REPOSITOR, Vol. 3, No. 4, Agustus 2021, Pp. 387-392
ISSN : 2714-7975
E-ISSN : 2716-1382

Analisis Klasifikasi SMS Spam Menggunakan *Logistic Regression*

Ferin Reviantika*[1], Yufis Azhar[2], Gita Indah Marthasari[3]
[1,2,3]Universitas Muhammadiyah Malang
ferinurin27@gmail.com*[1], yufis@umm.ac.id[2], gita@umm.ac.id[3]

Abstrak

SMS atau Short Message Service biasa nya terdapat pada telepon seluler. SMS dibagi menjadi dua kategori yaitu SMS spam dan SMS non spam (ham). SMS spam adalah SMS yang bersifat mengganggu pengguna telepon karena cenderung berisi pesan yang tidak penting seperti promo dan penipuan. Sedangkan SMS non spam (ham) cenderung berisi SMS yang penting, seperti sudah ada riwayat pesan dengan pengguna sebelumnya. Pada penelitian ini dilakukan klasifikasi SMS spam dan SMS non spam (ham) menggunakan metode logistic regression. Tujuan dari penelitian ini untuk membedakan atau mengklasifikasikan antara SMS spam dan non spam (ham). Dataset dalam penelitian ini berjumlah 1143 data, terdapat dua kolom yaitu kolom teks dan kolom label. Jumlah untuk pesan spam sebanyak 566 pesan dan jumlah untuk pesan non spam sebanyak 577. Metode yang diusulkan mendapat akurasi yang lebih baik yaitu 95%.

Kata Kunci: *SMS, SMS spam, SMS non spam (ham), Logistic regression, Klasifikasi.*

Abstract

SMS or Short Message Service is usually found on cell phones. SMS is divided into two categories, namely spam SMS and non spam SMS (ham). Spam SMS is SMS that is annoying to telephone users because it tends to contain unnecessary messages such as promos and scams. Meanwhile, non-spam SMS (ham) tends to contain SMS that are important, such as there has been a history of messages with previous users. In this study, the classification of SMS spam and SMS non-spam (ham) was carried out using logistic regression methods. The dataset used in this study amounted to 1143 consisting of two columns, namely the text column and the label column. The number for spam messages is 566 messages. Meanwhile, the number for non-spam messages was 577. The proposed method got better accuracy, namely 94%.

Keywords: *SMS, SMS spam, SMS non spam (ham), Logistic regression, Classification*

1. Pendahuluan

Short Message Service atau disingkat dengan SMS merupakan salah satu bagian dari layanan pesan perusahaan komunikasi yang ada pada telepon seluler atau handphone [1]. Rata – rata SMS hanya dapat mengirimkan pesan berupa teks dan hanya dapat digunakan untuk panggilan telepon. Fasilitas SMS pada telepon selular tidak hanya dapat bertukar pesan dengan orang terdekat saja, namun juga mengirimkan pesan kepada orang yang belum mengenal satu sama lain, seperti menawarkan produk, promosi, jasa dan masih banyak lagi. SMS juga banyak digunakan untuk kasus penipuan.

SMS merupakan salah satu sistem yang memiliki jumlah pengguna terbesar didunia. Di Indonesia sendiri pengguna SMS juga tergolong masih tinggi. Dikutip dari https://www.cnbcindonesia.com bahwa Indonesia punya 37 juta pelanggan yang teregistrasi dan 25 – 28 juta pengguna aktif. Tetapi SMS masih memiliki kelemahan dalam keamanan sistem. SMS dibagi menjadi dua kategori yaitu SMS spam dan SMS non spam atau ham. SMS spam adalah pesan – pesan yang tidak bermanfaat, sedangkan SMS non spam atau ham adalah pesan – pesan asli atau bersifat penting [1]. Hal ini memicu kekhawatiran dalam ruang lingkup pribadi karena biasanya SMS digunakan untuk pertukar pesan secara pribadi atau rahasia. Semakin banyaknya SMS spam yang tidak berguna sehingga terkadang mengganggu dan membuat ketidaknyamanan dalam penggunaan telepon seluler.

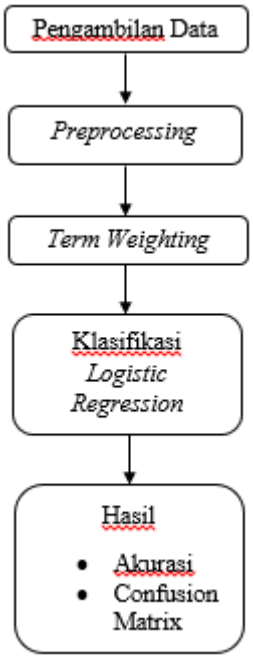
Klasifikasi adalah proses mencari model dengan cara mengategorisasikan sehingga dapat memprediksi kelas yang tidak berlabel [2]. Penelitian tentang klasifikasi SMS sudah banyak

dilakukan seperti penelitian terdahulu oleh Ika Novita dkk tentang klasifikasi teks spam menggunakan metode naïve bayes diperoleh hasil berupa akurasi sebesar 84,40%, recall 88.09% dan precision 45.76% [3]. Penelitian oleh Basuki dkk tentang klasifikasi teks *tweets* tindak kejahatan menggunakan berbahasa Indonesia dengan metode naïve bayes diperoleh hasil klasifikasi menggunakan metode *lexical* atau *bag of words* sebesar 79,25% sedangkan pada fitur fitur sintaktik adalah sebesar 88,1398% [4]. Penelitian oleh Sravya dkk yang berjudul *mobile sms spam filter techniques using machine learning techniques* dengan dilakukan menggunakan banyak metode salah satunya menggunakan *Logistic Regression*, akurasi yang didapatkan sangat tinggi yaitu 97,8% [5].

Berdasarkan uraian diatas dapat dilihat metode terdahulu yaitu *naïve bayes* masih memiliki hasil akurasi yang kurang tinggi dalam melakukan klasifikasi antara SMS spam dan non spam (ham). Dan metode *logistic regression* memiliki hasil yang lebih baik. Hal tersebut yang memicu dilakukannya penelitian ini untuk mempelajari klasifikasi SMS spam menggunakan metode yang berbeda, yaitu Logistic Regression [6].

2. Metode Penelitian

Metode penelitian adalah alur penelitian yang dilakukan pada penelitian menggunakan *logistic regression* untuk klasifikasi SMS spam, dimana hal ini meliputi beberapa tahapan dimulai dari *literature view*, pengambilan data, *preprocessing,* TF – IDF, analisis hasil, dan uji coba dataset. Gambar 1 berikut adalah gambaran penelitian.



Gambar 1. Flowchart dan tahap penelitian

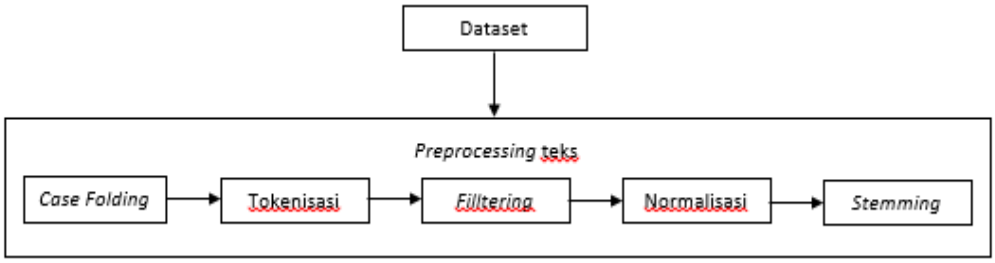
2.1 Pengambilan Data

Pada penelitian ini data diperoleh dari https://github.com yang berjudul klasifikasi spam SMS milik Kuncahyo Setyo Nugroho. Data ini di upload pada tanggal 15 November 2019 sebanyak 1143 *record.* Data terdiri dari dua atribut yaitu atribut nomor, atribut label yang dibedakan menjadi tiga kelas yaitu normal diindikasikan dengan nilai 0, penipuan diindikasikan dengan nilai 1, promo diindikasikan dengan nilai 2 dan atribut *message* atau contoh pesan SMS. Karena penelitian ini bertujuan untuk klasifikasi SMS spam dan non spam (ham), sehingga menyatukan kategori dari pesan spam dimana dalam dataset dijelaskan pada kolom label bernilai 1 dan 2 digabungkan menjadi berlabel 1 yaitu spam.

2.2 *Prepocessing*

Preprocessing merupakan proses awal dalam pengolahan data masukan sehingga akan menghasilkan data dengan format yang sesuai dan siap diproses pada tahap selanjutnya [7].

Tujuan dilakukannya *preprocessing* agar data yang digunakan lebih presisi dan memudahkan dalam menjalankan klasifikasi. Tahapan *preprocessing* dapat dilihat dalam Gambar 2.



Gambar 2. Tahap Preprocessing

1. *Case Folding*

Pada tahap penelitian ini akan mengubah atau mengkonversi *text* menjadi bentuk standart. Sehingga data masukan atau data primer akan diubah menjadi *lowercase* atau huruf kecil [8].

2. Tokenisasi

Untuk memudahkan *preprocess text* dalam tahap tokenisasi memisahkan teks menjadi token token [9]. Sehingga perlu menghapus tanda baca yang tidak perlu, menghapus semua *extra space,* menghapus semua kata yang berjumlah 1*, line breaks,* dan *tabs.*

3. *Filltering*

Dalam proses *Filltering* dilakukan proses *stopwords*, proses *stopword* akan menghapus kata-kata yang tidak memiliki arti yang signifikan atau biasa disebut *meaning less* seperti “ny”,”gk” ,“dg” , “sih” , “hehe” dll [10].

4. Normalisasi

Pada tahap ini akan dilakukan penyeragaman kata yang memiliki makna yang sama, dapat diakibatkan dari penulisan yang salah, penyingkatan kata, ataupun bahasa gaul [11].

5. *Stemming*

Proses *stemmer* yaitu mengembalikan kata yang berimbuhan menjadi kata induk [12].

2.3 *Term Weighting*

Pada Persamaan 1, *term weighting* pada penelitian ini adalah menggunakan nilai TF dan IDF. TF (*Term frequency)* adalah metode sederhana dalam pembobotan setiap *term* [13].

$$tf_{ij} = \frac{f_d(i)}{\mathrm{N}\, f_d\,(j)} \tag{1}$$

Sedangkan pada Persamaan 2, IDF adalah perhitungan *term* yang di distribusikan dalam sebuah dokumen [14].

$$idf(t,d) = \log\left(\frac{N}{df(t)}\right) \tag{2}$$

2.4 Klasifikasi *Logistic Regression*

Logistic Regression adalah bagian dari metode data mining yang penggunaannya untuk menganalisis data yang mendeskripsikan antara satu variabel respon (dependen) atau lebih variabel *predictor* [15]. Metode ini digunakan saat variabel *predictor* (y) memiliki skala kategorik atau nominal yang terdiri dari dua (biner) atau lebih kategori. Sehingga metode ini dibuat untuk memastikan bahwa, apa pun perkiraan yang terjadi, selalu berada di antara 0 dan 1 [16]. Model nilai *w* dapat dilihat pada Persamaan 3.

$$W = b + w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n \tag{3}$$

Pada Persamaan 4, nilai *w* yang diperoleh kemudian dipetakan menggunakan fungsi *logistic regression.*

$$P = \frac{1}{1+e^{-(w)}} \tag{4}$$

Persamaan 4 merupakan persamaan umum yang dapat diubah menjadi Persamaan 5.

$$P = \frac{1}{1+e^{-(b+w_1x_1+w_2x_2+w_3x_3+\cdots+w_nx_n)}} \quad (5)$$

2.5 Evaluasi

Dari proses klasifikasi akan diperoleh hasil yang akan sajikan dalam bentuk *confusion matrix,* akurasi , *precision, recall,* dan *f1-score. Confusion matrix* merupakan hasil dari proses klasifikasi berupa visualisasi data yang benar atau salah diprediksi [17]. Pada penelitian ini berisi hasil klasifikasi pesan spam dan non spam (ham). Nilai akurasi dihitung dari perhitungan jumlah prediksi benar yang sesuai (TP) ditambah jumlah prediksi benar tidak sesuai (TN) dibandingkan dengan jumlah prediksi benar yang sesuai (TP), jumlah prediksi benar tidak sesuai (TF), jumlah prediksi salah yang sesuai (FP) dan jumlah prediksi salah tidak sesuai (FN) [3], seperti pada Persamaan 6, Persamaan 7, dan Persamaan 8.

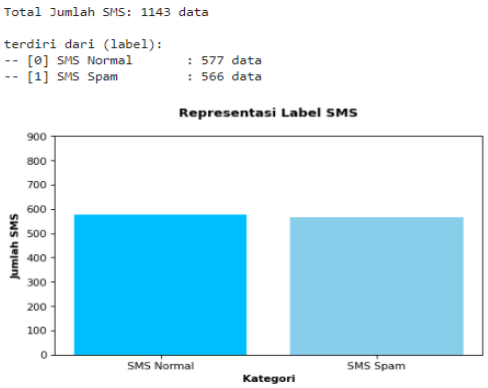
$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+TN} * 100\% \quad (6)$$

$$\text{Presisi} = \frac{TP}{TP+FP} * 100\% \quad (7)$$

$$\textit{Recall} = \frac{TP}{TP+FN} * 100\% \quad (8)$$

3. Hasil Penelitian dan Pembahasan

Penelitian ini menggunakan data sejumlah 1143 *record,* yang dibagi menjadi 577 data sebagai SMS non spam (ham) dan 566 data sebagai SMS spam. Seperti pada Gambar 3.



Gambar 3. Dataset SMS

Penelitian pada Gambar 4 ini juga menampilkan visualisasi kata. Visualisasi kata digunakan untuk mengetahui seberapa sering kata tersebut muncul dalam data. Sehingga semakin besar *font size* maka semakin sering kata tersebut muncul. Dimana kata nomor, promo, info dan hub adalah kata yang sering muncul pada SMS spam. Sedangkan kata nama, sudah, pakai, dan besok adalah kata yang sering muncul pada SMS non spam.

Gambar 4. Visualisasi kata yang sering mucul

Dari Gambar 5 dan Gambar 6, metode *logistic regression* menghasilkan nilai TP atau pesan yang benar diprediksi sebagai pesan Ham sebanyak 110, FP atau pesan yang sebenarnya bernilai Ham namun diprediksi Spam sebanyak 6, FN atau pesna yang bernilai Spam namun diprediksi Ham sebanyak 5,dan TN atau pesan yang benar diprediksi sebagai pesan Spam sebanyak 108. Sehingga dapat dapat diketahui nilai *precision, recall* dan *f1-score*.

	Predicted Ham	Predicted Spam
Actual Ham	110	6
Actual Spam	5	108

Gambar 5. Confusion matrix

	precision	recall	f1-score
0	0.96	0.95	0.95
1	0.95	0.96	0.95
accuracy			0.95

Gambar 6. Classification report

Akurasi adalah nilai dari perbandingan kasus yang mengidentifikasi benar dari jumlah kelas [18]. Sedangkan *precision* dan *recall* mengukur ketepatan dan kelengkapan model klasifikasi dalam penelitian ini [19]. Yang terakhir adalah *f1-score* adalah perbandingan nilai rata-rata dari nilai *precision* dan nilai *recall* [18]. Pada gambar 5 dapat dilihat hasil dari *precision, recall* dan *f1-score.* Nilai 0 pada gambar 5 menandakan hasil untuk pesan non spam (ham) dan nilai 1 menandakan hasil untuk pesan spam. Hasil dari *Classification report logistic regression* juga dapat dilihat pada tabel 1.

Penelitian ini juga membandingkan metode yang diusulkan dengan metode lain seperti *naïve bayes*. Dari Tabel 1 dapat dilihat metode *logistic regression* memiliki hasil yang lebih unggul dibandingkan metode *naïve bayes* diukur dari *precision*, *recall*, f1-*score*, dan *Accuracy*.

Tabel 1. Hasil perbandingan metode

Metode	*Precision* (%)		*Recall* (%)		F1-*score* (%)		*Accuracy* (%)
	0	1	0	1	0	1	
LR	0.96	0.95	0.95	0.96	0.95	0.95	0.95
NB	0.95	0.91	0.91	0.96	0.93	0.93	0.93

4. Kesimpulan

Berdasarkan hasil penelitian klasifikasi SMS spam dan Non spam dengan menggunakan metode *logistic regression,* didapatkan hasil yang baik dan tingkat akurasi yang tinggi pada metode *logistic regression* pada proses pengujian klasifikasi. Dengan menggunakan perbandingan data *train* dan *test* sebanyak 80:20 akurasi yang didapat lebih baik dari metode pembanding yaitu 95%.

Daftar Notasi

Contoh penulisan notasi dapat diuraikan dengan keterangan sebagai berikut:

$f_d(i)$ = frekuensi kemunculan *term i* pada dokumen *j*
N $f_d(j)$ = total *term* pada dokumen *j*
N = jumlah total dokumen
$df(t)$ = jumlah dokumen yang mengandung *term t*
x = nilai *term* dari perhitungan *term weighting*
w = bobot model yang dipelajari

Referensi

[1] B. Indiarto, “Klasifikasi Sms Spam Dengan Metode Naive Bayes Classifier Untuk Menyaring Pesan Melalui Selular,” vol. 8, no. 2, pp. 167–172, 2016.
[2] A. Setiawan, I. F. Astuti, and A. H. Kridalaksana, “Klasifikasi Dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (NBC) (Studi Kasus : Perpustakaan Daerah Provinsi Kalimantan Timur),” vol. 10, no. 1, 2015.
[3] I. N. Dewi, C. Supriyanto, F. I. Komputer, and U. D. Nuswantoro, “Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naïve Bayes,” vol. 2013, no. November, pp. 156–160, 2013.
[4] S. Basuki, S. Maghfiroh, and Y. Azhar, “Klasifikasi Tweets Tindak Kejahatan Berbahasa Indonesia Menggunakan Naive Bayes,” vol. 2, no. 7, pp. 933–944, 2020.
[5] G. S. Sravya and G. Pradeepini, “Mobile Sms Spam Filter Techniques Using Machine Learning Techniques,” vol. 9, no. 03, pp. 384–389, 2020.
[6] I. T. Utami, “Analisis Regresi Logistik Biner Untuk Mengklasifikasi Penderita Hipertensi Berdasarkan Kebiasaan Merokok Di RSU Mokopido Toli-Toli Binary Logistic Regression Analysis For Classification Of Hypertension Patients Based On Smoking Habits In Mokopido Toil-Toli Hospital,” vol. 7, no. 3, pp. 341–348, 2018.
[7] D. B. Setyohadi and F. A. Kristiawan, “Perbaikan performansi klasifikasi dengan preprocessing iterative partitioning filter algorithm 1,” vol. 14, no. 01, pp. 12–20, 2017.
[8] F. Ratnawati, “Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter,” 2018.
[9] A. H. Yunas and M. Fikry, “Klasifikasi Tweet E-Commerce dengan Menggunakan Metode Support Vector Machine,” vol. 4, no. 2, pp. 50–55, 2018.
[10] R. Melita *et al.*, “(TF-IDF) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus : Syarah Umdatil Ahkam),” vol. 11, no. 2, 2018.
[11] K. S. Setyawati *et al.*, “Aplikasi Sentiment Analysis Terhadap Pelaksanaan Pembelajaran Jarak Jauh Universitas Kristen Petra Dengan Metode Naive Bayes Classifier.”
[12] A. Y. Permana, “Implementasi Stemming Porter Kbbi Untuk Klasifikasi Topik Soal Ujian Nasional Bahasa Indonesia Menggunakan Algoritma Naive Bayes,” 2017.
[13] 2016 Kurniawati, “Term weighting berbasis indeks kelas menggunakan metode tf.idf.ics,” 2016.
[14] T. Trisna *et al.*, “Analysis And Detection Of Hoax Contents In Indonesian News Based On Machine Learning,” vol. 4, no. 1, 2019.
[15] C. Science, A. Bimantara, and T. A. Dina, “Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression,” vol. 4, no. 1, pp. 978–979, 2018.
[16] O. S. Balogun, T. J. Akingbade, and A. Akinrefon, “Evaluation Of Logistic Regression In Classification Of Drug Data In Kwara State,” no. May 2014, 2013.
[17] H. Saiyar, “Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes,” vol. 5, no. 5, pp. 498–502, 2018.
[18] M. Athaillah, Y. Azhar, and Y. Munarko, “Perbandingan Metode Klasifikasi Berita Hoaks Berbahasa Indonesia Berbasis Pembelajaran Mesin,” vol. 2, no. 5, pp. 675–682, 2020.
[19] D. H. Kalokasari, I. M. Shofi, and A. H. Setyaningrum, “Implementasi Algoritma Multinomial Naive Bayes Classifier Pada Sistem Klasifikasi Surat Keluar (Studi Kasus : DISKOMINFO Kabupaten Tangerang),” vol. 10, no. 2, 2017.