ISSN: 2503-054X
Vol. 8 No: 1, April 2023

Klasifikasi Penyalahgunaan Pesan singkat Menggunakan Algoritma *Naïve Bayes*

Elva Amaia
Universitas Negeri Makassar
Makassar, Indonesia
elvaamaliaaa@gmail.com

Andi Nurul Izzah
Universitas Negeri Makassar
Makassar, Indonesia
a.nurulizzah281@gmail.com

Andi Akram Nur Risal
Universitas Negeri Makassar
Makassar, Indonesia
akramandi@unm.ac.id

***Abstract*—Pesan singkat atau yang biasa disebut dengan SMS (*Short Message Service*) adalah pesan elektronik dengan memanfaatkan teknologi mengirim dan menerima pesan pada sebuah *device* atau *smartphone*. Saat ini penyeberan penerimaan pesan singkat sulit dikendalikan, dari nomor yang tidak dikenal. Pesan singkat dapat digolongkan menjadi beberapa kelas yaitu, pesan singkat normal, pesan singkat promo, dan pesan singkat penipuan. Karena banyaknya pesan singkat yang masuk, maka penilitian ini melakukan klasifikasi Penyalahgunaan pesan singkat menggunakan algoritma naïve bayes dengan PySpark. Tujuan dari penelitian ini adalah untuk membedakan atau mengklasifikasikan pesan singkat normal, promo, dan penipuan. Dataset pada penelitian ini menggunakan data berbahasa indonesia dengan jumlah 1143 data. Dari hasil pengujian berdasarkan metode yang diusulkan yaitu Algoritma naïve bayes mendapatkan nilai akurasi *precision* 94%, *recall* 92%, *f1-score* 93% dan *accuracy* sebesar 94%.**

Kata kunci — Klasifikasi, Naïve Bayes, Penyalahgunaan Pesan Singkat, PySpark.

I. PENDAHULUAN

Pesan singkat atau yang biasa disebut dengan SMS (Short Message Service) adalah pesan elektronik dengan memanfaatkan teknologi mengirim dan menerima pesan pada sebuah device atau smartphone. Pesan singkat juga salah satu media komunukasi yang masih banyak digunakan di kalangan masyarakat saat ini. Penggunaan pesan singkat saat ini dapat disalahgunakan oleh oknum untuk mencari dan memanipulasi korban demi sebuah keuntungan dan merugikan bagi korban atau masyarakat.

Pesan singkat merupakan salah satu sistem komunikasi paling populer di dunia. Di Indonesia, jumlah pengguna pesan singkat masih cukup tinggi. Pesan singkat memungkinkan pengguna untuk berkomunikasi dengan orang yang mereka kenal serta mengirim pesan kepada orang yang tidak mereka kenal, Seperti menawarkan produk, promosi, layanan dan sebagainya. Pesan singkat juga banyak digunakan dalam kasus penipuan. Sehingga, pesan singkat dapat digolongkan menjadi beberapa kelas yaitu pesan singkat normal pesan singkat promo, dan pesan singkat penipuan.

Penyalahgunaan pesan singkat menjadi penyebab utama dari ketidaknyamanan pengguna. Hal ini dikarekan pengguna harus meluangkan waktu untuk melihat dan menghapus Penyalahgunaan pesan singkat yang menghabiskan ruang memori ponsel. Berdasarkan permasalahan ini, penelitian ini menggunakan Teknik data mining dalam menganalisis pola pesan singkat dengan mengklasifikasikan pesan singkat menjadi 3 kategori pesan singkat normal, penipuan, dan promosi.

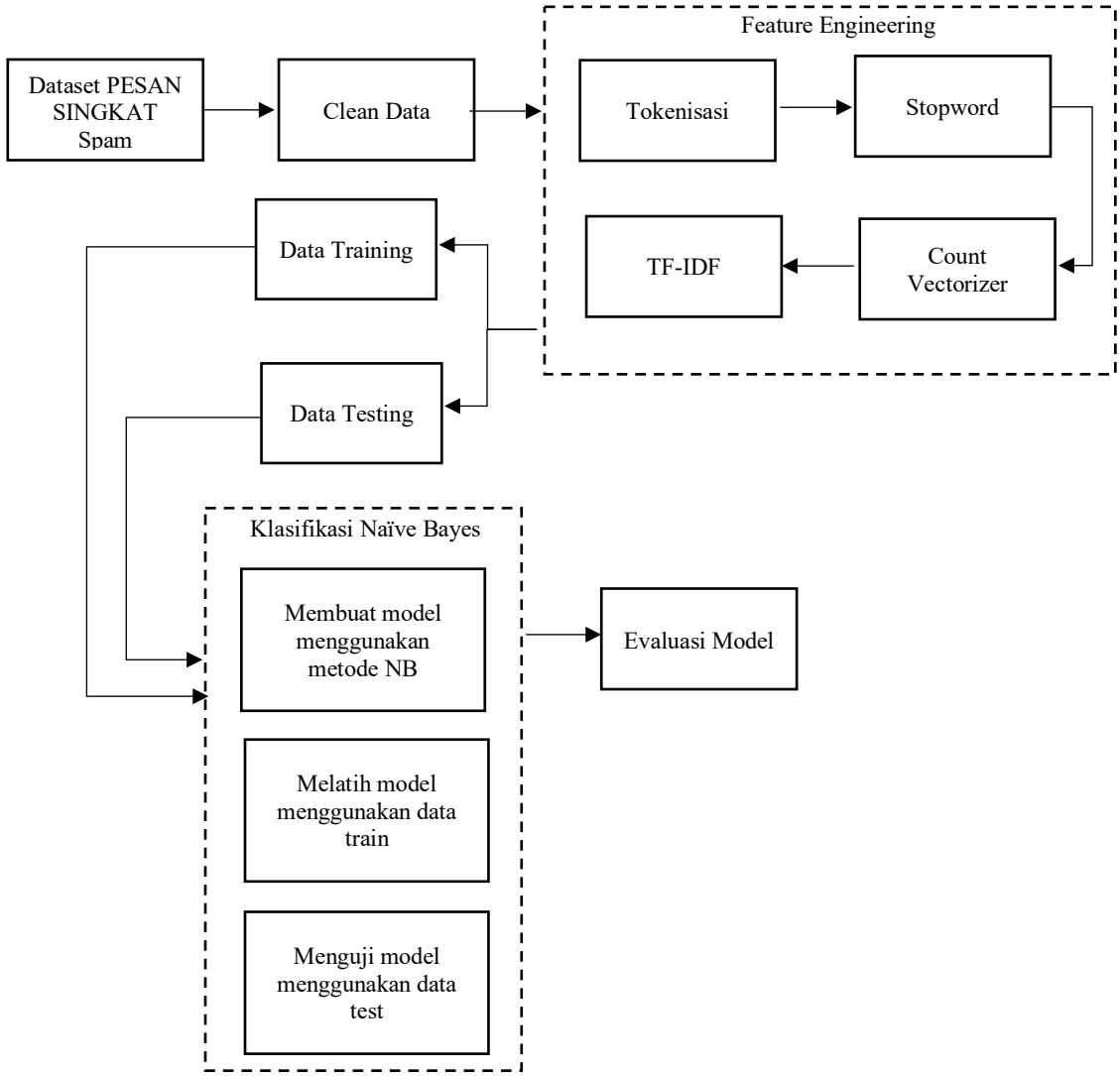
Beberapa penelitian sebelumnya mengenai klasifikasi dan Penyalahgunaan pesan singkat seperti, Devinta Nurul, dkk, mengklasifikasikan penyalahgunaan pesan singkat dengan menggunakan algoritma naïve bayes, SVM dan decision tree dan membandikan hasil algoritma tersebut terkait penyalahgunaan pesan singkat [1]. Hasil dari analisa pengujian pada perbandingan algoritma yang lebih baik dalam mengklasifikasi penyalahgunaan pesan singkat adalah algoritma naïve bayes, dikarenakan nilai recall dan f1-score tertinggi serta nilai accuracy mencapai 0.94. Penelitian lainnya yang dilakukan oleh Meriohengki memprediksi penyalagunaan email berbahasa inggris menggunakan algorima Naïve Bayes dan SVM berbasis PSO [2]. Penelitian tersebut bertujuan mencari nilai akurasi tertinggi. Perbandingan dua algoritma tersebut memiliki hasil yang cukup baik, akan tetapi algoritma SVM berbasis PSO mendapat nilai akurasi paling unggul. Sedangkan, Ferin Reviantika, dkk, melakukan analisis Penyalahgunaan pesan singkat menggunakan metode *logistic regressic* [3]. Pembagian data train dan test menjadi 80:20 menghasilkan nilai akurasi yang lebih baik yaitu 95%. Selanjutna, penelitian Agus Setiyono dengan judul klasifikasi Penyalahgunaan pesan singkat dengan Support Vector Machine (SVM) [4]. Penelitian Setiyono dan Pardede adalah membandingkan hasil algoritma SVM, Multinomial Naïve Bayes dan Decision Tree. Dataset diproses dengan 2 bagian sebagai data training dan data testing, disertai dengan set role yang bertujuan untuk klasifikasi data tersebut, kemudian untuk meningkatkan nilai akurasinya digunakan metode TF-IDF. Sehingga menghasilkan algoritma SVM sebagai algoritma terakurat dengan nilai akurasi sebesar 98.33%.

Berdasarkan penelitian-penelitian di atas, metode untuk klasifikasi yang mendapatkan hasil akurasi yang tinggi dipengaruhi oleh beberapa hal. Maka dari itu, Penelitian ini menerapkan klasifikasi konsep algoritma naïve bayes dengan menggunakan Pyspark yang merupakan platform untuk memproses data dalam skala besar.

II. METODE

Tahapan-tahapan penelitian ini dijelaskan pada Gambar 1 berikut.

ISSN: 2503-054X
Vol. 8 No: 1, April 2023



Gambar 1 Alur Penelitian

Penggunaan dataset penelitian ini adalah dataset penyalahgunaan pesan singkat sebanyak 1143 pesan singkat berbahasa Indonesia yang tersimpan dalam file berbentuk .csv (comma separated values). Berikut merupakan sampel dari dataset yang digunakan yang disajikan pada Tabel 1.

Tabel 1 Dataset Penyalagunaan pesan singkat Berbahasa Indonesia

teks	label
[PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K	2
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. S&K	2
2016-07-08 11:47:11.Plg Yth, sisa kuota Flash Anda 478KB. Download MyTelkomsel apps di http://tsel.me/tsel utk cek kuota&beli paket Flash atau hub *363#	2

**Sumber: github.com/ksnugroho*

A. *Clean Data*

Pada tahap *clean data*, untuk memperoleh hasil klasfikasi yang optimal, maka perlu dilakukan untuk optimasi terhadap dataset yang ada. Pada tahapan ini, akan dilakukan pembersihan data pada dataset. Pembersihan dilakukan untuk menghilangkan *noise* pada data. Adapun *noise* yang dimaksud diantaranya adalah symbol, tautan, angka, dan sebagaianya. Hal ini bertujuan untuk membersihkan data yang tidak diperlukan agar data yang dihasilkan memiliki standarisasi dengan baik dan juga untuk meningkatkan kualitas data. Tabel 2 menunjukkan hasil *cleaning data*.

Tabel 2 Hasil Tahap *Clean Data*

teks	clean text

ISSN: 2503-054X
Vol. 8 No: 1, April 2023

[PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K	promo beli paket flash my telkomsel app dpt extra kuota g lte extra nelpon mnthr buru cek tselmemytsel sk
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. S&K	gb hari hanya rp ribu spesial buat anda yang terpilih aktifkan sekarang juga di promo sd nov buruan aktifkan sekarang sk
2016-07-08 11:47:11.Plg Yth, sisa kuota Flash Anda 478KB. Download MyTelkomsel apps di http://tsel.me/tsel utk cek kuota&beli paket Flash atau hub *363#	plg yth sisa kuota flash anda kb download mytelkomsel apps di utk cek kuotabeli paket flash atau hub

Berdasarkan Tabel 2, upaya *preprocessing* yang dilakukan penelitian ini berupa *case folding*. Kolom *clean text* merupakan hasil dari tahapan *case folding* yang dilakukan pada dataset pesan singkat. Dimana tahapan *case folding* terdiri dari beberapa perintah seperti teks akan diubah menjadi *lower case*, menghapus tautan, angka, serta karakter tanda baca atau symbol. Tahapan ini bertujuan untuk memudahkan proses selanjutnya dengan menyamaratakan format dataset serta meminimalisir data *noise* atau data kosong pada data input.

B. *Feature Engineering*

Membangun fitur merupakan tahap yang sangat penting dalam membuat model, dimana tahap ini menentukan sukses tidaknya suatu model. Pada tahap ini akan mengekstrak dan mengubah fitur. Tahapan pertama adalah tokenisasi atau tokenizing yang merupakan tahapan lanjutan dari tahapan *clean data*. Pada proses tokenisasi akan memisahkan teks menjadi potongan-potongan berupa token. Tabel 3 memperlihatkan hasil tokenisasi.

Tabel 3 Hasil Tahap Tokenisasi

teks	**clean text**	**fitur**
[PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K	promo beli paket flash my telkomsel app dpt extra kuota g lte extra nelpon mnthr buru cek tselmemytsel sk	[promo, beli, paket, flash, my, telkomsel, app, dpt, extra, kuota, g, lte, extra, nelpon, mnthr, buru, cek, tselmemytsel, sk]
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. S&K	gb hari hanya rp ribu spesial buat anda yang terpilih aktifkan sekarang juga di promo sd nov buruan aktifkan sekarang sk	[gb, hari, hanya, rp, ribu, special, buat, anda, yang, terpilih, aktifkan, sekarang, juga, di, promo, sd, nov, buruan, aktifkan, sekarang, sk]
2016-07-08 11:47:11.Plg Yth, sisa kuota Flash Anda 478KB. Download MyTelkomsel apps di http://tsel.me/tsel utk cek kuota&beli paket Flash atau hub *363#	plg yth sisa kuota flash anda kb download mytelkomsel apps di utk cek kuotabeli paket flash atau hub	[plg, yth, sisa, kuota, flash, anda, kb, download, mytelkomsel, apps, di utk, cek, kuotabeli, paket flash, atau, hub]

Pada tabel diatas, dapat dilihat bahwa kolom fitur merupakan representasi dari tahap tokenisasi dari kolom *clean text*. Dimana kolom *clean text* berisikan data yang sudah dilakukan *pre-processing*. Tokenisasi menghasilkan potongan kata dari kalimat pada dataset pesan singkat. Tiap kata disebut sebagai token atau fitur yang kemudian akan dianalisa pada tahapan selanjutnya. *Tokenizing* akan mendukung tahap analisa teks atau pembobotan.

C. *Count Vectorizer*

Setelah data melewati tahap tokenisasi, kumpulan dokumen teks akan dikonvesi menjadi vector jumlah token. countVectorizer dapat digunakan untuk mengekstrak fitur dan menghasilkan CountVectorizerModel. Tahap ini akan menghasilkan model yang merepresentasikan vector dari token atau fitur. Hasil tahap count Vectorizer pada Tabel 4.

Tabel 2 Hasil Tahap *CountVectorizer*

teks	**features**
[PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K	(6420,[0,7,12,15,42,51,88,93,112,113,149,177,178,194,207,234,286,305,330,435,637,1060,5036],[2.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,2.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])
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. S&K	(6420,[0,1,16,26,28,48,62,71,87,125,148,170,194,237,245,394,998,1431,1548,1872,1883,2255,3421,5712],[1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,2.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])
2016-07-08 11:47:11.Plg Yth, sisa kuota Flash Anda 478KB. Download MyTelkomsel apps di http://tsel.me/tsel utk cek kuota&beli paket Flash atau hub *363#	(6420,[0,1,12,15,22,30,38,42,88,141,281,340,430,830,922,1531,1908,1988,2553,3064],[1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,2.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])

Pola teks dapat terbentuk pada tahapan *count vectorizer*. Proses ini bekerja menjadikan teks menjadi vektor. Pada tahapan ini terdapat juga proses *spark pipeline*.

ISSN: 2503-054X
Vol. 8 No: 1, April 2023

D. TF-IDF

Fitur asli akan melalui proses pemetaan fungsional pada tahap *feature extraction* yang akan menghasilkan fitur baru. Penelitian ini memanfaatkan *Term Frequency Inverse Document Frequency* (TF-IDF) untuk mengubah teks menjadi *vector* di tahap ektraksi fitur. *Vector* antar kata tersebut akan digunakan oleh metode TF-IDF untuk mencari nilai yang akan merepresentasikan tiap dokumen (kata) dari kumpulan data *training.* Sedangkan *cluster centroid* merupakan *prototype vector* yang menentukan kesetaraan antar dokumen dalam suatu *cluster* [5]. TF-IDF bekerja dengan mengkalkulasi bobot antar *vector* menggunakan langkah integrasi *term frequency* (tf) dan *inverse document frequency* (idf) [6]. Tahapan pada TF-IDF diantaranya yaitu mendapatkan jumlah munculnya kata yang diketahui (tf) kemudian dikalikan dengan banyak pesan dimana tiap fitur tesebut muncul pada dokumen (idf). Persamaan penentuan pembobotan TF-ID dibawah ini:

$$tf_{ij} = \frac{f_d(i)}{\max f_d(j)} \quad (1)$$

Banyaknya jumlah kata atau *term i* pada sebuah dokumen *j* disebut *Term Frequency*. Sedangkan *Inverse Document Frequency* merupakan frekuensi munculnya *term* pada semua dokumen yang ada. IDF berperan untuk mengurangi bobot dari suatu *term* jika kemunculannya banyak tersebar di seluruh dokumen yang kemudian dituliskan dalam bentuk persamaan berikut.

$$idf\ (t, D) = log \frac{N}{|\{d \in D\ :\ t\ \in\ d\}|} \quad (2)$$

Dimana *N* sebagai jumlah total dokumen dalam *corpus*, $N\ =\ |D|$. Kemudian,$|\{d \in D : t \in d\}| = df(t)$, adalah jumlah dokumen yang mengandung *term t*. nilai IDF yang diperoleh berbanding terbali dengan jumlah dokumen yang terdapat *term* tertentu didalamnya. Ketika suatu *term* memiliki nilai frekuensi yang rendah pada seluruh dokumen maka nilai IDF-nya akan lebih besar. Sebaliknya, nilai IDF pada *term* yang sering muncul akan lebih rendah.

E. *Naïve Bayes*

Pada *machine learning*, *naïve bayes* ialah metode klasifikasi yang bersumber dari teorema Bayes. Metode klasifikasi ini menggunakan teknik statistik dan probabilitas, yaitu peramalan peluang berdasarkan data historis. Dijelaskan bahwa dengan mempertimbangkan vektor informasi objek, *naïve bayes* akan menentukan probabilitas untuk setiap kelas keputusan dengan asumsi bahwa kelas keputusan benar. Frekuensi tabel keputusan digunakan untuk menentukan probabilitas yang digunakan untuk membuat prediksi akhir. Persamaan 3 merupakan teorema Bayes.

$$P(y|X)\ =\ \frac{P(X \backslash y)P(y)}{P(X)} \quad (3)$$

Variable y merupakan *class* variabel yang merepresentasikan pesan inputan, diklasifikasikan menjadi pesan singkat normal, *fraud*, atau promo. Sedangkan *X* merepresentasikan fitur-fitur dari pesan tersebut. Klasifikasi menggunakan *naïve bayes* dapat disebut dengan *multinomial naïve bayes* yang mana merupakan model penyederhanaan dari algoritma Bayes yang digunakan dalam mengklasifikasi teks. Dalam menentukan kelas dokumen pada rumus *Multinomial Naïve Bayes* yaitu berdasarkan frekuensi dan kata yang muncul didalam dokumen tersebut [6].

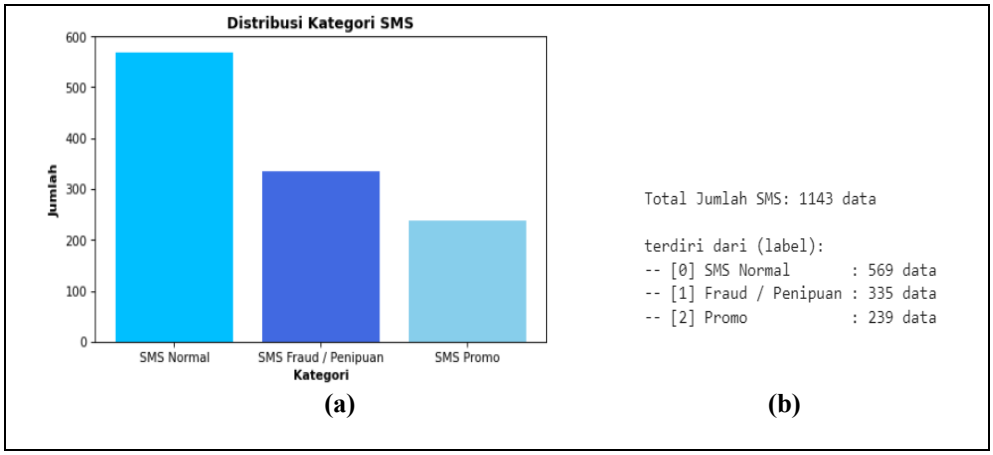
F. *PySpark*

Apache Spark digunakan dalam memproses data besar dalam memori. Fitur pengembangan API pada *PySpark* yang bersifat fleksibel memudahkan seorang analis data untuk mengolah data secara berulang dengan cepat yang akan diproses (*machine learning, streaming, sql* dsb). *PySpark* dapat membantu dalam menggunakan *Resilient Distributed Datasets* (RDD) dalam Bahasa pemrograman *Apache Spark* dan Python dengan memanfaatkan library Py4j. *Library* Py4j merupakan *library* yang telah terinterigasi dalam *PySpark*.

ISSN: 2503-054X
Vol. 8 No: 1, April 2023

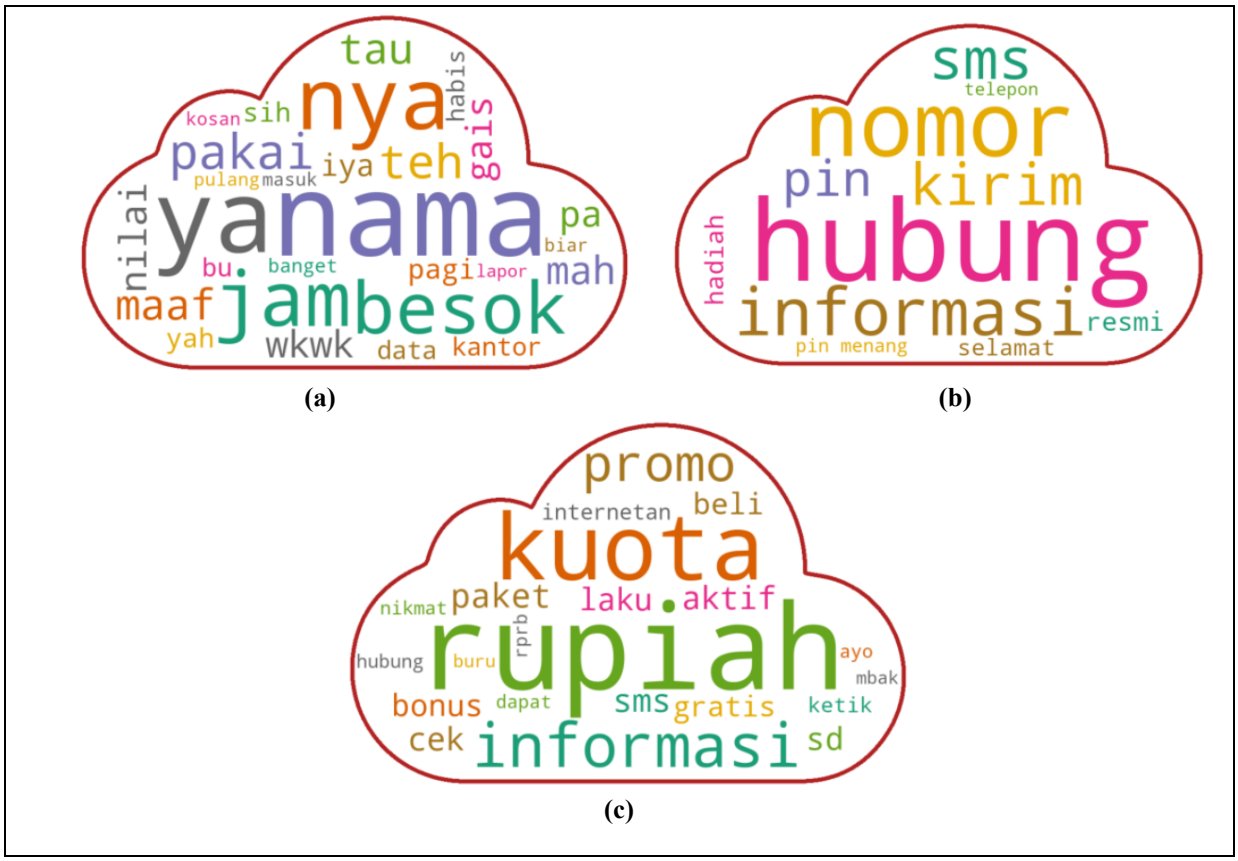
III. HASIL DAN PEMBAHASAN

Sumber data penelitian ini adalah *github.com* milik Ksnugroho yaitu dataset penyalahgunaan pesan singkat berbahasa Indonesia. Semua tahapan yang dilakukan menggunakan tool *PySpark* di *Google Colaboratory* dengan menerapkan algoritma Naïve Bayes. Dataset Penyalahgunaan pesan singkat berbahasa Indonesia ini memiliki 3 *class* yang terdiri dari 569 data berlabel 0 (pesan singkat Normal), 335 data berlabel 1 (pesan singkat *fraud* atau penipuan) dan 239 data berlabel 2 (pesan singkat promo). Visualisasi data tersebut dapat dilihat pada Gambar 2 (a) dan (b) berikut.



Gambar 2 Distribusi Pesan Singkat

Untuk mengetahui kata atau fitur yang sering muncul pada tiap label, dapat dilihat pada Gambar 3 sebagai visualisasi fitur. Pada Tabel 5 juga menunjukkan fitur lainnya dari setiap label yang memberikan informasi bahwa fitur-fitur yang menjadi ciri pada masing-masing label. Fitur yang memiliki ukuran paling besar merupakan fitur yang paling sering muncul pada pesan singkat *class* tersebut.



Gambar 3 (a) SMS Normal; (b) SMS *Fraud*; (c) SMS Promo

ISSN: 2503-054X
Vol. 8 No: 1, April 2023

Tabel 5 Fitur yang sering muncul tiap *class*

SMS Normal	SMS Penipuan (*Fraud*)	SMS Promo
Nama	Kirim	Rupiah
Bri	Hubung	Kuota
Nomor	Spasi	Lihat
Transfer	Ketik	Lokasi
Ya	Informasi	Tarif
Atay	Akun	Informasi
Jam	Hadiah	Hubung
Besok	Yuk	Promo
Pakai	Selamat	Rprb
Pagi	Format	Flash
Kirim	Admin	Internet
Mbak	Tukar	Yuks
Jins	Poin	Cek
Warna	Cek	Telepon
maaf	Kunjung	Khusus
Teh	Uinfo	Gratis
Gais	Utama	Tselmefl
Wkwk	Periode	Aktif
Nilai	Care	Update
Tau	Rejeki	Pakai
…	…	…
kfc	masalah	beli

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	169
1.0	0.95	0.93	0.94	98
2.0	0.97	0.86	0.91	84
accuracy			0.94	351
macro avg	0.95	0.93	0.94	351
weighted avg	0.94	0.94	0.94	351

Gambar 4 *Classification report*

Pada Gambar 4, menunjukkan hasil akhir klasifikasi penyalahgunaan pesan singkat berbahasa Indonesia menggunakan algoritma Naïve Bayes. Label 0.0 mewakilkan pesan singkat Normal, 1.0 mewakilkan pesan singkat penipuan dan 2.0 mewakilkan pesan singkat promo. Precision menjelaskan kesesuaian dokumen asli dan prediksi dari hasil klasifikasi. Gambar 3 menunjukkan klasifikasi pesan singkat promo memiliki nilai precision tertinggi sebesar 0.97 dan nilai precision terendah yaitu untuk klasifikasi pesan singkat normal sebesar 0.93.

Recall merepresentasikan proporsi data yang diprediksi dengan akurat berdasarkan jumlah total data secara keseluruhan di suatu kelas. Nilai *recall* tertinggi didapatkan oleh data berlabel pesan singkat normal sebesar 0.99, nilai yang mendekati sempurna. Sedangkan nilai *recall* terendah dimiliki data berlabel 2.0 atau pesan singkat promo sebesar 0.86. Kemudian untuk nilai f1-score menunjukkan perbandingan nilai rata-rata antar precision dan recall. Untuk class 0 atau pesan singkat normal mendapatkan nilai f1-score 0.96. Akurasi merupakan besaran dari perbandingan kasus yang mengidentifikasi benar dari jumlah class. Berdasarkan hasil classification report di atas hasil klasifikasi Penyalahgunaan pesan singkat berbahasa Indonesia menghasilkan nilai accuracy sebesar 94%.

IV. KESIMPULAN

Pengguna layanan pesan singkat pada *smartphone* (SMS) mengalami gangguan akibat pesan singkat yang berisikan penipuan dan atau promosi produk. Upaya untuk mengatasi permasalahan tersebut telah berhasil dilakukan dengan memanfaatkan algoritma Naïve Bayes. Berdasarkan hasil penelitian klasifikasi Penyalahgunaan pesan singkat berbahasa Indonesia dengan algoritme Naïve Bayes menggunakan PySpark, didapatkan hasil yang sangat baik dan tingkat akurasi yang dapat dikatakan tinggi. Penelitian data ini dibagi menjadi 2 bagian, data train dan data test perbandingan 70 : 30 mendapatkan nilai precision 94%, recall 92%, f1-score 93% dan

ISSN: 2503-054X
Vol. 8 No: 1, April 2023

accuracy sebesar 94%. Model yang dihasilkan algoritma Naïve Bayes berhasil membedakan 3 kategori pesan singkat antara lain pesan normal, pesan penipuan (*fraud*) dan pesan promo. Beberapa fitur yang mencirikan pesan singkat spam seperti 'kirim', 'hubung', 'spasi', 'ketik', 'rupiah', 'kuota', 'lihat', lokasi', 'tarif' dan 'informasi'. Kata-kata tersebut memiliki frekuensi paling tinggi pada pesan singkat kategori *fraud* dan promo.

Pengembangan lebih lanjut dari penelitian ini diharapkan dapat membandingkan beberapa algortima lainnya untuk melihat tingkat akurasi dalam memprediksi pesan singkat spam dan menggunakan algoritma asosiasi untuk melihat pola yang terbentuk dari pesan singkat spam.

DAFTAR PUSTAKA

[1] N. F. Devinta, A. Niken and Y. Ahmad, "Perbandingan Algoritma Naive Bayes, SVM dan Decision Tree untuk Klasifikasi Penyalagunaan Pesan Singkat," December 2020. [Online]. Available: https://doi.org/10.32767/jusim.v5i02.956.

[2] M. Mochamad, "Klasifikasi Algoritma Naïve Bayes dan SVM Berbasis PSO Dalam Memprediksi Spam Email Pada Hotline-Sapto," *Paradigma,* vol. 22, no. 1, 2020.

[3] R. Ferin, "Analisis Klasifikasi Penyalahgunaan pesan singkat Menggunakan Logistic Regression," vol. 04, no. 03, pp. 155-160, 2021.

[4] S. A and P. H, "KLASIFIKASI PENYALAHGUNAAN PESAN SINGKAT MENGGUNAKAN SUPPORT VECTOR MACHINE," *Jurnal Pilar Nusa Mandiri,* vol. 15, no. 2, pp. 275-280, 2019.

[5] P. M. Prihatini, I. K. G. Darma Putra, I. A. Dwi Giriantari and M. Sudarma, "Fuzzy-Gibbs Latent Dirichlet Allocation Model for Feature Extraction on Indonesian Documents," *Contemporary Engineering Sciences (CES),* vol. 10, no. 9, pp. 403-421, 2017.

[6] K. N.W. and dkk., "Seleksi Fitur Bobot Kata dengan Metode TF-IDF untuk Ringkasan Bahasa Indonesia," *Merpati,* vol. 6, no. 2, 2018.

[7] M. T, Y. E, B. B and F. M.A, "Penentuan Rating Review Film menggunakan Metode Multinomial Naïve Bayes Classifier dengan Feature Selection berbasis Chi-Square dan Galavotti-Sebastiani-Simi Coefficient," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya,* vol. 3, no. 1, pp. 447-453, 2019.

[8] E. Dragut, F. Fang, P. Sistla, C. Yu and W. Meng, "Stop Word and Related Problems in Web Interface Integration," *VLDB Endowment,* 2009.

[9] L. Mohan, J. Pant, P. Suyal and A. Kumar, "Support Vector Machine Accuracy Improvement with Classification," *12th Int. Conf. Comput. Intell. Commun. Networks, CICN,* p. 477–481, 2020.

[10] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Method," *Cambridge University Press,* 2000.

[11] L. GuangJun, S. Nazir, H. U. Khan and A. U. Haq, "Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms," *Security and Communication Networks,* vol. 2020, 2020.

[12] F. D. Pramakrisna, F. D. Adhinata and N. A. F. Tanjung, "Aplikasi Klasifikasi SMS Berbasis Web menggunakan Algoritma Logistic Regression," *TEKNIKA,* vol. 11, no. 2, pp. 90-97, 2022.

[13] m. J. Awan, R. A. Khan, H. Nobanee, A. Yasin, S. M. Anwar, U. Naseem and V. P. Singh, "A Recommendation Engine for Predictiong Movie Rating Using A Big Data Approach," *Electronics,* vol. 10, no. 1215, 2021.

[14] M. Chaudhury, A. Karami and M. A. Ghazanfar, "Large-Scale Music Genre Alanysis and Classification Using Machine Learning with Apache Spark," *Electronics,* vol. 11, no. 2567, 2022.

[15] F. Zamachsari, G. V. Saragih, Susafa'ati and W. Gata, "Analisis Sentimen Pemindahan Ibu Kota Negara dengan Feature Selection Algoritma Naive Bayes dan Support Vector Machine," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 4, no. 3, pp. 504-512, 2020.

[16] B. Guanwan, H. S. Pratiwi and e. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk menggunakan Metode Naive Bayes," *JEPIN (Jurnal Edukasi dan Penelitian Informatika),* vol. 4, no. 2, pp. 113-118, 2018.

[17] S. M. P. Tyas, B. S. Rintyarna and W. Suharso, "The Impact of Feature Extraction to Naïve Bayes Based Sentiment Analysis on Review Dataset of Indihome Services," *Jurnal Teknologi Informasi dan Komunikasi Digital Zone,* vol. 13, no. 1, pp. 1-10, 2022.

[18] L. Septiani and Y. Sibaroni, "Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia," *Jurnal Linguistik Komputasional,* vol. 2, no. 2, pp. 62-67, 2019.

[19] A.-M. Copaceanu, "Sentiment Analysis Using Machine Learning Approach," *Ovidius Univercity Annals : Economic Sciences Series,* vol. 21, no. 1, pp. 261-270, 2021.

[20] G. Al-Rawashdeh, R. Mamat and N. H. B. A. Rahim, "Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection," *IEEE,* vol. 7, pp. 143721-143734, 2019.