma Naïve Bayes* dengan menggunakan fitur N-gram telah berhasil melakukan klasifikasi sangat baik dengan nilai akurasi 90% hingga 94%, nilai *precision* 85% hingga 96%, *recall* 96% hingga 98, dan *F-Score* 91% hingga 97%.

Kata Kunci—Surel, Spam, Klasifikasi, Naïve Bayes

I. PENDAHULUAN

Surel merupakan sarana komunikasi dalam jaringan internal maupun internet untuk pertukaran informasi. Surel masih digunakan hingga saat ini karena kemudahan dalam hal penggunaannya. Saat ini selain digunakan untuk komunikasi surel juga digunakan untuk kebutuhan otentikasi aplikasi dan sinkronisasi media sosial seperti Instagram, Facebook dan Twitter. Penggunaan *surel* yang tinggi bisa berdampak positif dan berdampak negatif karena tidak semua orang menggunakan surel dengan baik dan bahkan ada banyak sekali penyalahgunaan surel sehingga berpotensi merugikan pengguna *surel* lainnya. *Surel* yang disalahgunakan ini disebut sebagai spam atau surel sampah yang mana memiliki konten tentang iklan, *Scam*, dan virus[1].

Surel *spam* yang beredar di kalangan pengguna sebenarnya memiliki pola tertentu hanya saja banyak sekali pengguna awam tidak banyak mengetahui. biasanya kasus yang banyak terjadi adalah surel spam berjenis iklan yang memenuhi kotak masuk surel korban padahal surel tersebut tidak diinginkan. *Spam* dapat menyebabkan ketidakefisienan *bandwidth* karena merupakan kapasitas dari sebuah jaringan agar dapat dilewati oleh paket data. Bagi banyak orang hal ini sangat mengganggu sehingga dibutuhkan penanganan mengatasi surel spam ini.

Permasalahan ini dapat diminimalisir dengan membuat sebuah model anti spam yang bertujuan untuk mengklasifikasikan *surel* dan memberikan informasi terhadap pengguna *surel* apabila terdapat pesan yang diprediksi sebagai pesan spam. Salah satu metode menciptakan anti spam adalah dengan metode Naive Bayes untuk mengklasifikasikan *surel spam* dan *non spam*. Penelitian menggunakan metode naïve bayes sebenarnya telah banyak dilakukan untuk mengklasifikasikan surel spam berbahasa Inggris namun penelitian untuk surel yang berbahasa Indonesia sangat jarang dilakukan. Oleh karena itu, penelitian ini akan mengukur metode tersebut seberapa baik performa metode Naive Bayes untuk menangani permasalahan *surel spam* yang berbahasa Indonesia.

II. LANDASAN TEORI

A. Surel Spam

Spam atau *stupid pointless annoying messages* merupakan serangan pesan yang dikirimkan ke sejumlah pengguna layanan pesan yang tidak secara khusus meminta pesan tersebut. *Spam* juga dapat didefinisikan sebagai pengiriman pesan secara berulang-ulang. Berikut merupakan tipe-tipe *surel spam* [2]:

1. Iklan: digunakan untuk mempromosikan suatu barang atau layanan yang dimiliki suatu perusahaan maupun individu perorangan.
2. Phising: menyamar sebagai perusahaan besar/lembaga terpercaya untuk memikat para korban untuk mengunjungi situs web palsu yang tertera dalam pesan dan mengambil data pribadi korban.
3. Malware: memperdaya korban dengan mengirimkan sebuah file yang berisikan sebuah virus malware.
4. Scam: upaya penyamaran yang dilakukan untuk mendapatkan simpati korban sehingga bisa mendapatkan sesuatu hal yang berharga seperti data maupun uang.

Perbedaan *Spam* dan *Non Spam* dapat dilihat dari struktur surel sebagai berikut:

1. Subject: merupakan judul topik yang mewakili isi surel biasanya dalam surel spam terdapat kata-kata “Ada Diskon” yang sering dijumpai pada korban yang terkena serangan surel spam.
2. Body: merupakan inti dari pesan surel yang diberikan dan isi surel spam sangat mudah dikenali dengan melihat kata-kata yang dikirimkan oleh pengirim

B. *Crowdsourcing*

Crowdsourcing merupakan suatu aktivitas yang dilakukan untuk mendapatkan sebuah ide, data ataupun informasi untuk menyelesaikan masalah yang kompleks dengan tidak memandang latar belakang pendidikan, kewarganegaraan, agama, amatir maupun profesional setiap individu diperbolehkan untuk ikut berpartisipasi dengan pengetahuan dan pengalaman sehingga dalam permasalahan yang ada dapat ditangani secara cepat, tepat, dan hemat biaya.

Pengguna nantinya akan mendapatkan kepuasan dari hasil yang telah didapat baik itu ekonomi, pengkauan sosial, maupun pengembangan keterampilan individu [3].

C. *Text Mining*

Text mining adalah proses yang dilakukan untuk menggali data dengan format teks. *Text Mining* mempunyai tujuan untuk mengambil kata dan memperoleh sebuah informasi sehingga dari hasil yang didapat bisa dilakukan sebuah analisis yang memiliki nilai untuk kepentingan tertentu. Terdapat beberapa tahapan proses dalam implementasi text mining yaitu *text preprocessing,* teks tranformasi, seleksi fitur, dan *pattern discovery*[4].

D. *N-Gram*

N-Gram merupakan model probabilistik yang dikembangkan untuk memprediksi urutan item selanjutnya pada item yang berurutan. Item dapat berupa karakter/huruf, kata dan lain sebagainya. Penggunaan *N-Gram* pada item kata digunakan untuk mengambil potongan kata berdasarkan nilai n yang ditentukan. Berikut merupakan contoh penggunaan *N-gram* pada kalimat "Saya sedang membaca jurnal penelitian tersebut" dapat dituliskan dalam metode N-Gram sebagai berikut:

- *Unigram*(n=1): Saya, sedang, membaca, penelitian, jurnal, tersebut
- *Bigram*(n=2): Saya sedang, membaca jurnal, penelitian tersebut
- *Trigram*(n=3): Saya sedang membaca, jurnal penelitian tersebut.

dan seterusnya.

Metode *N-Gram* juga memiliki keunggulan yaitu tidak sensitive terhadap kesalahan penulisan yang ada pada suatu data[5].

E. *Naïve Bayes*

Merupakan sebuah algoritma yang dikemukakan oleh ilmuwan inggris Thomas Bayes untuk pengklasifikasian yang dihitung dari gabungan probabilitas dengan melakukan penjumlahan antar frekuensi dan kombinasi nilai dari data yang ada. Kelebihan dalam menggunakan algoritma Naïve Bayes adalah metode ini hanya membutuhkan training yang sedikit[4].

Berikut merupakan persamaan Naïve Bayes:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

Keterangan:

X : Data dengan kelas yang belum diketahui.

H : Hipotesis data merupakan suatu kelas spesifik

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (posterior probabilitas)

P(H) : Probabilitas hipotesis H (prior probabilitas)

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

Dalam bayes memiliki aturan jika P(H1|x) < P(H2|x), maka x diklasifikasikan sebagai h2. Pernyataan P(H1|x) mengindikasikan probabilitas hipotesis h1 berdasarkan kondisi x terjadi, begitu pula dengan h2. Sebenarnya dapat klasifikasi dari x sesuai dengan probabilitas terbesar di antara probabilitas x terhadap semua kelas.

F. *Performance Evaluation Measure*

PEM memiliki tujuan untuk mengevaluasi model yang telah dibuat dan merepresentasikan prediksi kondisi sebenarnya(aktual) dari data yang dihasilkan oleh algoritma yang digunakan. banyak perhitungan untuk mendapat hasil dari nilai PEM yaitu[4]:

- *Precision*

Mengukur tingkat kepastian (*exactness*) atau jumlah data *testing* yang diklasifikasikan dengan benar oleh model klasifikasi yang dibangun[6].

Rumus *Precision(pre):*

$$pre = \frac{TP}{FP + TP} \qquad (2)$$

- *Accuration*

merupakan pembanding antara infromasi yang dijawab oleh sistem dengan benar oleh keseluruhan informasi

Rumus *Accuration(acc):*

$$acc = \frac{TN}{FN + FP + TN + TP} \qquad (3)$$

- *Recall*

Recall mengukur sensitivitas atau rasio dari data untuk setiap label yang diklasifikasikan dengan benar terhadap data yang salah diklasifikasikan ke label lainnya.

Rumus Recall(rec):

$$rec = \frac{TP}{FN + TP} \qquad (4)$$

Keterangan:

TP (*true positive*): data bernilai positif yang diprediksi benar sebagai positif

TN (*true negative*): data bernilai negatif yang diprediksi benar sebagai negatif

FP (*false positive*): data bernilai negatif yang diprediksi salah sebagai positif

FN (*false negative*): data bernilai positif yang diprediksi salah sebagai negatif

- *F-Score*

F-Score merupakan perbandingan rata-rata presisi dan *recall* yang dibobotkan.

Rumus F-Score:

$$Fscore = \frac{2xRecallxPrecison}{Recall + precision} \qquad (5)$$

G. Penelitian Terkait

Terdapat penelitian terkait klasifikasi surel spam yang telah dilakukan sebelumnya. Hengki, et al.[7] menggunakan algoritma Naïve Bayes dan SVM dengan berbasis PSO untuk mengklasifikasikan *surel* berbahasa Inggris dan keduanya memiliki akurasi yang sangat baik. Chandra et al.[8] melakukan perbandingan antara metode Pos Tagger dan Naïve Bayes dengan menggunakan dataset Bahasa Inggris. dalam mengklasifikasikan dari hasil penelitian yang dilakukan Naïve Bayes memiliki skor yang lebih tinggi dibandingkan Pos Tagger.

Menurut Juang[9] dari hasil penelitian yang sudah dilakukan bahwa algoritma naïve bayes dapat mengklasifikasikan suatu pesan ke dalam dua kelas yaitu spam dan non spam. Dari pengklasifikasian tersebut sangat dipengaruhi oleh proses *training.*

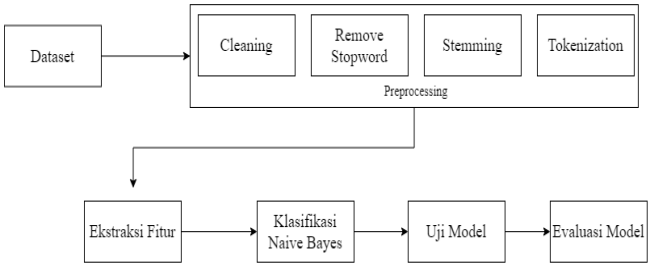
Hayuningtyas[2] melakukan penelitian menggunakan algoritma Naïve Bayes dan menggunakan *confusion matrix* sebagai evaluasi model. Hasil pengujian yang dilakukan Naïve Bayes sudah baik dalam mendeteksi *spam* karena memiliki akurasi 75,9% .

Alwani, et al[10] melakukan penelitian untuk membangun *surel spam filtering* berbahasa Arab. Dalam penelitiannya mereka menggunakan fitur N-gram dan menggunakan algoritma Bayesian. Keakuratan *spam filtering* berbahasa arab yang mereka bangun sangat baik mencapai 80%.

Abdulhamid, et al [11] melakukan perbandingan antar algoritma untuk *surel spam detection* dalam penelitiannya mereka membandingkan algoritma *Bayesian Logistic Regression, Hidden Naïve Bayes, Radial Basis Function (RBF) Network, Voted Perceptron, Lazy Bayesian Rule, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Naïve Bayes, Multilayer Perceptron, Random Tree, dan J48*. Hasil yang didapat algoritma *Rotation Forest* mendapatkan akurasi paling tinggi dan algoritma yang memiliki akurasi paling rendah untuk *spam detection* adalah algoritma *REP Tree.*

III. METODOLOGI PENELITIAN

Tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut:



Gambar 1. Alur Penelitian

A. Dataset

Data surel berbahasa Indonesia dikumpulkan secara manual dengan meminta bantuan kepada setiap individu yang terindikasi terkena serangan surel spam dan yang memiliki surel *Non Spam*. Data dikumpulkan dengan format .txt dengan jumlah data yang didapat sebanyak 617 data surel dengan rincian 317 Surel *Spam* dan 300 Surel *Non Spam*. Data yang diperoleh dibagi menjadi 2 bagian *training* 80% dan *testing* 20%.

B. *Pre-processing*

Setelah data diperoleh, dilakukan proses *preprocessing* yang bertujuan untuk menjadikan data menjadi data yang lebih terstruktur. Tahapan ini menggunakan bantuan *library* pada bahasa Python. Tahap-tahap *preprocessing* yang dilakukan sebagai berikut:

1. *Cleaning*: Membersihkan data yang didapat dari tanda baca atau *punctuation*, *hashtag* maupun *mention*, dan mengubah semua kata menjadi *lowercase.*

TABEL I. *CLEANING*

Sebelum	Sesudah
saya sudah menginstal = virus Trojan pada Sistem Operasi	saya sudah menginstal virus Trojan pada sistem operasi
Absensi bisa langsung menemui saya di ruangan,	absensi bisa langsung menemui saya di ruangan

2. *Remove Stopwords*: Penghapusan kata-kata yang kurang memiliki makna yang berarti seperti kata: dan, saya, atau. Pada proses ini menggunakan bantuan *library nltk* pada bahasa pemrograman *Python.*

TABEL II. *REMOVE STOPWORDS*

Sebelum	Sesudah
saya sudah menginstal virus Trojan pada sistem operasi	menginstal virus trojan sistem operasi
absensi bisa langsung menemui saya di ruangan	absensi langsung menemui ruangan

3. *Stemming:* Mengubah kata-kata yang memiliki imbuhan menjadi sebuah kata dasar aslinya. Pada proses *stemming* menggunakan bantuan *library Sastrawi* pada bahasa pemrograman *Python.*

TABEL III. *STEMMING*

Sebelum	Sesudah
menginstal virus trojan sistem operasi	install virus trojan sistem operasi
absensi langsung menemui ruangan	absen langsung nemu ruang

4. *Tokenization:* Memecah dokumen menjadi bagian - bagian yang lebih kecil sehingga memudahkan untuk analisis. Pada proses ini menggunakan *library nltk* pada Bahasa pemograman *Python.*

TABEL IV. Tokenization

Sebelum	Sesudah
install virus trojan sistem operasi	[‘install’, ‘virus’, ‘trojan’, ‘sistem’,’operasi’]
absen langsung nemu ruang	[‘absen’, ‘langsung’,’nemu’, ‘ruang’]

C. Ekstraksi Fitur

Tahapan yang dilakukan setelah tahap preprocessing bertujuan untuk mempermudah proses klasifikasi. Tahapan ini menggunakan fitur N-gram yaitu memecah kalimat sesuai dengan nilai n yang telah ditentukan. Misal apabila n=2 maka disebut dengan bigram maka kalimat akan dibagi menjadi dua kata pada setiap bagian. Metode selanjutnya adalah dengan menggunakan *term frequency* yaitu digunakan untuk melakukan perhitungan pembobotan dengan menghitung frekuensi kemunculan term tertentu pada suatu dokumen.[6]

D. Klasifikasi Naïve Bayes

Metode yang digunakan untuk mengklasifikasikan data surel untuk mendapatkan prediksi spam atau non spam adalah *Naïve Bayes*. Setelah melakukan *Preprocessing* dan ekstraksi fitur selanjutnya melakukan *training* dengan jumlah data sebanyak 80% dari jumlah data yang ada dengan menggunakan bantuan *library sklearn* pada bahasa pemrograman *python.*

E. Uji Model

Proses ini dilakukan setelah proses *training* selesai dilakukan. Dengan menguji model menggunakan data testing sebanyak 20% dari seluruh jumlah data. Proses ini memiliki tujuan untuk mengetahui akurasi model yang telah di buat.

F. Evaluasi Model

Evaluasi dan Validasi Hasil dilakukan dengan menghitung nilai *precision* dengan rumus nomor (2), *recall* dengan rumus nomor (3), dan *f-score* dengan rumus nomor (4). *precision* mengukur tingkat kepastian yang diklasifikasikan dengan benar, *recall* rasio prediksi dari data pada setiap label yang diklasifikasikan dengan benar terhadap data yang salah diklasifikasikan ke label lainya. *F-score* merupakan *harmonic mean* antara nilai *precision* dan nilai *recall.*

IV. HASIL DAN PEMBAHASAN

A. Perhitungan Akurasi pada model klasifikasi

Berikut merupakan hasil perhitungan akurasi model dapat dilihat pada Tabel V.

TABEL V. Hasil Akurasi Model Klasifikasi pada data *testing*

N-gram	Akurasi
1	0,90
2	0,94
3	0,93
4	0,92
5	0,92

Berdasarkan Tabel V, Hasil yang didapat menunjukan bahwa model dengan menggunakan algoritma Naïve Bayes dan menggunakan fitur N-gram memiliki akurasi yang cukup baik. Dari percobaan yang dilakukan nilai n=2 memiliki akurasi yang paling tinggi yaitu 94% dibandingkan dengan n=1 memiliki akurasi 90%, n=3 memiliki akurasi 93%, n=4 memiliki akurasi 92%, dan n=5 memiliki akurasi 92%.

B. Perhitungan *Precision, Recall*, dan *F-Score*

Hasil perhitungan evaluasi menggunakan *precision, recall, f-score* dalam penelitian ini dapat dilihat dalam Tabel VI.

TABEL VI PERHITUNGAN *PRECISION, RECALL,* dan *F-SCORE*

N-gram	*Precison*	*Recall*	*F-Score*
1	0,85	0,98	0,91
2	0,96	0,98	0,97
3	0,93	0,96	0,95
4	0,95	0,96	0,95
5	0,95	0,96	0,95

Berdasarkan hasil yang ditampilkan pada Tabel VI, untuk nilai *F-Score* pada table tersebut nilai *F-Score* tertinggi diperoleh n=2 yaitu 97% dan pada nilai *precision* dengan nilai n=2 96%. Pada nilai *recall* untuk n=1, n=2, memiliki nilai yang sama yaitu 97%

C. Hasil Klasifikasi

Berikut merupakan contoh hasil klasifikasi yang dilakukan model yang telah dibuat dapat dilihat pada Tabel VII.

Tabel VII. Contoh Hasil Klasifikasi

Contoh Hasil Klasifikasi Spam
Halo! Saya punya kabar buruk untuk Anda. Dua bulan lalu, saya menerima akses ke semua peranti elektronik yang Anda gunakan untuk menelusuri internet. Setelah itu, saya mulai melacak aktivitas Anda di seluruh web. Sekarang, saya akan ungkapkan bagaimana itu bisa terjadi Saya membuat situs palsu (domain.com) dan mengirim undangan otorisasi. ... Prediksi: Spam
Contoh Hasil klasifikasi Non Spam
Selamat siang pak Rwen, Apakah boleh mengirimkan CVnya sekali lagi? Mungkin menggunakan format PDF? Karena file Ms.Word ini tidak dapat dibuka Terima kasih, Coffee Rider Prediksi: Non Spam

Berdasarkan Tabel VII, Model yang telah dibangun dapat mengklasifikasikan dengan benar.

V. KESIMPULAN

Dari hasil penelitian ini telah berhasil dibangun sebuah model untuk melakukan deteksi surel *Spam* dan *non Spam* berbahasa Indonesia dan Algoritma *Naïve Bayes* memiliki akurasi yang sangat baik. Berdasarkan eksperimen yang dilakukan perpaduan fitur *N-Gram* dengan nilai n=2 dan algoritma *Naïve* Bayes akurasi yang paling tinggi yaitu 94%

REFERENSI

[1] S. N. D. Pratiwi dan B. S. S. Ulama, "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, hal. 344–349, 2016.

[2] R. Y. Hayuningtyas, "Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 2, no. 1, hal. 53–60, 2017.

[3] M. Andriansyah, T. Oswari, dan B. Prijanto, "Crowdsourcing : Konsep Sumber Daya Kerumunan dalam Abad Partisipasi Komunitas Internet," hal. 1–6, 2016.

[4] A. Imron, "ANALISIS SENTIMEN TERHADAP TEMPAT WISATA DI KABUPATEN REMBANG MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER," 2019.

[5] S. A. Sugianto, L. Liliana, dan S. Rostianingsih, "Pembuatan Aplikasi Predictive Text Menggunakan Metode N-gram-based," *J. Infra*, vol. 1, no. 2, hal. 1–6, 2013.

[6] A. F. Hidayatullah dan M. R. Ma'arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) Agustus*, 2016, hal. 1907–5022.

[7] M. Hengki dan M. Wahyudi, "Klasifikasi Algoritma Naïve Bayes dan SVM Berbasis PSO Dalam Memprediksi Spam Email Pada Hotline-Sapto," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 1, hal. 61–67, 2020, doi: 10.31294/p.v22i1.7842.

[8] W. N. Chandra, G. Indrawan, dan I. N. Sukajaya, "Spam Filtering Dengan Metode Pos Tagger Dan Klasifikasi Naïve Bayes," *J. Ilm. Teknol. Inf. Asia*, vol. 10, no. 1, hal. 47–55, 2016.

[9] D. Juang, "Analisis Spam dengan Menggunakan Naïve Bayes," *J. Teknovasi*, vol. 03, no. 1998, hal. 51–57, 2016.

[10] A. Al-Alwani dan M. Beseiso, "Arabic Spam filtering using Bayesian Model," *Int. J. Comput. Appl.*, vol. 79, no. 7, hal. 11–14, 2013, doi: 10.5120/13752-1582.

[11] S. M. Abdulhamid, M. Shuaib, O. Osho, I. Ismaila, dan J. K. Alhassan, "Comparative Analysis of Classification Algorithms for Email Spam Detection," *Int. J. Comput. Netw. Inf. Secur.*, vol. 10, no. 1, hal. 60–67, 2018, doi: 10.5815/ijcnis.2018.01.07.