JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

Klasifikasi SMS *Spam* Berbahasa Indonesia Menggunakan Algoritma Multinomial Naïve Bayes

Herwanto[*], Nuke L Chusna, Muhammad Syamsul Arif

Fakultas Teknik, Program Studi Teknik Informatika, Universitas Krisnadwipayana, Jakarta, Indonesia
Email:[1,*]herwanto@unkris.ac.id, [2] nuke_anshori@yahoo.com, [3]syamsularif0904@gmail.com
Email Penulis Korespondensi: herwanto@unkris.ac.id

Abstrak−Berdasarkan laporan yang disampaikan oleh *Truecaller Insights Report* 2020 Indonesia menempatkan posisi ke-6 paling banyak mendapat pesan *spam*, salah satu penerapan *spam* ada pada SMS. SMS *spam* mengandung pesan yang tidak diinginkan atau tidak diminta, termasuk iklan, penipuan dan lain sebagainya. Adanya pesan *spam* ini menimbulkan ketidaknyamanan dari sisi pengguna saat menerima SMS *spam* bahkan ada yang menjadi korban tindak kejahatan setelah merespon SMS tersebut. Untuk meminimalisir ketidaknyamanan dan tindak kejahatan yang disebabkan oleh pesan *spam*, maka tujuan penelitian ini adalah perlu dilakukan penyaringan SMS *spam* atau SMS *filtering* dengan cara mengklasifikasikan SMS *spam* menggunakan algoritma *Multinomial Naïve Bayes* dengan mencari kombinasi parameter terbaik sehingga dapat meningkatkan performa model yang di bentuk. Hasil pengujian model mendapatkan nilai *precision* tertinggi pada model MNB dan SVM sebesar 93%, nilai *recall* tertinggi pada model SVM sebesar 94%, nilai *f1-score* tertinggi pada model SVM sebesar 94% dan nilai *accuracy* tertinggi pada model SVM sebesar 95%. Serta waktu pengujian tercepat pada model MNB sebesar 2.66 ms.

Kata Kunci: Klasifikasi SMS *Spam*; *Multinomial Naïve Bayes*; SMS *Spam* Berbahasa Indonesia; *Text Mining*; SMS

Abstract−Based on a report submitted by Truecaller Insights Report 2020, Indonesia placed sixth position with the most spam messages, one of the spam applications is SMS. Spam SMS contains unwanted or unsolicited messages, including advertisements, scams and so on. The existence of this spam message causes inconvenience from the user's side when receiving spam SMS, and some even become victims of crime after responding to the SMS. To minimize inconvenience and crime caused by spam messages, the purpose of this study is to filter SMS spam or SMS filtering by classifying SMS spam using the Multinomial Naïve Bayes algorithm by looking for the best combination of parameters to improve the performance of the model that is formed. The results of model testing get the highest precision value in the MNB and SVM models by 93%, the highest recall value in the SVM model at 94%, the highest f1-score value in the SVM model at 94%, the highest accuracy value in the SVM model at 95%, and the fastest test time on the MNB model is 2.66 ms.

Keywords: Classification of SMS Spam; Multinomial Nave Bayes; Indonesian SMS Spam; Text Mining; SMS

1. PENDAHULUAN

Teknologi komunikasi dan informasi semakin berkembang dan menghasilkan berbagai macam media komunikasi serta informasi yang dapat diolah, sehingga informasi yang terkumpul merupakan aset yang dapat dimanfaatkan untuk dianalisis dan menghasilkan pengetahuan atau informasi berharga untuk masa yang akan datang. Berdasarkan penelitian yang dilakukan oleh digital marketing Emarketer memperkirakan pada tahun 2018 adanya peningkatan jumlah pengguna aktif *smartphone* di Indonesia mencapai lebih dari 100 juta orang. Sehingga menempatkan Indonesia menjadi negara dengan peringkat keempat terbesar di dunia dalam pengguna aktif *smartphone* setelah Cina, India, dan Amerika [1]. Meningkatnya penggunaan *smartphone* di Indonesia dapat menimbulkan ancaman privasi data setiap penggunanya. Salah satu faktor yang menyebabkan terjadinya pelanggaran informasi dan privasi adalah karena para pengguna *smartphone* memiliki *security awareness* atau kurangnya tingkat kesadaran keamanan dalam menggunakan *smartphone* yang baik dan aman [2]. Sehingga dengan adanya faktor tersebut dapat menimbulkan indikasi atau peluang terjadinya tindak kejahatan oleh pihak yang tidak bertanggung jawab, salah satunya melalui pesan *spam.*

Berdasarkan laporan yang disampaikan oleh *Truecaller Insights Report* 2020, Indonesia menjadi negara Asia dengan jumlah pesan *spam* paling banyak pada tahun 2020. *Truecaller* mencatat pesan *spam* di Indonesia turun 34% dari 27,9% pada tahun 2019 menjadi 18,3% dan turun peringkat ke urutan 6 dalam 20 negara paling banyak mendapat pesan *spam*. Layanan keuangan menjadi jenis *spam* paling besar di Indonesia dengan proporsi 52% selanjutnya asuransi 25%, *operator seluler* 11%, *scam* 9% dan tagihan hutang 3% [3]. *spamming* merupakan perbuatan dengan mengirimkan pesan elektronik yang tidak diinginkan atau diminta serta tanpa adanya persetujuan penerimanya, sehingga dapat melanggar privasi dan hukum dalam bentuk penyalahgunaan data pribadi tanpa persetujuan yang dapat mengakibatkan kerugian [4]. Pengiriman informasi yang terindikasi melakukan *spam* (*spammer*) dapat dilakukan secara sengaja dengan mengirimkan pesan *spam* untuk berbuat kejahatan atau melakukan kegiatan untuk mempromosikan suatu produk. Salah satu penerapan *spam* ada pada SMS (*Short Message Service)*.

SMS merupakan fasilitas untuk mengirim atau menerima pesan singkat berupa teks melalui perangkat telepon genggam. SMS *spam* mengandung pesan yang tidak diinginkan atau tidak diminta, termasuk iklan, penipuan dan lain sebagainya. Saat ini pemerintah telah memberlakukan kebijakan tentang registrasi kartu prabayar melalui peraturan menteri Kominfo Nomor 14 Tahun 2017 bertujuan untuk membuat pelanggan jasa telekomunikasi memiliki data yang valid, serta dapat menanggulangi modus kejahatan dan memberikan

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

perlindungan terhadap kepentingan pelanggan jasa telekomunikasi [5]. Akan tetapi kebijakan tersebut masih menimbulkan adanya indikasi SMS yang mengandung *spam*. Faktor lain yang dapat menyebabkan adanya pesan *spam* adalah pemberian izin atau *permission* kepada aplikasi berbahaya sehingga aplikasi tersebut dapat mengakses data *user*, dengan adanya pesan *spam* ini menimbulkan ketidaknyamanan dari sisi pengguna saat menerima SMS *spam* bahkan ada yang menjadi korban tindak kejahatan setelah merespon SMS tersebut. Untuk meminimalisir ketidaknyamanan dan tindak kejahatan yang disebabkan oleh pesan *spam*, maka perlu dilakukan penyaringan SMS *spam* atau SMS *filtering* dengan cara mengklasifikasikan SMS *spam* menggunakan suatu metode untuk memperoleh keakurasian hasil prediksi yang optimal. Klasifikasi merupakan suatu teknik untuk menilai objek data serta mengelompokkan objek berdasarkan atribut – atribut atau ciri objek ke dalam salah satu kategori yang telah didefinisikan. Klasifikasi melakukan pembelajaran model berdasarkan data latih yang telah di berikan label atau kelas target. Untuk mempermudah dalam pengklasifikasian data SMS diperlukan suatu sistem menggunakan metode *text mining* sebagai salah satu alternatif untuk menyelesaikannya. *Text mining* adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, dimana *text mining* merupakan variasi dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar [6], serta mengimplementasikan *hyperparameter* yang merupakan *variable* untuk mempengaruhi *output* dari model.

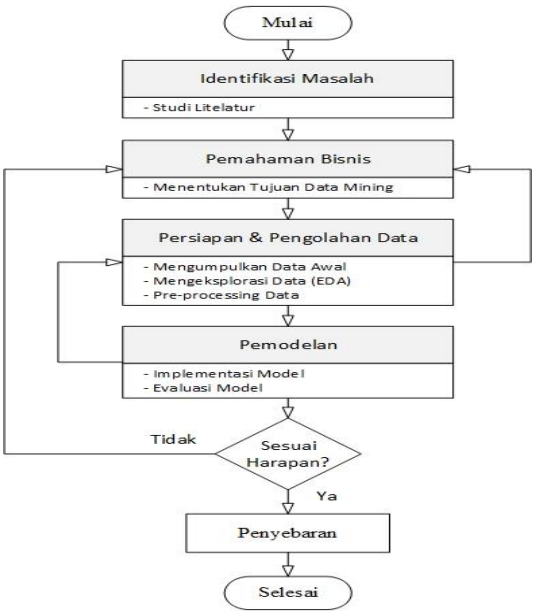
Beberapa penelitian terdahulu terkait dengan klasifikasi SMS *spam* telah dilakukan oleh Widyawati dan Sutanto penelitian tersebut membahas mengenai perbandingan algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) dalam klasifikasi SMS *spam* berbahasa Indonesia. Hasil penelitian tersebut dapat menghasilkan *accuracy* sebesar 93% untuk algoritma *Naïve Bayes* [1]. Algortima *Naïve Bayes* juga unggul dibandingkan SVM, Decision Tree [7], ID3 (*Iterative Dychotomizer Version 3*) dan TAN (*Tree Augmented Naïve*) [8] pada klasifikasi SMS *spam*. Penelitian – penelitan yang telah dilakukan menunjukan bawah algoritma *Naïve Bayes* memiliki akurasi yang cukup baik untuk melakukan sebuah prediksi.

Berdasarkan penelitian – penelitian yang dilakukan sebelumnya dengan tujuan untuk membandingkan beberapa algoritma tertentu dalam mengklasifikasikan SMS *spam*, maka pada penelitian ini akan difokuskan untuk mengidentifikasi jenis pesan yang mengandung SMS normal, SMS *fraud*/penipuan dan SMS promosi dengan mengoptimalkan *hyperparameter* algoritma *Multinomial Naïve Bayes* untuk mendapatkan model dengan performa dan akurasi yang baik.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Metodologi yang digunakan pada penelitian ini CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang merupakan standarisasi proses untuk menyelesaikan permasalahan pada penelitian dalam bidang *data mining* [10] [12]. Metode ini dapat digunakan baik untuk melakukan data *mining* mapun *text mining*. Perbedaan diantara data mining dan text mining adalah pada format datanya [11] [14]. Format input data pada data mining terstruktur sedangkan text mining tidak terstruktur [15]. Artinya model diekstraksi dari teks yang tidak terstruktur pada text mining sedangkan pada data mining digunakan data yang terstruktur [13]. Adapun tahapan penelitian yang akan dilakukan pada penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Diagram Alir Penelitian

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

2.1.1 Identifikasi Masalah

Pada identifikasi masalah terdapat beberapa metode yang diterapkan dengan tujuan agar analisis lebih terarah untuk mengidentifikasi masalah. Kumpulan informasi yang di hasilkan dari setiap metode akan menghasilkan rangkuman yang akan digunakan untuk menyelesaikan masalah. Adapun metode yang digunakan dalam identifikasi masalah yaitu studi litelatur yang merupakan metode pengumpulan data dengan cara membaca serta menelusuri pustaka dari kumpulan buku, jurnal maupun makalah untuk mendapatkan teori yang relevan dengan masalah penelitian. Penulis juga mempelajari dan menelusuri sumber – sumber tulisan atau penelitian yang dibuat sebelumnya terkait dengan apa yang penulis teliti. Tahapan ini menjadi acuan dalam menentukan metode pemecahan masalah.

2.1.2 Pemahaman Bisnis

Pada tahapan ini memahami dan menentukan tujuan untuk di proses dalam kegiatan data mining. Kemudian menterjemahkan pengetahuan tersebut ke dalam pendefinisian masalah. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut. Pada penelitian ini di perlukan pemahaman tentang latar belakang dan tujuan pada proses yang berkaitan dengan SMS *filtering*. Sedangkan tujuan dari penelitian ini adalah mencari pola (*pattern*) SMS untuk memprediksi SMS yang mengandung *spam* dan tidak mengandung *spam*.

2.1.3 Persiapan, Pemahaman dan Pengolahan Data

a. Pengumpulan Data (*Data Collection*)

Pada tahapan ini dilakukan pengumpulan data yang dibutuhkan sebagai bahan baku informasi yang akan diolah untuk mencapai tujuan penelitian. Adapun rincian data yang diperoleh berdasarkan jenis SMS dapat dilihat pada tabel 1.

Tabel 1. Rincian Jenis SMS

No	Jenis SMS	Jumlah
1	SMS Normal	569
2	SMS *Fraud* atau Penipuan	335
3	SMS Promosi	239
	Total	1143

b. Pemahaman Data

Pada tahap pemahaman data memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan (*summary*) dan mengidentifikasi potensi masalah dalam data. Ringkasan atau *summary* pada data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan. Dalam proses ini dilakukan explorasi data dengan menggunakan metode *Exploratory Data Analysis* (EDA) dengan menerapkan teknik aritmatika dan teknik grafis dalam meringkas data mempermudah pemahaman data dan mempermudah melihat *trend* dalam data.

c. *Pre-processing*

Pre-processing mencakup seluruh kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Tahapan ini dapat dilakukan berulang untuk mendapatkan data yang sesuai. Pada tahapan ini mencangkup langkah – langkah sebagai berikut :

1. *Case Folding*
   Proses *case folding* merupakan tahapan yang digunakan untuk mengubah setiap huruf pada kata menjadi bentuk yang standar, menghilangkan angka serta tanda baca, menghilangkan *whitespace* (karakter kosong), menghilangkan kalimat yang tidak digunakan dan menangani bentuk kata *slangs* atau singkatan menjadi kata aslinya berdasarkan *dictionary*.
2. *Stemming*
   Stemming merupakan proses ekstraksi atau normalisasi sebuah kata untuk mencari kata dasar (*root word*) berdasarkan hasil proses *filtering* dengan menghilangkan imbuhan *affixes* (semua imbuhan), *suffix* (imbuhan akhiran), *prefix* (imbuhan awalan) dan *confixes* (kombinasi imbuhan awalan dan akhiran).
3. *Tokenizing*
   Proses *tokenizing* berfungsi untuk merubah *text* atau *document* yang di input, sehingga dapat dipisahkan dan menjadi bagian kecil yang dapat di representasikan sebagai *token*. *Tokenizing* dalam penelitian ini menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Untuk dapat menghitung nilai TF-IDF dapat menggunakan persamaan berikut:

$$tf(t,d) = f(t,d) \quad (1)$$

$$idf(t) = ln\frac{1+n}{1+df(t)} + 1 \quad (2)$$

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \quad (3)$$

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{{v_1}^2 + {v_2}^2 + \cdots + {v_n}^2}} \quad (4)$$

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

4. *Stopword*
   Stopword merupakan proses *filtering* atau menghilangkan kata – kata yang tidak memiliki arti penting. Pada setiap *term* yang terdapat pada proses tokenisasi akan dicocokan ke dalam daftar kata *stopword* yang merupakan *dictionary* (kamus data) atau kumpulan kata yang tidak memiliki arti penting yang telah ditentukan sebelumnya.

2.1.4 Pemodelan (*Modeling*)

Algoritma yang digunakan untuk membangun model adalah *Multinomial Naïve Bayes* (MNB). Algortima ini merupakan pengembangan dari algoritma *Naïve Bayes* (NB) yang sering digunakan untuk pengklasifikasian data *teks*. Metode *Multinomial Naïve Bayes* (MNB) memiliki kesederhanaan dan efektifitas dalam melakukan proses klasifikasi kategori *document* atau *teks* dengan cara membandingkan gabungan kata dan kategori dengan metode *Naïve Bayes* [9].

a. Input Data
Tahapan *input data* merupakan proses untuk memasukan data yang telah dilakukan tahapan *pre-processing* ke dalam skenario model yang akan dilakukan uji coba. Input data terbagi menjadi dua bagian yaitu data *training* dan data *testing.* Data *training* adalah data SMS yang telah di identifikasi sesuai dengan jenis SMS sedangkan data *testing* adalah data yang akan diuji untuk membuktikan data SMS mengandung jenis SMS normal, fraud/penipuan dan promosi serta digunakan untuk mengetahui performa algoritma yang sudah di latih. Adapun contoh teks SMS yang akan di proses dapat dilihat pada tabel 2.

Tabel 2. Contoh Teks SMS

No	Isi Pesan	Keterangan
1	Selamat Anda Terpilih M-dapatkan HADIAH WHATSAPP Rp.175jta pin WHA012 Silahkan Klik : bit.ly/pt-whatsapp52	SMS Penipuan
2	Sudah dengan tentang 321? Ini adalah layanan berita dan informasi GRATIS dari XL! Cukup gunakan ponsel Anda untuk menelpon 321 hari ini. Yuk, coba sekarang ! MC759A	SMS Promosi

b. *Input Hyperparameter*
Pada tahapan ini merupakan proses membuat skenario *hyperparameter* yaitu *variable* yang akan mempengaruhi *ouput* model. Skenario *hyperparameter* dilakukan dengan mencari parameter yang tepat agar algoritma *Naïve Bayes* dapat mempelajari pola dalam data dengan baik

c. *Training*
Pada tahapan ini penerapan uji coba prediksi dengan algoritma *Naïve Bayes* untuk di implementasikan pada data *training*. Proses *training* bertujuan untuk melatih algoritma dan memberikan petunjuk data dalam memvalidasi model untuk mencari model yang terbaik.

d. Evaluasi Model
Setelah melakukan serangkaian proses, selanjutnya dilakukan evaluasi kinerja model dengan menggunakan metode *K-Fold Cross Validation*. *K-Fold Cross Validation* atau *Cross Validation* merupakan salah satu teknik yang digunakan untuk melakukan evaluasi model. Pada *Cross Validation*, *dataset* dibagi sebanyak K lipatan, dalam setiap iterasi setiap lipakan akan dipakai satu kali sebagai data uji dan lipatan sisanya dipakai sebagai data latih. Metode selanjutnya yaitu dengan *confusion matrix* yang merupakan metode untuk menampilkan dan membandingkan nilai aktual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan untuk menghasilkan matrik evaluasi seperti *accuracy*, *precision*, *recal* dan *f1-score*. Sehingga penggunaan *confusion matrix* ini akan memberitahu seberapa baik model yang dibuat. *Confusion matrix* memiliki empat nilai yang dihasilkan dalam tabel diantaranya TP (*True Positive*), TN(*True Negative*), FP (*False Positive*) dan FN (*False Negative*).

2.1.5 Penyebaran (*Deployment*)

Setelah melalui serangkaian proses, selanjutnya pada hasil pengolahan data yang telah di proses menjadi pengetahuan atau informasi dapat di implementasikan dan digunakan untuk mengklasifikasikan jenis SMS yang diterima oleh pengguna.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan *dataset* SMS berbahasa Indonesia yang bersumber dari situs *website* kaggle. Data yang diperoleh disajikan dalam format .csv dan berjumlah 1143 *record* dengan 2 kolom. Adapun isi tabel SMS berbahasa Indonesia dapat disajikan pada gambar 3.

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

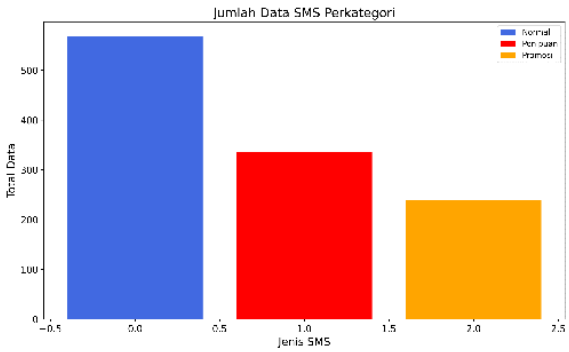
	sms	kategori
0	[PROMO] Beli paket Flash mulai 1GB di MY TELKO...	2
1	2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A...	2
2	2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ...	2
3	2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ...	2
4	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an...	2
...	...	...
1138	Yooo sama2, oke nanti aku umumin di grup kelas	0
1139	sebelumnya ga ad nulis kerudung. Kirain warn...	0
1140	Mba mau kirim 300 ya	0
1141	nama1 beaok bwrangkat pagi...mau cas atay tra...	0
1142	No bri atas nama kamu mana	0

1143 rows × 2 columns

Gambar 2. Isi SMS Berbahasa Indonesia

3.2 Pemahaman Data

Tahapan pemahaman data perlu dilakukan untuk memberikan memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan (*summary*) dan mengidentifikasi potensi masalah dalam data sebelum melakukan tahapan *pre-processing* dan *modeling*. Untuk memahami isi data dapat menggunakan metode *Exploratory Data Analysis* (EDA) yaitu teknik untuk mengeksplorasi data dengan menerapkan teknik aritmatika dan teknik grafis dalam meringkas dan mempermudah pemahaman data. Adapun hasil dari proses *Exploratory Data Analysis* dapat disajikan dalam gambar 4.



Gambar 3. Diagram Jumlah Data SMS Perkategori

Berdasarkan hasil dari eksplorasi data SMS berbahasas Indonesia, maka dapat disimpulkan total data pada setiap jenisnya memiliki data yang beragam atau tidak seimbang (*Imbalance*). Jenis SMS normal memiliki total data sebesar 556, jenis SMS penipuan memiliki total data 335 dan jenis SMS promosi memiliki total data 239.

Gambar 4. *Wordcloud* Data SMS Berbahasa Indonesia

Pada gambar 5 melakukan teknik visualisasi data teks dengan metode *wordcloud* yang memiliki fungsi untuk memunculkan frekuensi kata – kata, semakian sering suatu kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam *wordcloud*. Melalui *wrodcloud* tersebut dapat diketahui bahwa kata penghubung dan kata bantu seperti “di”, “ke”, “dan”, “yang”, “ini”. Serta kata – kata lainnya yang tidak

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

memiliki informasi atau arti penting yang tampak dominan. Pada pengolahan data teks, kata – kata tersebut akan sering diabaikan sehingga grafik yang dihasilkan lebih informatif.

3.3 *Pre-processing* Data

Pada tahapan *pre-processing* data dalam penelitian ini memiliki beberapa tahapan pemrosesan yaitu *Case Folding*, *Stemmimg*, *Tokenizing* [16] [17].

3.3.1 Case Folding

Pada proses *case folding* memiliki tahapan – tahapan yang dapat dilakukan. Tahapan awal merupakan proses untuk merubah setiap huruf pada kata menjadi huruf kecil (*lowercase*) dan menghilangkan *whitespace* (karakter kosong). Selanjutnya pada tahapan *case folding* dilakukan proses untuk menghilangkan tanda baca, angka dan kata yang tidak digunakan. Pada penelitian ini kata yang akan dihilangkan meliputi kata yang mengandung sebuah *link* situs *website*. Pada proses selanjutnya dalam tahapan *case folding* adalah menangani bentuk kata *slangs* atau kata singkatan menjadi kata aslinya berdasarkan *dictionary* yang telah di tentukan. Adapun cuplikan data dalam proses *case folding* dapat dilihat pada tabel 3.

Tabel 3. Data Setelah Proses *Case Folding*

No	Teks SMS
1	pelanggan yang terhormat simcard anda sebagai pemenang mendptkan hadiah mobil toyota all new yaris pin jf dari gebyar undian care untuk informasi lenkap klik
2	selamat kepada pelanggan xxx anda mendapat hadiah ke dalam program poin plus plus indosat nomor pin fg informasi klik
3	selamat anda mendapatkan satu unit mobil nissan juke dari pt indofood pin anda tdn untuk informasi pengambilan hadiah kunjungi website

3.3.2 *Stemming*

Stemming merupakan proses untuk mencari kata dasar (*root word*) dengan menghilangkan semua imbuhan yang ada. Proses *stemming* yang dilakukan pada penelitian ini menggunakan *library* sastrawi yang menerapkan algoritma Nazief & Andriani. Adapun cuplikan data yang telah dilakukan proses *stemming* dapat dilihat pada tabel 4.

Tabel 4. Data Setelah Proses *Stemming*

No	Teks SMS
1	ini hujan malam ruang tengah sama kamar kayak tidak bisa di nyalain harus tunggu kering baru bisa nyala moga hujan renti
2	ibu lagi tidur aku mau tanya tentang daging kalau ada daging mau di presto dahulu juga

3.3.3 *Tokenizing*

Tokenisasi berfungsi untuk merubah *text* atau *document* menjadi bagian kecil yang di reperentasikan sebagai *token.* Pada penelitian ini proses tokenisasi menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Tokenisasi yang dilakukan berdasarkan hasil pembagian partisi data dengan komposisi 75:25 dan proses tokenisasi memperoleh 491 token dengan jumlah data 857 *rows*. Berikut contoh hasil dari proses tokenisasi yang telah di urutkan sesuai abjad dapat dilihat pada tabel 5.

Tabel 5. Cuplikan Hasil Tokenisasi

No	*Token*
1	aamiin
2	abang
3	adik
4	agen
5	agustus
	…
491	your

3.3.4 Stopword

Berdasarkan hasil eksplorasi data teks dengan metode *wordcloud* pada gambar 5. Kata – kata yang sering muncul dan jarang sekali muncul memiliki sedikit sekali informasi dan tidak memiliki arti penting, sehingga dapat dilakukan proses *stopword* untuk dapat disaring dan dihilangkan dari *document.*

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

```
# Menambahkan Stopword Pendukung yang Dibutuhkan
new_stp = ['aa', 'at','agt', 'an', 'asa', 'bal', 'bfg', 'co', 'com', 'cx','da'

stopwords = stopwords + new_stp
print(len(stopwords))
print(stopwords)
```

```
864
['a', 'ada', 'adalah', 'adanya', 'adapun', 'agak', 'agaknya', 'agar', 'akan',
```

Gambar 5. Total Daftar *Stopword*

Berdasarkan pada gambar 6 total daftar kata *stopword* yang digunakan adalah 864 kata. Selanjutnya proses implementasi *stopword* dengan menerapkan frekuensi *filtering* yang terdiri dari menentukan nilai *minimum document frekuency* sebesar 0,003 atau 0.3% dan nilai *maximum document frekuency* sebesar 0,99 atau 99% untuk dapat digunakan pada proses menentukan jumlah token yang akan digunakan pada tahapan tokenisasi.

3.4 Pemodelan (*Modeling*)

Untuk membuat dan menentukan sebuah model MNB (*Multinomial Naïve Bayes*) digunakan beberapa perbandingan algoritma dan beberapa parameter diantarnya adalah menentukan nilai *minimum document frequency* untuk menghilangkan token dengan range *minimum* yang telah di tentukan, *maximum document frequency* untuk menghilangkan token dengan range *maximum* yang telah di tentukan, menentukan nilai *alpha* yang merupakan *smoothing factor* untuk mengatasi permasalahan peluang token yang bernilai nol dan menentukan *fit prior* yang digunakan untuk mempelajari probabilitas hipotesis atau sampel.

Tabel 6. Parameter Pengujian Model MNB

Parameter	Keterangan
Input	Data SMS Berbahasa Indonesia
Min DF	0.001 – 0.005 *frequency*
Max DF	0.1 – 0.99 *frequency*
Alpha	0.01 – 5.0
Fit Prior	True or False

Pada pengujian tahap pertama, dilakukan untuk mendapatkan nilai *minimum document frequency* dengan menggunakan pembagian partisi data pada komposisi 75% data *training* dan 25% data *testing*. Berdasarkan hasil percobaan didapatkan hasil akhir pada tabel 7.

Tabel 7. Hasil Pengujian *Min DF*

Minimum Document Frequency				
Frequency	*Macro Avg Precision*	*Macro Avg Recall*	*Macro Avg F1-Score*	*Accuracy*
0.001	0.92	0.92	0.92	0.93
0.002	0.92	0.92	0.92	0.93
0.003	0.93	0.93	0.93	0.94
0.004	0.92	0.91	0.92	0.93
0.005	0.91	0.91	0.91	0.92

Setelah melakukan pengujian tahap pertama, dilakukan pengujian tahap kedua untuk mendapatkan nilai *maximum document frequency* yang optimal. Pengujian tahap kedua dilakukan dengan menggunakan komposisi data 75:25 dan nilai *minimum document frequency* 0.003 berdasarkan hasil pengujian tahap pertama. Adapaun hasil percobaan yang dilakukan didapatkan hasil akhir pada tabel 8.

Tabel 8. Hasil Pengujian *Max DF*

Maximum Document Frequency				
Frequency	*Macro Avg Precision*	*Macro Avg Recall*	*Macro Avg F1-Score*	*Accuracy*
0.1	0.91	0.92	0.91	0.92
0.5	0.93	0.93	0.93	0.94
0.6	0.93	0.93	0.93	0.94
0.7	0.93	0.93	0.93	0.94
0.8	0.93	0.93	0.93	0.94
0.99	0.93	0.93	0.93	0.94

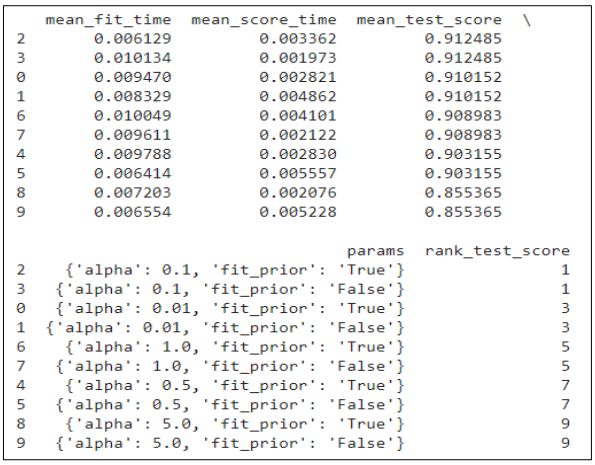
Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

Setelah melakukan pengujian tahap kedua, selanjutnya dilakukan pengujian tahap ketiga dan tahap keempat untuk mendapatkan nilai *alpha* dan *fit prior* yang optimal. Pengujian tahap ketiga dilakukan dengan menggunakan komposisi data 75:25 dan nilai *minimum document frequency* 0.003 berdasarkan hasil pengujian tahap pertama dan nilai *maximum document frequency* 0.99 berdasarkan pengujian tahap kedua. Adapaun hasil percobaan yang dilakukan didapatkan hasil akhir pada tabel 9 dan gambar 7.

Tabel 9. Hasil Pengujian *Alpha* dan *Fit Prior*

Alpha & Fit Prior					
Alpha	*Fit Prior*	*Macro Avg Precision*	*Macro Avg Recall*	*Macro Avg F1-Score*	*Accuracy*
0.01	True	0.930505	0.922175	0.925568	0.937062
0.01	False	0.917043	0.900131	0.906710	0.919580
0.1	True	0.934834	0.924773	0.929071	0.940559
0.1	False	0.921372	0.902749	0.910436	0.923076
0.5	True	0.931826	0.928591	0.930118	0.940559
0.5	False	0.912477	0.891332	0.899232	0.912587
1.0	True	0.931826	0.928591	0.930118	0.940559
1.0	False	0.919089	0.898990	0.907065	0.919580
5.0	True	0.844040	0.902959	0.865354	0.884615
5.0	False	0.914523	0.891791	0.901338	0.912587



```
   mean_fit_time  mean_score_time  mean_test_score  \
2       0.006129         0.003362         0.912485
3       0.010134         0.001973         0.912485
0       0.009470         0.002821         0.910152
1       0.008329         0.004862         0.910152
6       0.010049         0.004101         0.908983
7       0.009611         0.002122         0.908983
4       0.009788         0.002830         0.903155
5       0.006414         0.005557         0.903155
8       0.007203         0.002076         0.855365
9       0.006554         0.005228         0.855365

                                  params  rank_test_score
2    {'alpha': 0.1, 'fit_prior': 'True'}                1
3   {'alpha': 0.1, 'fit_prior': 'False'}                1
0   {'alpha': 0.01, 'fit_prior': 'True'}                3
1  {'alpha': 0.01, 'fit_prior': 'False'}                3
6    {'alpha': 1.0, 'fit_prior': 'True'}                5
7   {'alpha': 1.0, 'fit_prior': 'False'}                5
4    {'alpha': 0.5, 'fit_prior': 'True'}                7
5   {'alpha': 0.5, 'fit_prior': 'False'}                7
8    {'alpha': 5.0, 'fit_prior': 'True'}                9
9   {'alpha': 5.0, 'fit_prior': 'False'}                9
```

Gambar 6. Hasil Pengujian *Alpha* dan *Fit Prior* Pada GridSearchCV

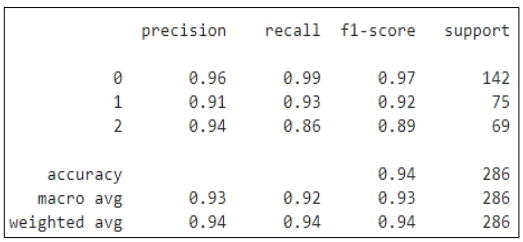
3.4.1 Hasil Perbandingan Algoritma

Selain menentukan *hyperparameter* terbaik dalam mendapatkan model yang optimal, selanjutnya perlu dilakukan pengujian untuk membandingkan serta menentukan algoritma yang akan digunakan pada penelitian ini. Adapun beberapa algoritma yang akan digunakan dapat dilihat pada tabel 10.

Tabel 10. Algoritma Perbandingan

No	Algoritma
1	*Multinomial Naïve Bayes*
2	*Support Vector Machine*
3	*Random Forest Classifier*

Maka berikut adalah perbandingan *accuracy* dari ketiga algoritma yang dapat dilihat pada gambar 8, gambar 9 dan gambar 10.



```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97       142
           1       0.91      0.93      0.92        75
           2       0.94      0.86      0.89        69

    accuracy                           0.94       286
   macro avg       0.93      0.92      0.93       286
weighted avg       0.94      0.94      0.94       286
```

Gambar 7. Hasil Matrik Evaluasi Algoritma MNB

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

	precision	recall	f1-score	support
0	0.99	0.95	0.97	152
1	0.92	0.96	0.94	74
2	0.87	0.92	0.89	60
accuracy			0.95	286
macro avg	0.93	0.94	0.94	286
weighted avg	0.95	0.95	0.95	286

Gambar 8. Hasil Matrik Evaluasi Algoritma SVM

	precision	recall	f1-score	support
0	0.98	0.95	0.96	151
1	0.90	0.93	0.91	74
2	0.87	0.90	0.89	61
accuracy			0.93	286
macro avg	0.92	0.93	0.92	286
weighted avg	0.94	0.93	0.93	286

Gambar 9. Hasil Matrik Evaluasi Algortima RFC

Berdasarkan hasil yang diperoleh gambar 8, gambar 9 dan gambar 10 nilai *accuracy* pada setiap algoritma memiliki selisih yang tidak terlalu jauh. Untuk nilai *accuracy* tertinggi terdapat pada algoritma SVM sebesar 0.95. Selanjutnya *accuracy* tertinggi pada urutan kedua terdapat pada algoritma MNB sebesar 0.94 dan untuk *accuracy* terendah terdapat pada algoritma RFC sebesar 0.93.Selanjutnya dilakukan pengujian waktu yang dihasilkan dari proses pada masing – masing algoritma. Berikut adalah hasil pengujian yang dapat dilihat pada gambar 11.

```
%timeit model_mnb.fit(sms_train_vect, label_train)
%timeit model_svm.fit(sms_train_vect, label_train)
%timeit model_rf.fit(sms_train_vect, label_train)

100 loops, best of 5: 2.66 ms per loop
1 loop, best of 5: 3.45 s per loop
1 loop, best of 5: 422 ms per loop
```

Gambar 10. Hasil Waktu Pengujian Algoritma

Berdasarkan hasil yang diperoleh pada gambar 11 algoritma MNB memiliki waktu pengujian yang lebih cepat yaitu sebesar 2.66 ms dalam 1 *loops* dari total 100 *loops* pengujian, sedangkan untuk algoritma RFC memiliki total waktu pengujian yang dihasilkan sebesar 422 ms dalam 1 *loops* dan untuk algoritma SVM memiliki total waktu pengujian yang jauh lebih lambat dari kedua algoritma dengan nilai waktu pengujian sebesar 3.45 s dalam 1 *loops*. Setelah dilakukan pengujian model berdasarkan matrik evaluasi dan waktu pengujian yang dihasilkan pada masing – masing algoritma, maka dengan demikian pada penelitian ini algoritma yang akan digunakan untuk mengklasifikasikan SMS *spam* adalah algoritma *Multinomial Naïve Bayes* (MNB).

4. KESIMPULAN

Berdasarkan pengujian sistem menggunakan metode *Multinomial Naïve Bayes* (MNB) pada identifikasi SMS berbahasa Indoensia yang mengandung *spam* dan tidak mengandung *spam* didapatkan kesimpulan model optimal yang didapatkan dari algoritma Multinomial Naïve Bayes (MNB) dengan partisi 75:25, minimum document frequency 0.003 atau 0.3 %, maximum document frequency 0.99 atau 99%, alpha 0.1 dan fit prior True menghasilkan nilai rata – rata matrik evaluasi precision 0.93%, recall 0.92%, f1-score 0.93 dan accuracy 0.94%. Serta menghasilkan waktu pengujian lebih cepat sebesar 2.66 ms dalam 1 loops dari total 100 loops pengujian dibandingkan 2 algoritma yang telah diuji sebelumnya serta model mampu melakukan klasifikasi SMS berbahasa Indonesia yang mengandung *spam* dan tidak mengandung *spam.*

UCAPAN TERIMAKASIH

Terima kasih kami ucapkan kepada pimpinan prodi Teknik Informatika UNKRIS yang telah mendukung dalam terlaksananannya penelitian ini.

REFERENCES

[1] Widyawati and Sutanto, “Perbandingan Algoritma Naive Bayes Dan Support Vector Machine (SVM),” *J. Sains Teknol.*, vol. 3, no. 2, pp. 178–194, 2019.

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1316-1325
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3119

[2] M. R. Akhyari and A. R. Pratama, "Kesadaran akan Ancaman Serangan Berbasis Backdoor di Kalangan Pengguna Smartphone Android," *Automata*, vol. 2, no. 1, pp. 1–7, 2021.
[3] K. F. Kok, "Top 20 Countries Affected by Spam Calls in 2020," *truecaller*, 2020. [Online]. Available: https://truecaller.blog/2020/12/08/truecaller-insights-top-20-countries-affected-by-spam-calls-in-2020-2/. [Accessed: 08-Dec-2020].
[4] M. A. F. Syahril, "Privasi Yang Terpublikasi," pp. 1–14, 2021.
[5] B. Susilo, "Pengaruh Penggunaan Media Sosial Terhadap Kesadaran Registrasi Kartu Prabayar Di Pontianak," *SENSITEK*, pp. 121–126, 2018.
[6] Apriliana, N. Ransi, and J. Nangi, "Implementasi Text Mining Klasifikasi Skripsi Menggunakan Metode Naïve Bayes Classifier," *Semant. Vol.3, No.2, Jul-Des 2017*, vol. 3, no. 2, pp. 187–194, 2017, doi: 10.1007/978-1-4471-7307-6_20.
[7] D. N. Fitriana, N. A. Setifani, and A. Yusuf, "Perbandingan Algoritma Naïve Bayes, Svm, Dan Decision Tree Untuk Klasifikasi SMS Spam," *JUSIM (Jurnal Sist. Inf. Musirawas)*, vol. 5, no. 02, pp. 167–174, 2020, doi: 10.32767/jusim.v5i02.956.
[8] A. S. Dharma, O. Y. Silitonga, and H. J. Manurung, "Perbandingan Algoritma Naive Bayes, ID3 dan TAN Pada Klasifikasi SMS Spam," *J. Marit. Educ.*, vol. 1, no. 2, pp. 30–34, 2019.
[9] N. Hayatin, "Implementasi Multinomial Naïve Bayes Untuk Klasifikasi Data Tweets Mengandung Term," *SENTRA*, pp. 344–349, 2020.
[10] Chapman, P., Kerber R., Clinton J., Khabaza T., Reinartz T., Wirth R. – "The CRISP-DM Process Model", Discussion Paper, 2000.
[11] E. M. Silval, H. A. do Pradol, E. Femedal, Text mining: crossing the chasm between the academy and the industry", Paper from: Data Mining III, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors), 2002
[12] Lukasza, Kurgan, and Petrmusilek, " A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, Vol. 21, 2006
[13] S. A. Salloum, M. Al-Emran, A. A. Monem, K. Shaalan, "A Survey of text mining in social media: facebook and twitter perspectives", Ad-vances in Science, Technology and Engineering Systems Journal, Vol. 2, 2017
[14] W. Hua, Z. Wang, H. Wang, K. Zheng and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE Transactions On Knowledge And Data Engineering, 2016.
[15] J. Zhu, Member, K. Wang, Y. Wu, Zhongyi Hu, and H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transactions On Knowledge And Data Engineering, Vol. 28, 2016.
[16] Anil Kumar Soni, Avinash Kumar, Robin Prakash Mathur, "Enhancing the Stemming Algorithm in Text Mining", International Journal of Applied Engineering Research, Vol. 10, 2015
[17] C.Ramasubramanian1, R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", IJARCCE Vol. 2, Issue 12, 2013.

Submitted: **04/07/2021**; Accepted: **05/10/2021**; Published: **26/10/2021**