JNATIA Volume 3, Nomor 2, Februari 2025 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya e-ISSN: 3032-1948

Komparasi Ekstraksi Fitur BoW dan TF-IDF untuk Klasifikasi SMS Menggunakan Naive Bayes

I Komang Dwiprayoga[a1], Made Agung Raharja[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]komangyoga835@gmaill.com
[2]made.agung@unud.ac.id

Abstract

Short Message Service (SMS) has become one of the most popular communication media. However, the ease and speed of sending SMS is also utilized by irresponsible parties to send spam messages. These spam messages not only annoy users but can also cause financial losses and theft of personal data. The purpose of this research is to compare feature extraction methods that have the best performance such as TF-IDF and Bag of Word tested with Multinomial Naive Bayes machine learning algorithm. For the first research stage, load dataset, data balancing, data preprocessing, feature extraction, modeling with machine learning algorithms, and then testing and comparing confusion matrix models on each feature extraction. The results of this study show that the use of BoW feature extraction has better performance than the TF-IDF feature extraction model with an accuracy value of 94.44%.

Keywords: *Back of Words,TF-IDF, Multinomial Naive Bayes, Sms, Text Classification*

1. Pendahuluan

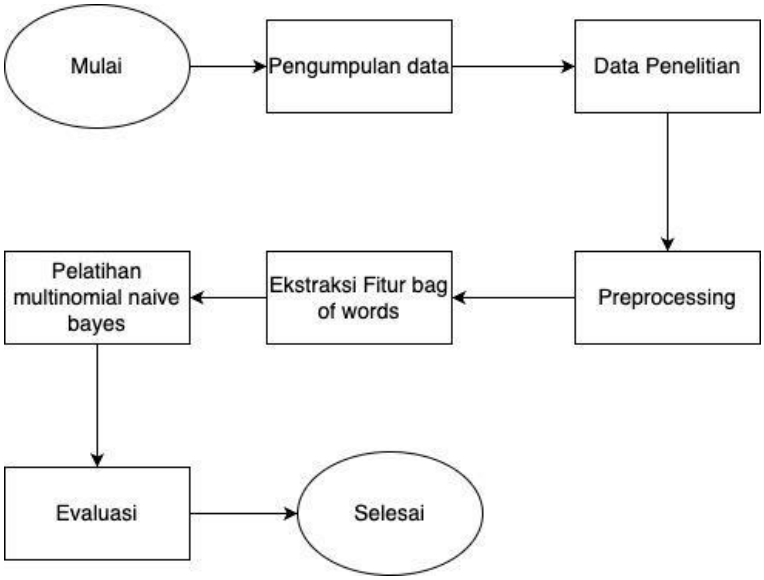
Era ini, data dan informasi merupakan komoditi utama yang dapat diperjualbelikan yang dengan mudah dapat diakses oleh pengguna dan pelanggan. Semuanya itu membawa masyarakat ke dalam suasana yang disebut oleh John Naisbitt, Nana Naisbitt dan Douglas Philips Sebagai "Zona Mabuk Teknologi" [1]. Spam, juga disebut sebagai unsolicited commercial email atau unsolicited bulk email telah menyebabkan beberapa masalah komunikasi dalam kehidupan sehari hari kita. Kerugian yang disebabkan karena spam antara lain spam menempati sumber daya yang besar (termasuk bandwidth jaringan, ruang penyimpanan), contoh kasus spam bisa berupa iklan perjudian maupun pornografi. Spam atau junk email adalah penyalahgunaan dalam pengiriman berita elektronik untuk menampilkan berita, iklan, dan keperluan lainnya yang mengakibatkan ketidaknyamanan bagi para pengguna [2]. Penipuan melalui SMS menjadi ancaman serius karena skema penipuan semakin kompleks dan dapat disesuaikan. Serangan pencurian identitas yang merugikan secara finansial hingga phishing, metode yang menyamar sebagai upaya peretasan dengan menggunakan pesan palsu [3]. Tujuan dilakukan penelitian ini adalah untuk meningkatkan akurasi dari metode yang digunakan sebelumnya oleh peneliti dengan menggunakan metode multinomial naive bayes dengan ekstraksi fitur bag of words. Spam atau penipuan ini dapat diklasifikasikan dengan melihat pola dan kata-kata yang digunakan. Pesan yang masuk ke ponsel dapat diklasifikasikan lagi menjadi tiga jenis pesan, yakni spam/penipuan, promo, dan normal.

2. Metode Penelitian

Pada tahap ini diawali dengan mengumpulkan data penelitian yang akan digunakan, selanjutnya tahap load dataset ke python notebook, kemudian dari data mentah tersebut akan dilakukan data balancing atau peyeimbangan data, kemudian dilanjutkan dengan tahap preprocessing data untuk membersihkan data dari kata-kata yang tidak diperlukan. Kemudian dilakukan pemisahan data latih dan data uji sebelum dilakukan tahap ekstraksi fitur. Kemudian tahap selanjutnya tahap perhitungan ekstraksi fitur menggunakan bag of words dan TF-IDF. Kemudian lanjut ke tahap

p-ISSN: 2986-3929
e-ISSN: 3032-1948

pelatihan model menggunakan multinomial naive bayes. Tahap terakhir adalah proses evaluasi menggunakan confusion matrix untuk menentukan akurasi dari model yang telah dilatih sebelumnya.



Gambar 1. Alur Penelitian

2.1. Pengumpulan Data

Pada penelitian ini, data yang digunakan adalah data sekunder yang diambil langsung di github peneliti yang sebelumnya juga melakukan penelitian terkait. Dataset yang diambil berupa teks atau pesan sms dengan banyak dataset berjumlah 1143 data yang memiliki tiga label, yakni berawal dari sms normal dengan jumlah 569 data yang direpresentasikan dengan index 0, sms penipuan atau spam dengan jumlah 335 data yang direpresentasikan dengan index 1, dan sms promo dengan jumlah 239 data yang direpresentasikan dengan index 2. Gambaran data yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Gambaran Dataset

id	Teks	Label
1	[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	Yooo sama2, oke nanti aku umumin di grup kelas	0
3	""ROXI CELL" Hanya dengan Rp.100rb Anda bisa jadi agen pulsa elektrik ke semua Operator GSM/CDMA, v5=4300, v10=8300, Utk daftar ketik MTRONIK kirim ke 087870870707	1

Tabel 2. Distribusi Data

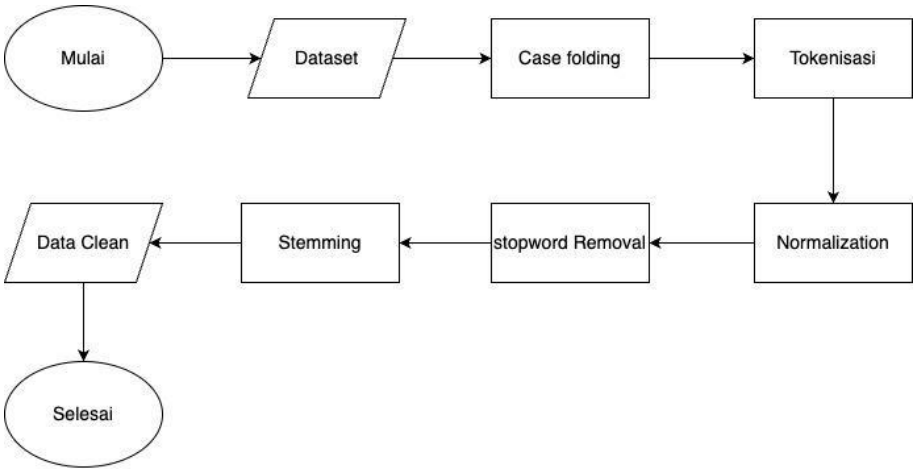
id	Label	Data
1	SMS Normal	569
2	SMS Spam	335
3	SMS Promo	239

2.2. Data Balancing

Jumlah data di setiap label tidak seimbang, yakni 569 data untuk label normal, 335 data untuk sms spam, dan 239 data untuk sms promo. Oleh karena itu, perlu dilakukannya proses penyeimbangan data di setiap label. Pada penelitian ini, peneliti menggunakan teknik random under sampling untuk meyeimbangkan jumlah data di setiap label. Random under sampling merupakan. Pendekatan undersampling dan oversampling adalah teknik standar yang digunakan dalam menangani data yang tidak seimbang, namun keduanya memiliki keterbatasan masing-masing. Misalnya, undersampling menyebabkan lebih banyak penghapusan sampel data yang pada akhirnya menyebabkan masalah kekurangan data, dengan peningkatan kemungkinan kehilangan data penting. sementara oversampling menyebabkan duplikasi data asli, sehingga menyebabkan overfitting kelas minoritas [4].

2.3. Preprocessing

Text preprocessing adalah tahap pertama dalam klasifikasi teks yang mengubah data teks asli yang tidak terstruktur menjadi data yang terstruktur sekaligus juga untuk mengidentifikasi fitur dari teks yang paling signifikan untuk membedakan antara kategori teks [5]. Tahap ini akan menghasilkan data teks yang siap digunakan untuk proses selanjutnya. Adapun tahapan dalam text preprocessing ini ditunjukkan pada Gambar 2.



Gambar 2. Alur Text Preprocessing

2.4. Bag of Words

Bag of Words (BoW) adalah sebuah teknik ekstraksi fitur untuk merepresentasikan dokumen teks ke dalam bentuk matriks. Teknik ini bekerja dengan cara mempelajari seluruh kosakata dari dokumen, lalu memodelkan tiap dokumen dengan menghitung jumlah kemunculan tiap katanya [6].

2.5. TF-IDF

TF-IDF adalah metode pembobotan kata yang digunakan untuk menentukan pentingnya sebuah kata dalam sebuah dokumen. Ada dua komponen dalam perhitungan nilai TF – IDF, yaitu TF (Term Frequency) dan IDF (Inverse Document Frequency). TF menentukan pentingnya suatu kata relatif terhadap kemunculannya dalam suatu dokumen, sedangkan IDF menentukan suatu kata penting dalam suatu dokumen jika tidak sering muncul di dokumen lain [7]. Rumus dari TF-IDF adalah sebagai berikut:

$$w\,(d,t) \;=\; TF\,(d,t) \;\times\; \log\left(\frac{N}{df(t)}\right) \tag{1}$$

2.6. Multinomial Naive Bayes

Multinomial Naïve Bayes merupakan implementasi algoritma Naïve Bayes yang umumnya digunakan dalam pemrosesan teks dengan mengikuti prinsip distribusi multinomial [3]. Penggunaan model distribusi multinomial menunjukkan bahwa vektor fitur dalam suatu dokumen dibentuk dari frekuensi kemunculan setiap kata pada dokumen tersebut. Adapun algoritma perhitungannya [8]. Algoritma ini akan menghitung probabilitas sebuah dokumen d terhadap kelas C yang ditunjukkan pada Persamaan 2 [3].

2.7. Evaluasi

Tahap evaluasi dilakukan dengan menguji model dengan data uji yang telah dibagi sebelumnya dengan porsi 20%, yakni 229 data uji dengan menggunakan confusion matrix. Hasil evaluasi akan menghasilkan nilai akurasi, recall, precision, dan f1 score. Adapun rumus yang digunakan untuk melakukan perhitungan dari akurasi, recall, presisi, dan f1 score adalah sebagai berikut:

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \times 100\% \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{FP+FN}} \quad (4)$$

$$\text{F1} - \text{Score} = \frac{2(\text{Precision+Recall})}{(\text{Precision+Recall})} \quad (5)$$

3. Hasil dan Diskusi

3.1. Data Balancing

Dataset dengan jumlah data 1143 akan di-reduce menjadi 771 data. Data sms spam yang tadinya berjumlah 569 data berkurang menjadi 239 data, begitu juga dengan jumlah data sms spam yang tadinya 335 data menjadi 239 karena mengikuti data minoritas atau dalam penelitian ini yang menjadi data minoritas adalah sms promo sejumlah 239 data. Hasil penyeimbangan data menggunakan teknik random undersampling dapat dilihat pada tabel 2 berikut:

Tabel 3. Hasil Undersampling

id	Teks	Label
0	Maaf, Keyword yang anda masukkan ke 234 salah. Info lanjut hub call center: 200	0
1	Gais aku sudah brgkt. Mau matiin hp	0
2	"Kan bapaknya ngomong2 ya, trs ditanya siapa itu, aku blg lg rame ma sibuk dikampus nanti yaa"	0
3	Gimana nuy menghadapi beliau?	0
4	"Yaudah sekarang mah eta harddisk di laptop maneh keluarin terus dijadiin harddisk external aja.. cari temen yang punya case nya.. entar datanya copy-in, paling maneh ngerjain di luar dulu aja ga di laptop itu"	0
712	Paket Flash anda 10 MB utk 1 hari akan berakhir pd 08/07/2016. Tarif non paket berlaku setelah tgl tsb. Info hubungi *363#. Cek Kuota *889#.	2
713	Terima kasih telah menjadi Sahabat IndosatOoredoo. Mau tahu RAHASIA2 TERHEBOH ARTIS-ARTIS TOP + bonus gratis nelpon 60 menit?Tlp *700*1#	2

id	Teks	Label
714	Pembelian Dave's Single Burger Gratis Small Fries & Minum. Periode 5-11 September 2016. Tukarkan SMS ini di Wendy's terdekat! Penawaran terbatas.Promo *606#	2
715	Selamat Paket 50 SMS ke semua operator Anda telah aktif. Cek kuota di *888#	2
716	Pakai XL tdk perlu repot setting bisa langsung internetan. Yuk aktifkan paket Internet di My.XL.co.id sekarang.	2

Tabel 4. Hasil Distribusi Menggunakan Undersampling

id	Label	Data
1	SMS Normal	239
2	SMS Spam	239
3	SMS Promo	239

3.2. Preprocessing

Dataset yang berjumlah 771 data dilakukan tahap preprocessing atau pembersihan teks, tahap ini dilakukan untuk meningkatkan performa akurasi dari model yang akan dilatih setelahnya. Hasil dari text preprocessing dapat dilihat dari tabel 2 berikut ini:

Tabel 5. Preprocessing

id	Proses	Teks
1	Data mentah	Besok kekantor saya tunggu jam 9 ya.
2	Case Folding	besok kekantor saya tunggu jam ya
4	Normalization	besok kekantor saya tunggu jam ya
5	Stopword Removal	besok kekantor tunggu jam ya
6	Stemming	besok kantor tunggu jam ya

3.3. Ekstraksi Fitur

Setelah mendapatkan data yang bersih dari tahap preprocessing, maka tahap selanjutnya ke tahap ekstraksi fitur menggunakan Bag of Words dan TF-IDF. Perhitungan ekstraksi fitur menggunakan bantuan dari library scikit-learn dengan modul countvectorizer dan Tfidfvectorizer. Hasil dari perhitungan didapatkan sebanyak 3255 fitur dari 573 data latih. Fitur yang didapat terbilang banyak mengingat jumlah data latih yang digunakan hanya 573 data yang berupa teks. Berikut merupakan hasil dari ekstraksi fitur menggunakan BoW dan TF-IDF:

Ambil data pada dataset yang telah melalui preprocessing:
Dokumen 1 = "transfer tipon nomor ubah nomor rekening terima kasih"
Dokumen 2 = "bayar transfer harap hubung atas sayabpk ardiansah nomor telepon ubah bayar"
Dokumen 3 = “nomor papa tolong kirimin pulsa nomor papa papa kantor polisi telepon papa"

Tabel 6. Hasil Pengujian Ekstraksi Fitur BoW

Doc	ardiansah	atas	bayar	……..	tipon	tolong	transfer	ubah
Doc1	0	0	0	……..	1	0	1	1
Doc2	1	1	2	……..	0	0	1	2

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Doc	ardiansah	atas	bayar	…….	tipon	tolong	transfer	ubah
Doc3	0	0	0	…….	0	1	0	0

Tabel 7. Hasil Pengujian Ekstraksi Fitur TF-IDF

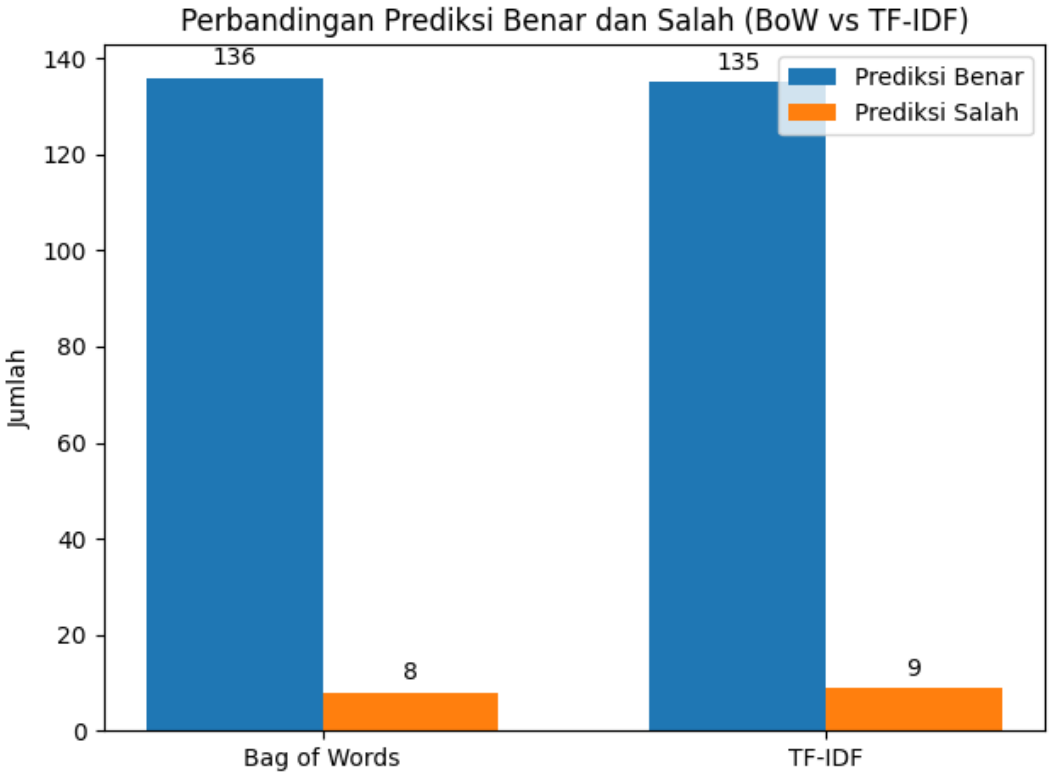
Doc	ardiansah	atas	bayar	…….	tipon	tolong	transfer	ubah
Doc1	0.00000	0.00000	0.00000	…….	0.32493	0.00000	0.32493	0.26303
Doc2	0.301647	0.301647	0.512236	…….	0.00000	0.00000	0.49580	0.49580
Doc3	0.00000	0.00000	0.00000	…….	0.00000	0.00000	0.00000	0.00000

3.4. Pemodelan

Pemodelan multinomial naive bayes dilakukan dengan bantuan library scikit-learn dari python. Model akan dilatih dengan data latih yang telah melalui proses ekstraksi fitur menggunakan Bag of Words dan TF-IDF. Pemodelan juga menggunakan hyperparameter alpha terbaik yang nanti dicari menggunakan Gridsearch. Setiap ekstraksi fitur akan melalui proses evaluasi menggunakan Gridsearch untuk mencari hyperparameter terbaik dan dengan proses cross validation dengan banyak fold sebanyak 5 kali fold.

3.5. Evaluasi

Tahap terakhir adalah tahap pengujian dari model yang telah dilatih dengan menggunakan confusion matrix. Data yang diujikan merupakan data uji yang telah dibagi sebelumnya sebanyak 144 data uji. Setelah melalui proses pengujian, maka didapat nilai akurasi, presisi, recall, dan f1-score terhadap model yang telah dilatih. Hasil pengujian menggunakan confusion matrix dapat dilihat pada gambar 3 dan tabel 8.



Gambar 3. Hasil Prediksi BoW dan TF-IDF

Pada gambar 3, dapat dianalisis, untuk ekstraksi fitur Bag of Words dapat memprediksi 136 data dengan benar dan 8 data yang diprediksi salah, sedangkan pada ekstraksi fitur TF-IDF memprediksi sebanyak 135 data diprediksi benar dan 9 data diprediksi salah. Terdapat perbedaan yang tidak terlalu signifikan dari hasil prediksi kedua ekstraksi fitur.

Tabel 8. Hasil Performa Ekstraksi Fitur

BoW					**TF-IDF**			
	Precision	**Recall**	**F1-Score**	**Acc**	**Precision**	**Recall**	**F1-Score**	**Acc**
SMS Normal	97,50%	90,70%	93,98%		97,50%	90,70%	93,98%	
SMS Spam	90%	97,78%	93,75%	94,44%	89,80%	95,65%	92,63%	93,75%
SMS Promo	96,30%	94,55%	95,41%		94,55%	94,55%	94,55%	

Berdasarkan tabel 8 di atas, hasil akurasi yang diperoleh menggunakan ekstraksi fitur BoW lebih besar dibandingkan dengan menggunakan TF-IDF, yakni 94,44%. Nilai precision, recall dan F1-Score untuk sms normal didapatkan dengan menggunakan BoW dengan nilai 97,50%, 90,70%, 93,98%, Hasil akurasi untuk precision, recall, dan F1-Score untuk sms spam didapatkan dengan menggunakan BoW dengan nilai 90%, 97,78%, 93,75%. Untuk sms promo didapat nilai precision, recall, dan F1-Score secara berturut-turut 96,30%, 94,55%, dan 95,41%.

4. Kesimpulan

Berisi pernyataan yang menjawab persoalan pada bagian sebelumnya dan karya penelitian yang akan datang [5]. Berdasarkan penelitian yang sudah dilakukan, maka peneliti dapat menarik kesimpulan yang diperoleh, di antara kedua perbandingan ekstraksi fitur, yakni Bag of Words dan TF-IDF yang menghasilkan akurasi dengan performa yang baik adalah ekstraksi fitur Bag of Words dengan nilai akurasi 94,44% sedangkan akurasi yang dihasilkan oleh TF-IDF adalah 93,75%. Perbedaan nilai akurasi antara Bag of Words dan TF-IDF tidak terlalu signifikan, mengingat pada penelitian yang telah dilakukan penulis melakukan penyeimbangan data di setiap label menggunakan teknik random undersampling. Peningkatan performa akurasi dapat dilakukan dengan beberapa cara, seperti menambah data pada setiap label dengan syarat data harus seimbang dan melakukan seleksi fitur.

Daftar Pustaka

[1] L. Mamoto, "Peranan Hukum Pidana dalam Menanggulangi Penipuan Lewat SMS Serta Penegakan Hukumnya," Lex Crimen, vol. 5, no. 7, 2016.
[2] Al Amien, J., Mukhtar, H., & Rucyat, M. A. (2020). Jurnal Computer Science and Information Technology (CoSciTech).
[3] Aksenta, Almasari, et al. Literasi Digital: Pengetahuan & Transformasi Terkini Teknologi Digital Era Industri 4.0 dan Sociaty 5.0. PT. Sonpedia Publishing Indonesia, 2023.
[4] N. U. Maulidevi and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 6, pp. 3413-3423, 2022.
[5] F. Alzami et al., "Document preprocessing with TF-IDF to improve the polarity classification performance of unstructured sentiment analysis," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, pp. 235-242, 2020.
[6] W. T. H. Putri and R. Hendrowati, "Penggalian Teks Dengan Model Bag of Words Terhadap Data Twitter," Jurnal Muara Sains, Teknologi, Kedokteran, dan Ilmu Kesehatan, vol. 2, no. 1, pp. 129-138, 2018.
[7] B. Harjito, K. N. Aini, and B. Murtiyasa, "Klasifikasi dokumen berkonten serangan jaringan menggunakan multinomial naive bayes," in Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASTIK), vol. 1, no. 1, pp. 112-118, Oct. 2018.
[8] N. Rezaeian and G. Novikova, "Persian text classification using naive bayes algorithms

and support vector machine algorithms," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 8, no. 1, pp. 178-188, 2020.