

KLASIFIKASI SMS SPAM MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Andrei Wara Putera¹, Suriati², Yuyun Dewi Lestari³

^{1,2,3} Program Studi Teknik Informatika, Fakultas Teknik dan Komputer, Universitas Harapan Medan,

e-mail: ¹andre.wara97@gmail.com, ²suriati1980@gmail.com, ³yuyun.dl@gmail.com

ABSTRAK

Penggunaan media *short message service* (SMS) untuk komunikasi masih sangat banyak. Hal itu terjadi dikarenakan beberapa faktor, seperti tarif yang murah, bonus yang diberikan serta kemudahan dalam penggunaan. Namun, faktor-faktor tersebut menjadikan layanan SMS dimanfaatkan untuk melakukan tindakan kriminal, salah satunya adalah SMS penipuan. Untuk mengatasi hal tersebut, diperlukan sebuah sistem yang dapat mengklasifikasi SMS yang termasuk spam atau bukan spam. Dalam penelitian ini *Dataset* SMS yang digunakan adalah *Dataset* SMS berbahasa Indonesia. Untuk pembobotan teks menggunakan metode *TF-IDF* dan *Cosine Similarity* untuk metode perhitungan jarak. Hasil dari penelitian ini berupa aplikasi yang mampu mengklasifikasi SMS Spam Bahasa Indonesia Menggunakan Algoritma *K-Nearest Neighbor*. Hasil pengujian menunjukkan sistem yang dibangun dapat mengklasifikasi SMS sesuai dengan kategori, yaitu normal, penipuan, penipuan dengan baik.

Kata kunci: SMS Spam; *TF-IDF*; *K-Nearest Neighbor*; *Cosine Similarity*

ABSTRACT

The use of Short Message Service (SMS) media for communication is still very much. This is due to several factors, such as low rates, bonuses provided and ease of use. However, these factors make SMS services used to commit criminal acts, one of which is SMS fraud. To overcome this, we need a system that can classify SMS as spam or not spam (ham). In this study, the SMS dataset used is the Indonesian language SMS dataset. For text weighting, use the TF-IDF method and Cosine Similarity for the distance calculation method. The results of this study are an application that is able to classify Indonesian SMS Spam using the K-Nearest Neighbor Algorithm. The test results show that the system built can classify SMS according to categories, namely normal, fraud, fraud well.

Keywords: SMS Spam; *TF-IDF*; *K-Nearest Neighbor*; *Cosine Similarity*

1. PENDAHULUAN

Perkembangan Teknologi Informasi semakin memudahkan manusia untuk melakukan pertukaran informasi satu sama lain. Hal tersebut disebabkan karena munculnya teknologi-teknologi dibidang komunikasi. Salah satu dari teknologi komunikasi tersebut adalah SMS. SMS merupakan layanan pengiriman pesan berupa teks dengan jumlah karakter singkat antar perangkat Telepon Selular [1]. SMS masih banyak digunakan oleh masyarakat sampai sekarang, hal itu terjadi dikarenakan beberapa faktor,

seperti tarif yang murah, bonus yang diberikan serta kemudahan dalam penggunaan [2]–[3].

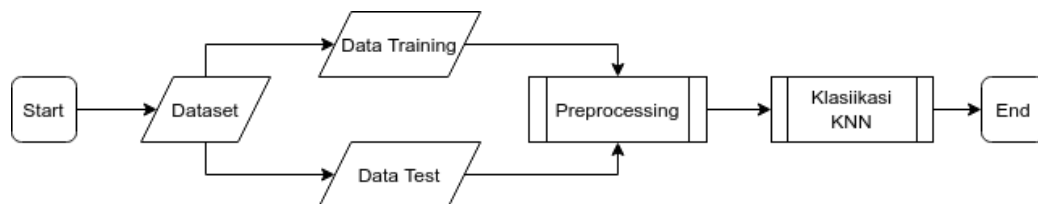
Layanan SMS tidak hanya digunakan untuk menyampaikan pesan kepada sesama pengguna yang sudah saling mengenal, tetapi juga untuk mengirimkan informasi kepada orang yang belum dikenal dengan tujuan untuk menawarkan produk, jasa, dan bahkan tindakan kejahatan [5]. Tindakan tersebut adalah melakukan pengiriman SMS yang tidak dikehendaki atau diistilahkan sebagai *Spam*.

Pengguna Layanan SMS tidak menginginkan SMS spam karena sangat mengganggu bahkan sampai membahayakan sehingga dapat menyebabkan kerugian bagi pengguna layanan SMS. Contoh SMS *Spam* adalah SMS yang biasanya bertujuan untuk pemasaran, promosi, pengiklanan, penipuan dan lain-lain [4]. Untuk mengatasi hal tersebut, diperlukan sebuah sistem yang mampu melakukan klasifikasi untuk membedakan pesan SMS yang mengandung *Spam* dan bukan *Spam*. Sehingga kerugian di sisi pengguna dapat diminimalisir.

Beberapa penelitian sudah dilakukan untuk klasifikasi SMS *Spam*, antara lain : (1) Perbandingan algoritma C4.5, KNN, NBC, dan SVM untuk klasifikasi SMS *spam*, akurasi tertinggi dihasilkan oleh Metode SVM sebesar 94,60%, metode C4.5 sebesar 85,86%, KNN sebesar 77,50% dan NBC sebesar 86,10 % [3]; (2) klasifikasi SMS pada aplikasi SMS *Gateway* yang digunakan oleh LKBN ANTARA menggunakan metode KNN dengan nilai akurasi sebesar 96,15% [6]; (3) klasifikasi *Email Spam* dengan membandingkan kinerja algoritam SVM dan KNN dimana Metode KNN memberikan hasil performansi klasifikasi terbaik saat $k = 3$ dengan hasil akurasi=92.28%, $precision=92.3\%$ dan $recall=92.3\%$ dan $error=7.72\%$ dari total 6000 *email* [7]; (4) Perancangan Aplikasi Deteksi *Spam Twitter* menggunakan Metode *Naive Bayes* dan KNN pada Perangkat Bergerak *Android*, *Naive Bayes* dan *KNN* dapat mendeteksi kandungan *Spam* dan *Ham* masing-masing dengan akurasi 82% dan 71% [8].

2. METODE PENELITIAN

Penelitian ini terdiri dari tahapan-tahapan utama yaitu: *input* penelitian adalah *file dataset* berekstensi .txt yang berisi SMS *spam* yang masing-masing akan dipisahkan menjadi *dataset* pelatihan dan *dataset* pengujian. Kemudian *dataset* pelatihan (*training dataset*) dan *dataset* pengujian (*testing dataset*) akan melalui tahap *preprocessing*, yaitu *tokenizing*, *case folding*, *stopword removal*, *stemming*. Setelah tahapan tersebut dilakukan, maka *data testing* akan diidentifikasi menggunakan metode KNN, dengan *output* berupa hasil dari teks SMS yang berhasil dikenali. Untuk lebih jelasnya, tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Dataset

Data yang digunakan dalam penelitian ini adalah *Dataset* SMS Spam bahasa Indonesia dan *Dataset* SMS spam lainnya dari sumber yang relevan [9]. *Dataset* tersebut terdiri dari 3 kelas yaitu sms normal, penipuan, dan promo. Data yang digunakan sebanyak 50 data dan dipilih secara acak. Contoh data dari masing-masing kelas dapat dilihat pada Tabel 1

Tabel 1 Contoh Dataset SMS

Data SMS	Label
Maaf jika ada janji yg belum terpenuhi, jika ada janji boleh mengingatkan saya.	0
Dana Tunai (KTA) bunga 0,99% hingga 300 jt. Syarat KTP & CC. Bisa Dgn BPKB rate 0.99% 3 hari CAIR . hub AYU 081584650877 (WA).	1
2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat Anda yang terpilih. Aktifkan sekarang juga di *550*905#. Promo sd 30 Nov 2015. Buruan aktifkan sekarang.	2

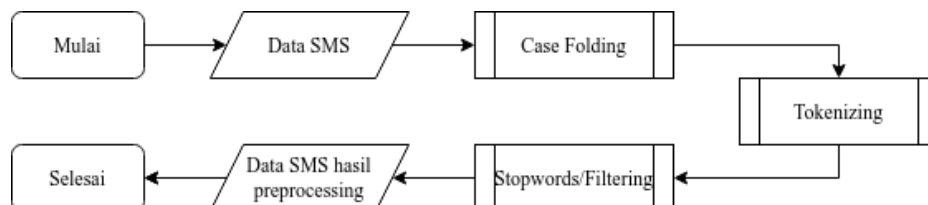
Sebagai contoh dapat dilihat pada Tabel 2, terdapat 6 dokumen SMS (dokumen 1 sampai dengan dokumen 6) sebagai data latih dan satu buah dokumen SMS (dokumen 7) sebagai data uji sebagai berikut:

Tabel 2 Dataset SMS

No.	SMS
1.	ayah isikan pulsa xl 50rb di no 0818643929 sekarang penting
2.	Belikan dulu mama pulsa simpati 20 ribu ini nomor barunya mama 081223052854. kalau bisa kirim sekarang secepatnya penting. Ini hpnya orang mama pinjam
3.	"Di transfer saja Uang'nya ke rekening BNI ini, a/n. ISMAWATI. Rek: 0248147033. SMS saja klo sdh kirim."
4.	Aku senin udah ke tempat kerja. Minggu2 depan aku gaktau bisa/ngga :(
5.	Aku teh cuma bawa celana jeans yg tua tp jilbabnya yg muda -_
6.	"Ayam aja Pak, terus satu paket itu sama isinya ap aja?"
7.	"Ini bpk tolong belikan dulu pulsa 20ribu di no barunya bpk karna lagi ada masalah di kantor polisi dan jangan dulu tlp/sms nanti bpk yg tlp, penting."

Preprocessing

Dataset SMS yang akan digunakan harus melalui tahapan *preprocessing* terlebih dahulu. Tahapan *preprocessing* dapat dilihat pada Gambar 2.



Gambar 2. Tahapan Preprocessing

Case Folding

Pada tahapan ini teks diubah menjadi huruf kecil. Hasil dari tahapan *case folding* teks SMS pada Tabel 3 dapat dilihat pada Tabel 4.

Tabel 4 Hasil Case Folding

No.	SMS
1.	ayah isikan pulsa xl ribu di nomor sekarang penting
2.	belikan dulu mama pulsa simpati ribu ini nomor barunya mama kalau bisa kirim sekarang secepatnya penting ini hpnya orang mama pinjam
3	di transfer saja uangnya ke rekening bni ini atas nama ismawati rekening sms saja kalau sudah kirim
4.	aku senin sudah ke tempat kerja minggu depan aku tidak tau bisa tidak
5.	aku teh cuma bawa celana jeans yang tua tapi jilbabnya yang muda
6.	ayam saja pak terus satu paket itu sama isinya apa saja
7.	ini bapak tolong belikan dulu pulsa ribu di nomor barunya bapak karena lagi ada masalah di kantor polisi dan jangan dulu telepon sms nanti bapak yang telepon penting

Tokenizing

Karakter selain huruf, rangkaian angka, atau rangkaian angka dengan huruf akan dihilangkan. Setiap kata, rangkaian angka, maupun rangkaian angka dengan huruf disebut sebagai token. Hasil dari tahapan *tokenizing* dapat dilihat pada Tabel 5

Tabel 5 Hasil Tokenizing

No.	Terma yang mewakili dokumen
1	ayah isikan pulsa xl ribu di nomor sekarang penting
2	belikan dulu mama pulsa simpati ribu ini nomor barunya mama kalau bisa kirim sekarang secepatnya penting ini hpnya orang mama pinjam
3	di transfer saja uangnya ke rekening bni ini atas nama ismawati rekening sms saja kalau sudah kirim
4	aku senin sudah ke tempat kerja minggu depan aku tidak tau bisa tidak
5	aku teh cuma bawa celana jeans yang tua tapi jilbabnya yang muda
6	ayam saja pak terus satu paket itu sama isinya apa saja
7	ini bapak tolong belikan dulu pulsa ribu di nomor barunya bapak karena lagi ada masalah di kantor polisi dan jangan dulu telepon sms nanti bapak yang telepon penting

Stopword Removal

Beberapa kata yang *termasuk* dalam daftar *stopwords* adalah yang, di, ke, dari, adalah, dan, atau, dan lain sebagainya. Hasil dari tahap ini dapat dilihat pada Tabel 6.

Tabel 6 Hasil Stopword Removal

No.	Terma yang mewakili dokumen
1.	ayah isikan pulsa xl nomor penting
2.	belikan mama pulsa simpati nomor barunya mama kirim secepatnya penting hpnya orang mama pinjam

3. transfer uangnya rekening bni nama rekening sms kirim
4. senin tempat kerja minggu
5. cuma bawa celana jeans tua jilbabnya muda
6. ayam satu paket isinya
7. tolong belikan pulsa nomor barunya masalah kantor polisi telepon sms telepon penting

Stemming

Proses *stemming* merupakan lanjutan dari proses *stopword removal* pada Tabel 6. Hasil proses *stemming* dapat dilihat pada Tabel 7

Tabel 7 Hasil Stemming

No.	Term yang mewakili dokumen
1.	ayah isi pulsa xl nomor penting
2.	beli mama pulsa simpati nomor baru mama kirim cepat penting hp orang mama pinjam
3.	transfer uang rekening bni rekening sms kirim
4.	senin tempat kerja minggu
5.	bawa celana jeans tua jilbab muda
6.	ayam paket isi
7.	tolong beli pulsa nomor baru masalah kantor polisi telepon sms telepon penting

Pembobotan Kata dengan TF-IDF

Proses pembobotan *term* dilakukan pada seluruh *data training* dan *data Testing*. *Data training* diberi label D1 hingga D6 dan data uji diberi label Q. Hasilnya diperlihatkan pada Tabel 8. Proses pembobotan kata didahului dengan melakukan proses perhitungan nilai TF-IDF dengan menggunakan rumus:

$$w_{ij} = tf_{ij} \times \log\left(\frac{D}{df_i}\right) \quad (1)$$

Keterangan:

w_{ij} = Bobot kata ke-i pada dokumen ke-j

tf_{ij} = merupakan jumlah *term* ke-i pada dokumen ke-j

D = Jumlah dokumen keseluruhan

df_i = merupakan jumlah dokumen yang mengandung *term* ke-I

Tabel 8 Pelabelan Dokumen

Dokumen	Term yang mewakili dokumen	Label
D1	ayah isi pulsa xl nomor penting	Penipuan
D2	beli mama pulsa simpati nomor baru mama kirim cepat penting hp orang	Penipuan
D3	transfer uang rekening bni rekening sms kirim	Penipuan
D4	senin tempat kerja minggu	normal
D5	bawa celana jeans tua jilbab muda	normal

D6	ayam paket isi	normal
Q	tolong beli pulsa nomor baru masalah kantor polisi telepon sms telepon penting	?

Banyaknya jumlah kata atau *term frequency* (tf) yang muncul untuk kata “ayah” pada setiap dokumen adalah sebagai berikut: tf(ayah) Q = 0, tf(ayah) D1 = 1, tf(ayah) D2 = 0, tf(ayah) D3 = 0, tf(ayah) D4 = 0, tf(ayah) D5 = 0, tf(ayah) D6 = 0 sehingga nilai df(ayah) dapat dihitung sebagai berikut:

$$df(ayah) = \sum tf$$

$$df(ayah) = tfQ + tfD1 + tfD2 + tfD3 + tfD4 + tfD5 + tfD6$$

$$df(ayah) = 0 + 1 + 0 + 0 + 0 + 0 + 0$$

$$df(ayah) = 1$$

Nilai idf untuk *term* “ayah” dapat dihitung seperti berikut ini

$$idf(ayah) = \log(n/df)$$

$$idf(ayah) = \log(7/1)$$

$$idf(ayah) = \log(6)$$

$$idf(ayah) = 0,845098$$

Hasil perhitungan TF-IDF dari seluruh *term* pada dokumen dapat dilihat pada Tabel 9

Tabel 9 Hasil Perhitungan TF-IDF

<i>Term</i>	Q	D1	D2	D3	D4	D5	D6	df= \sum tf	idf= $\log(n/df)$
ayah	0	1	0	0	0	0	0	1	0,845098
ayam	0	0	0	0	0	0	1	1	0,845098
bawa	0	0	0	0	0	1	0	1	0,845098
beli	1	0	1	0	0	0	0	2	0,544068
bni	0	0	0	1	0	0	0	1	0,845098
celana	0	0	0	0	0	1	0	1	0,845098
cepat	0	0	1	0	0	0	0	1	0,845098
hp	0	0	1	0	0	0	0	1	0,845098
isi	0	1	0	0	0	0	1	2	0,544068
jeans	0	0	0	0	0	1	0	1	0,845098
jilbab	0	0	0	0	0	1	0	1	0,845098
kantor	1	0	0	0	0	0	0	1	0,845098
kerja	0	0	0	0	1	0	0	1	0,845098
kirim	0	0	1	1	0	0	0	2	0,544068
mama	0	0	3	0	0	0	0	3	0,367977
minggu	0	0	0	0	1	0	0	1	0,845098
muda	0	0	0	0	0	1	0	1	0,845098
nomor	1	1	1	0	0	0	0	3	0,367977
orang	0	0	1	0	0	0	0	1	0,845098
paket	0	0	0	0	0	0	1	1	0,845098
pinjam	0	0	1	0	0	0	0	1	0,845098
polisi	1	0	0	0	0	0	0	1	0,845098
pulsa	1	1	1	0	0	0	0	3	0,367977
rekening	0	0	0	2	0	0	0	2	0,544068

Beli	0.000	0.296	0.000	0.000	0.000	0.000
Bni	0.000	0.000	0.000	0.000	0.000	0.000
celana	0.000	0.000	0.000	0.000	0.000	0.000
cepat	0.000	0.000	0.000	0.000	0.000	0.000
Hp	0.000	0.000	0.000	0.000	0.000	0.000
Isi	0.000	0.000	0.000	0.000	0.000	0.000
jeans	0.000	0.000	0.000	0.000	0.000	0.000
jilbab	0.000	0.000	0.000	0.000	0.000	0.000
kantor	0.000	0.000	0.000	0.000	0.000	0.000
kerja	0.000	0.000	0.000	0.000	0.000	0.000
kirim	0.000	0.000	0.000	0.000	0.000	0.000
mama	0.000	0.000	0.000	0.000	0.000	0.000
Minggu	0.000	0.000	0.000	0.000	0.000	0.000
muda	0.000	0.000	0.000	0.000	0.000	0.000
nomor	0.135	0.135	0.000	0.000	0.000	0.000
orang	0.000	0.000	0.000	0.000	0.000	0.000
paket	0.000	0.000	0.000	0.000	0.000	0.000
pinjam	0.000	0.000	0.000	0.000	0.000	0.000
polisi	0.000	0.000	0.000	0.000	0.000	0.000
pulsa	0.135	0.135	0.000	0.000	0.000	0.000
rekening	0.000	0.000	0.000	0.000	0.000	0.000
ribu	0.135	0.135	0.000	0.000	0.000	0.000
senin	0.000	0.000	0.000	0.000	0.000	0.000
simpati	0.000	0.000	0.000	0.000	0.000	0.000
sms	0.000	0.000	0.296	0.000	0.000	0.000
The	0.000	0.000	0.000	0.000	0.000	0.000
tempat	0.000	0.000	0.000	0.000	0.000	0.000
Tlp	0.000	0.000	0.000	0.000	0.000	0.000
tolong	0.000	0.000	0.000	0.000	0.000	0.000
transfer	0.000	0.000	0.000	0.000	0.000	0.000
Tua	0.000	0.000	0.000	0.000	0.000	0.000
uang	0.000	0.000	0.000	0.000	0.000	0.000
Xl	0.000	0.000	0.000	0.000	0.000	0.000
Jumlah	0.4062	0.7022	0.2960	0	0	0

setelah dilakukan perhitungan perkalian skalar kemudian dilakukan proses perhitungan panjang vektor untuk masing-masing dokumen. Hasil perhitungan panjang vektor keseluruhan dapat dilihat pada Tabel 12

Tabel 12 Hasil Perhitungan Panjang Vektor

<i>Term</i>	Q	D1	D2	D3	D4	D5	D6
Ayah	0.000	0.714	0.000	0.000	0.000	0.000	0.000
Ayam	0.000	0.000	0.000	0.000	0.000	0.000	0.714
Bawa	0.000	0.000	0.000	0.000	0.000	0.714	0.000
Beli	0.296	0.000	0.296	0.000	0.000	0.000	0.000
Bni	0.000	0.000	0.000	0.714	0.000	0.000	0.000
celana	0.000	0.000	0.000	0.000	0.000	0.714	0.000
cepat	0.000	0.000	0.714	0.000	0.000	0.000	0.000
Hp	0.000	0.000	0.714	0.000	0.000	0.000	0.000
Isi	0.000	0.296	0.000	0.000	0.000	0.000	0.296
jeans	0.000	0.000	0.000	0.000	0.000	0.714	0.000
jilbab	0.000	0.000	0.000	0.000	0.000	0.714	0.000
kantor	0.714	0.000	0.000	0.000	0.000	0.000	0.000
kerja	0.000	0.000	0.000	0.000	0.714	0.000	0.000
kirim	0.000	0.000	0.296	0.296	0.000	0.000	0.000
mama	0.000	0.000	1.219	0.000	0.000	0.000	0.000
Minggu	0.000	0.000	0.000	0.000	0.714	0.000	0.000
muda	0.000	0.000	0.000	0.000	0.000	0.714	0.000
nomor	0.135	0.135	0.135	0.000	0.000	0.000	0.000
orang	0.000	0.000	0.714	0.000	0.000	0.000	0.000
paket	0.000	0.000	0.000	0.000	0.000	0.000	0.714
pinjam	0.000	0.000	0.714	0.000	0.000	0.000	0.000
polisi	0.714	0.000	0.000	0.000	0.000	0.000	0.000
pulsa	0.135	0.135	0.135	0.000	0.000	0.000	0.000
rekening	0.000	0.000	0.000	1.184	0.000	0.000	0.000
ribu	0.135	0.135	0.135	0.000	0.000	0.000	0.000
senin	0.000	0.000	0.000	0.000	0.714	0.000	0.000
simpati	0.000	0.000	0.714	0.000	0.000	0.000	0.000
sms	0.296	0.000	0.000	0.296	0.000	0.000	0.000
teh	0.000	0.000	0.000	0.000	0.000	0.714	0.000
tempat	0.000	0.000	0.000	0.000	0.714	0.000	0.000
tlp	1.184	0.000	0.000	0.000	0.000	0.000	0.000
tolong	0.714	0.000	0.000	0.000	0.000	0.000	0.000
transfer	0.000	0.000	0.000	0.714	0.000	0.000	0.000
tua	0.000	0.000	0.000	0.000	0.000	0.714	0.000
uang	0.000	0.000	0.000	0.714	0.000	0.000	0.000
xl	0.000	0.714	0.000	0.000	0.000	0.000	0.000
Jumlah	4.325	2.131	5.788	3.919	2.857	4.999	1.724
Akar	2.080	1.460	2.406	1.980	1.690	2.236	1.313

Langkah selanjutnya adalah menghitung jarak menggunakan *Cosine Similarity* dari Q dengan D1, D2 dan seterusnya sampai dengan D6, sebagai berikut:

$$\begin{aligned} \text{Cos (Q, D1)} &= 0.4062 / (2.080 * 1.460) = 0 \\ \text{Cos (Q, D2)} &= 0.7022 / (2.080 * 2.406) = 0 \\ \text{Cos (Q, D3)} &= 0.2960 / (2.080 * 1.980) = 0 \\ \text{Cos (Q, D4)} &= 0 / (2.080 * 1.690) = 0 \\ \text{Cos (Q, D5)} &= 0 / (2.080 * 2.236) = 0 \\ \text{Cos (Q, D6)} &= 0 / (2.080 * 1.313) = 0 \end{aligned}$$

Untuk hasil perhitungan jarak keseluruhan dapat dilihat pada Tabel 13

Tabel 13 Jarak Tetangga Dengan *Cosine Similarity*

Jarak	(Q, D1)	(Q, D2)	(Q, D3)	(Q, D4)	(Q, D5)	(Q, D6)
Nilai	0.134	0.140	0.072	0	0	0

Setelah mendapatkan jumlah tetangga terdekat dengan *Cosine Similarity*, lalu menentukan jumlah tetangga terdekat (jumlah *K*). Pada penelitian ini digunakan *K* = 1. Tetangga terdekat yaitu (Q, D2) = 0.140. Berdasarkan kategori, Dokumen D2 berada dalam kategori Penipuan, maka data uji Q termasuk ke dalam kategori SMS penipuan.

3. HASIL DAN PEMBAHASAN

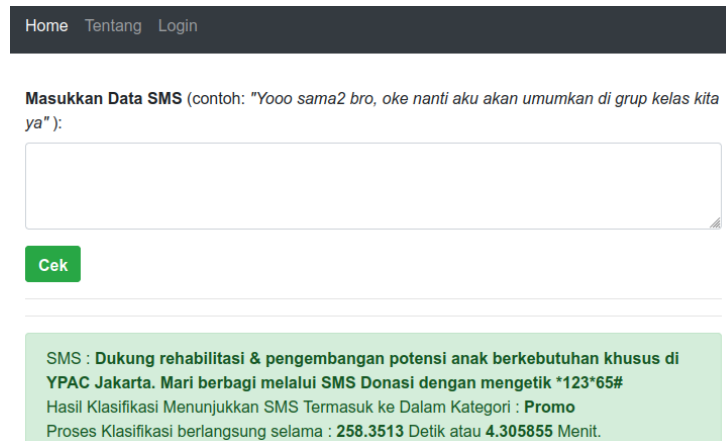
Hasil pengujian dilakukan untuk melihat sejauh mana sistem yang dibangun sesuai dengan proses perancangan yang telah dilakukan. Data SMS yang akan diuji dimasukkan melalui form pengujian (Gambar 3).

Gambar 3 Form input data SMS

Hasil pengujian klasifikasi SMS untuk kategori normal dapat dilihat pada Gambar 4.

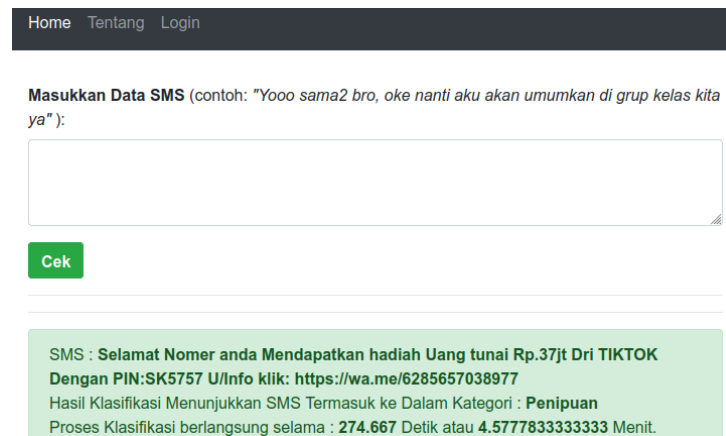
Gambar 4 Hasil pengujian SMS kategori normal

Hasil pengujian klasifikasi SMS untuk kategori Promosi dapat dilihat pada Gambar 5



Gambar 5 Hasil pengujian SMS kategori promosi

Hasil pengujian klasifikasi SMS untuk kategori Penipuan dapat dilihat pada Gambar 6



Gambar 6 Hasil pengujian SMS kategori penipuan

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, Algoritma KNN dapat digunakan untuk mengklasifikasi SMS Spam Bahasa Indonesia dengan baik. Aplikasi web klasifikasi SMS Spam Bahasa Indonesia menggunakan algoritma KNN berhasil diimplementasikan. Hasil pengujian menunjukkan sistem yang dibangun dapat mengklasifikasi SMS SPAM sesuai dengan kategori, yaitu normal, penipuan, penipuan dengan baik.

DAFTAR PUSTAKA

- [1] I. K. Siregar and F. Taufik, "PERANCANGAN APLIKASI SMS ALERT BERBASIS WEB," *JIMP - J. Inform. Merdeka Pasuruan*, vol. 2, no. 2, 2017, doi: 10.37438/jimp.v2i2.68.

- [2] R. Minarni, "Implementasi Algoritma Base64 Untuk Mengamankan Sms Pada Smartphone," *Build. Informatics, Technol. Sci.*, vol. 1, no. 1, 2019, doi: 10.47065/bits.v1i1.3.
- [3] C. S. Wahyuni and M. Munar, "Aplikasi Pemilihan Kepala Desa Di Kecamatan Gandapura Menggunakan Sms Gateway Dan E-Voting," *J. TIKA*, vol. 6, no. 01, 2021, doi: 10.51179/tika.v6i01.406.
- [4] E. Zuviyanto, T. B. Adji, and N. A. Setiawan, "Perbandingan Algoritme-Algoritme Pembelajaran Mesin pada Klasifikasi SMS Spam," in *Seminar Nasional Inovasi dan Aplikasi Teknologi di Industri*, 2018.
- [5] B. Indiarso, "Klasifikasi Sms Spam Dengan Metode Naive Bayes Classifier Untuk Menyaring Pesan Melalui Selular," *J. Telemat. MKOM*, vol. 8, no. 2, 2016.
- [6] Subhan Subhan, "Klasifikasi Konten Web Radikal Di Indonesia menggunakan Web Content Mining Dan Algoritma K-Nearest Neighbor," *J. Informasi, Sains dan Teknol.*, vol. 4, no. 2, 2021, doi: 10.55606/isaintek.v4i2.3.
- [7] S. N. D. Pratiwi and 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, 2016.
- [8] F. A. Prayoga, A. Pinandito, and R. S. Perdana, "Rancang Bangun Aplikasi Deteksi Spam Twitter menggunakan Metode Naive Bayes dan KNN pada Perangkat Bergerak Android," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, 2018.
- [9] Y. Wibisono, "Dataset Klasifikasi Bahasa Indonesia (SMS Spam) & Klasifikasi Teks dengan Scikit-Learn," 2018. <https://yudiwbs.wordpress.com/2018/08/05/dataset-klasifikasi-bahasa-indonesia-sms-spam-klasifikasi-teks-dengan-scikit-learn/> (accessed Sep. 10, 2022).
- [10] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/796>