Klasifikasi *Short Message Service Spam* Menggunakan Algoritma *Naïve Bayes Classifier*

Jannio Alvares[1], Uyok Angoro Saputro*[2]
[1,2]Fakultas Ilmu Komputer, Program Studi Informatika, Universitas Amikom Yogyakarta
E-mail: [1]jannio.11@students.amikom.ac.id, *[2]**uyock@amikom.ac.id**

Abstrak

Short Message Service atau yang lebih dikenal dengan SMS merupakan salah satu komunikasi dengan media teks melalui perangkat telepon seluler atau smartphone. Saat ini perkembangan teknologi informasi memungkinkan pengaksesan data lebih praktis, cepat, dan efisien, oleh karena itu dapat dimanfaatkan sebagian orang untuk melakukan spam untuk melakukan promosi terhadap suatu toko atau jasa. Spam adalah singkatan dari Sending and Posting Advertisement in Mass, yang artinya mengirim pesan secara massal. Selain untuk promosi ada juga yang menyalahgunakannya untuk melakukan tindakan kejahatan seperti penipuan berhadiah. Sebagian orang mungkin dapat mengenali apakah SMS tersebut berupa penipuan atau tidak, namun bagi orang awam yang baru saja mendapat SMS seperti itu tentu kebingungan dan mungkin akan tergiur. Oleh karena itu, tujuan penelitian ini adalah untuk mengklasifikasi jenis-jenis SMS apakah termasuk pesan normal, penipuan, atau promo. Metode yang digunakan dalam penelitian ini adalah Naïve Bayes Classifier. Sebelum data digunakan dilakukan proses preprocessing terlebih dahulu, kemudian dilakukan proses feature extraction menggunakan TF-IDF. Data yang sudah diproses akan dibagi menjadi dua bagian dan akan dilakukan proses cross validation untuk menghasilkan kinerja yang baik. Berdasarkan hasil yang diperoleh, Multinomial Naïve Bayes yang menggunakan proses cross validation mendapat akurasi mencapai 97,7%.

Kata Kunci*— Klasifikasi, SMS, Spam, Naive Bayes, TF-IDF*

1. PENDAHULUAN

Saat ini teknologi informasi dan komunikasi semakin berkembang, SMS atau Short Message Service merupakan hal yang sudah dikenal oleh masyarakat di dunia. Dengan SMS, komunikasi antara ponsel dapat dilakukan di mana saja, kapan saja dan oleh siapa saja. Karena sifatnya yang umum, maka SMS sering kali digunakan sebagai perantara untuk menyebarkan spam sesuai dengan kepentingan masing-masing individu.

Secara definisi SMS spam adalah pesan teks yang tidak diinginkan atau tidak diminta yang dikirim secara berturut-turut. Namun pada saat ini SMS spam secara umum dapat dikatakan sebagai pesan teks yang tidak diinginkan karena biasanya hanya dikirim sebanyak satu kali ke setiap nomor telepon. SMS spam yang dikirim dapat berupa iklan atau promosi ataupun sebuah pesan penipuan yang mengatakan anda mendapatkan hadiah dengan mengatasnamakan perusahaan atau lembaga tertentu. Beberapa orang mungkin akan tergiur dengan hal tersebut sehingga mencoba menghubungi spammer dan akhirnya tertipu. Sebagian orang yang tertipu mungkin merasa rugi karena telah mengeluarkan uang yang ditransfer ke spammer sebagai syarat pengiriman hadiah.

Di Indonesia banyak terjadi kasus penipuan lewat SMS, salah satunya dikutip dari Dektiknews pada 2 Maret 2021, dua pelaku penipuan via SMS yang meraup 200 juta sebulan dibekuk polisi. Kedua pelaku tersebut menipu dengan modus undian dengan mengirimkan pesan secara acak bahwa penerima pesan memenangkan undian, kemudian akan dijelaskan oleh pelaku bagaimana cara mendapatkan hadiahnya yakni dengan mentransfer sejumlah uang sebagai

persyaratan. Selain itu pada tahun 2019 sebuah aplikasi mobile yang bernama Truecaller merilis daftar negara dengan SMS spam terbanyak dan Indonesia termasuk di dalamnya pada urutan ke 10. Untuk mengatasi masalah tersebut dapat digunakan salah satu cara yaitu dengan menggunakan teknik Machine Learning yang dapat mengklasifikasi SMS apakah itu SMS normal, SMS promo ataupun SMS penipuan. Selain menggunakan Machine Learning, juga dibutuhkan teknik Text Mining untuk memproses data sebelum dilakukannya proses machine learning agar data menjadi efisien.

Untuk melakukan klasifikasi dalam penelitian ini, algoritma yang digunakan adalah Naïve Bayes Classifier. Naive Bayes merupakan suatu metode Machine Learning yang populer dan sering digunakan untuk melakukan klasifikasi, selain itu algoritma ini mudah diimplementasikan, memiliki kinerja yang baik, dan prosesnya cepat [1]. Kemudian dari segi akurasi tidak jauh berbeda dari algoritma lain seperti SVM (Support Vector Macine) dan Logistic Regression [2], oleh karena itulah algoritma Naïve Bayes digunakan pada penelitian ini.

1.1. Short Message Service (SMS)

Short Message Service (SMS) adalah layanan dasar yang memungkinkan pertukaran pesan teks singkat antar pelanggan. 2 Pesan teks pendek pertama diyakini telah dikirim pada tahun 1992 melalui saluran sinyal jaringan GSM Eropa. Sejak percobaan yang sukses ini, penggunaan SMS telah menjadi subyek pertumbuhan yang luar biasa [3]. SMS di standarisasi oleh badan Bernama ETSI (European Telecommunication Standards Institute), fitur SMS ini dapat mengirimkan sampai dengan 160 karakter yang dapat berupa alfabet dan angka melalui jaringan GSM hanya dalam hitungan detik. SMS sebenarnya dirancang sebagai bagian dari Global System for Mobile communications (GSM), tetapi sekarang tersedia di berbagai standar jaringan seperti Code Division Multiple Access (CDMA) [4].

1.2. Spam

Spam sebenarnya merupakan singkatan dari Sending and Posting Advertisement in Mass, yaitu pengiriman pesan promosi secara massal. Tujuan spam awalnya sama seperti pengertiannya yaitu untuk melakukan promosi atau iklan, namun saat ini spam banyak disalahgunakan untuk kepentingan pribadi yang dapat merugikan orang lain misalnya penipuan atau phishing, dan ada juga yang menggunakannya untuk menyebar malware atau virus. Spam SMS memiliki efek yang lebih besar pada pengguna daripada spam email karena pengguna melihat setiap SMS yang mereka terima, sehingga spam SMS mempengaruhi pengguna secara langsung [4].

1.3. Text Mining

Text Mining adalah proses mencari atau mengekstraksi informasi yang berguna dari data teks. Text Mining juga dikenal sebagai Text Data Mining karena proses Text Mining sama dengan Data Mining, namun ada pembedanya yaitu Data Mining dirancang untuk menangani data terstruktur sedangkan Text Mining dapat menangani kumpulan data tidak terstruktur atau semi-terstruktur seperti email, file HTML, full text document, dan lain-lain [5].

1.4. Natural Language Processing

Natural Language Processing (NLP) merupakan salah satu bidang dalam Text Mining seperti information retrieval, information extraction, dan data mining [6]. Selain itu Natural Language Processing adalah cabang ilmu Artificial Intelligence atau kecerdasan buatan dan linguistik yang ditujukan untuk membuat komputer memahami pernyataan atau kata-kata yang ditulis dalam bahasa manusia, bahasa yang dimaksud adalah bahasa biasa yang diucapkan atau

ditulis oleh orang atau manusia untuk komunikasi umum [7]. Bahasa manusia mempunyai struktur dan tata bahasa yang berbedabeda, sehingga sebelum diproses biasanya perlu dilakukan preprocessing. Preprocessing bertujuan untuk mengurangi jumlah data, menemukan hubungan antar data, menormalisasi data, dan mengekstrak fitur untuk data [8].

1.5. TF-IDF

Term Frequency- Inverse Document Frequency adalah gabungan dari dua metode yaitu TF (Term Frequency) dan IDF (Inverse Document Frequency). Metode ini merupakan metode dalam statistik yang digunakan untuk menilai pentingnya sebuah kata untuk dokumen. Ide pokoknya adalah jika suatu kata atau frasa sering muncul dalam suatu dokumen dan jarang ditemukan di dokumen lain, maka kata atau frasa tersebut dianggap memiliki kemampuan pembedaan kelas yang baik dan layak untuk diklasifikasi [9].

1.6. Naïve Bayes Classifier

Naïve Bayes Classifier adalah salah satu algoritma yang dapat digunakan dalam melakukan klasifikasi dokumen yang memanfaatkan metode probabilitas dan statistik, yaitu memprediksi probabilitas di masa depan berdasarkan masa sebelumnya. Algoritma ini mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas [10]. Naïve bayes juga termasuk dalam supervised learning di mana setiap data harus mempunyai label agar dapat dilakukan pelatihan. Persamaan dasar dari algoritma ini dapat dilihat pada Persamaan (1) berikut ini

$$\mathrm{P(A|B)} = \frac{P\ (A|B).P(A)}{P(B)} \quad (1)$$

Keterangan :

A	: Hipotesis data merupakan suatu class spesifik
B	: Data dengan class yang belum diketahui
P(A|B)	: Probabilitas hipotesis A berdasar kondisi B
P(A)	: Probabilitas hipotesis A
P(B|A)	: Probabilitas B berdasarkan kondisi pada hipotesis A
P(B)	: Probabilitas B

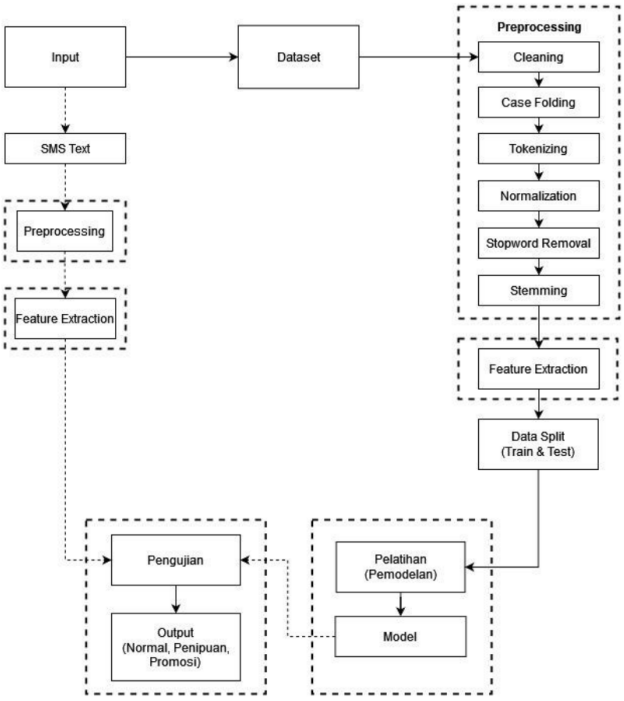
1.7. Cross Validation

Cross Validation adalah teknik yang digunakan untuk menilai bagaimana kinerja dari pengklasifikasi ketika akan mengklasifikasikan data baru. Satu iterasi cross validation melibatkan partisi sampel data menjadi dua himpunan bagian yang saling melengkapi yaitu data train dan data test [11]. Metode ini digunakan karena dapat mengurangi waktu komputasi dan tetap menjaga akurasi yang diperoleh.

2. METODE PENELITIAN

2.1. Gambaran Umum Penelitian

Gambaran umum penelitian tersaji pada gambar 1 sebagai berikut.



Gambar 1. Gambaran Umum Penelitian

2.2. Dataset

Dalam penelitian ini digunakan dataset SMS spam berbahasa Indonesia. Dataset ini berisi 1143 SMS yang dikategorikan secara manual ke dalam tiga kategori menggunakan angka. Tiga kategori tersebut yaitu normal, penipuan, dan promo. Rincian dataset dapat dilihat pada Tabel I.

Tabel 1 Dataset

Kategori	**Jumlah**
Normal	569
Promo	335
Penipuan	239
Jumlah Total	1143

2.3. Preprocessing

Prepocessing adalah tahapan penting pada proses data mining. Proses ini berfungsi untuk membersihkan dan menyiapkan dataset sehingga ideal untuk digunakan dan tidak mengganggu hasil proses. Pada tahapan ini ada proses yang terjadi yaitu cleaning, case folding, normalization, stopword removal, dan stemming.

2.4. Feature Extraction

Feature extraction mengacu pada proses mengubah data mentah menjadi data numerik yang dapat diproses sambil mempertahankan informasi yang terdapat dalam kumpulan data asli.. Tujuan dari tahapan ini yaitu memperkecil jumlah data, menaikkan presisi pengolahan, dan mengambil informasi terpenting dari data yang diolah. Salah satu metode dalam feature extraction

adalah TF-IDF (Term Frequency-Inverse document frequency). TF-IDF dapat menyimpan informasi dari kata-kata dalam teks dan merepresentasikan pentingnya tiap kata dengan kombinasi TF dan IDF. Term Frequency adalah menentukan bobot pada suatu kata berdasarkan jumlah kemunculannya. Nilai jumlah kemunculan diperhitungkan dalam pemberian bobot terhadap suatu kata, semakin besar jumlah kemunculannya semakin besar pula bobotnya. Kemudian Document Frequency adalah menentukan bobot dalam suatu dokumen berdasarkan jumlah kemunculannya. Untuk mendapatkan hasil pembobotan, dilakukan perkalian antara TF dan IDF. Sebelum melakukan perhitungan TF-IDF, dokumen harus melalui tahap preprocessing terlebih dahulu.

2.5. Pemodelan Naïve Bayes

Data train yang telah melalui proses preprocessing dan feature extraction akan dilanjut menuju proses pelatihan data dengan menerapkan algoritma Naïve Bayes Classifier. Pada proses ini dilakukan klasifikasi data ke dalam kelas yang telah ditentukan sebelumnya yaitu SMS normal, SMS penipuan, dan SMS promo. Ketika proses klasifikasi selesai maka akan menghasilkan model disertai nilai akurasinya, model ini nantinya akan digunakan untuk proses pengujian. Proses ini juga menggunakan metode k-fold cross validation untuk membandingkan kinerja atau performa dari metode tanpa k-fold dan metode yang menggunakan k-fold. Pada proses pengujian data yang diinput juga akan melalui tahap preprocessing dan feature extraction kemudian akan dilakukan prediksi untuk menentukan kategori atau kelas dari SMS yang telah diinput sebelumnya.

2.6. Pengukuran Performa

Pengukuran performa dilakukan dengan menghitung akurasi, precision, recall, dan f1-score berdasarkan confusion matrix yang telah dibuat. Confusion matrix merupakan salah satu metode yang digunakan untuk mengevaluasi performa dari suatu model berdasarkan tabel matriks yang terbentuk.

3. HASIL DAN PEMBAHASAN

3.1. Impelemntasi Preprocessing

Tahap preprocessing dilakukan untuk mempersiapkan data sehingga kualitas data lebih baik sebelum digunakan untuk tahap selanjutnya. Data yang akan diproses berjumlah 1143 mencakup 569 SMS normal, 239 SMS penipuan, dan 335 SMS promo. Ada beberapa tahapan yang akan dilakukan dalam proses ini yaitu cleaning, case folding, tokenizing, normalization, stopword removal, dan stemming.

3.2. Implementasi Feature Extraction

Pada tahap features extraction, data hasil preprocessing ditransformasikan ke dalam bentuk angka dengan menggunakan metode TF-IDF. Tahap ini akan menghasilkan sebuah matriks yang akan digunakan dalam melakukan klasifikasi SMS spam. Data hasil preprocessing yang berupa kumpulan token digabungkan kembali menjadi sebuah kalimat utuh, kemudian dilakukan features extraction. Proses feature extraction dilakukan dengan metode TF-IDF menggunakan modul TfidfVectorizer yang terdapat dalam library sklearn.features_extraction.text. Sebelum feature extraction dilakukan, proses preprocessing dijalankan dan disimpan dalam bentuk kata yang utuh atau tidak dalam kumpulan token atau kata lagi. Hal ini dilakukan untuk mempermudah proses feature extraction karena dalam proses itu

sendiri kalimat akan dipecah menjadi kumpulan token. Hasil dari proses ini berupa matriks dari kumpulan kata yang sudah diubah menjadi angka melalui perhitungan TF-IDF.

3.3. Implementasi Pembagian Data

Pada tahap ini dilakukan proses pembagian data menjadi 2 yaitu data train dan data test. Proses pembagian data dilakukan dengan perbandingan 80% train dan 20% test. Pembagian data dilakukan menggunakan function train_test_split dari library sklearn.model_selection. Pembagian data tidak dilakukan secara acak dan dengan persentase di mana data test berjumlah 20% dari total data, dan data train berjumlah 80% dari total data.

3.4. Implementasi Naïve Bayes

Pada tahap ini klasifikasi dilakukan dengan menggunakan model Naïve Bayes Classifier. Data yang digunakan untuk melatih model pada tahap ini adalah data train. Proses pelatihan model menggunakan metodde cross validation sehingga data train akan dipecah lagi untuk memperoleh data validasi. Pembagian secara default tidak dilakukan pengacakan, ini berfungsi untuk menghasilkan keluaran data yang sama pada setiap sesi program sehingga hasilnya akan konsisten. Metode cross validation diimplementasikan menggunakan modul KFold dari library sklearn.model_selection. Sedangkan untuk model Naïve Bayes Classifier, menggunakan modul MultinomialNB dari library sklearn.naive_bayes.

3.5. Implementasi Confusion Matrix

Confusion matrix dibuat berdasarkan hasil prediksi pada proses pelatihan atau pemodelan. dengan menggunakan function confusion_matrix dari library sklearn.metrics. Pada tahap ini dilakukan pembuatan dua confusion matrix, yang pertama dari data hasil predict X_test (hasil pembagian data) dan yang kedua dari data hasil predict x_test dari hasil k-fold cross validation. Kedua matriks ini akan disimpan dalam bentuk gambar (.png) dan akan ditampilkan pada sistem.

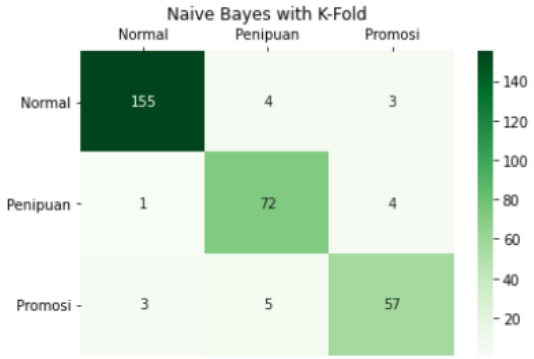
3.6. Pengujian

Berikut adalah tabel hasil pengujian klasifikasi SMS spam menggunakan Multinomial Naïve Bayes dan cross validation yang terdapat pada tabel 2.

Tabel 2 Tabel hasil pengujian klasifikasi SMS

K-Fold	Akurasi	F1-Score
Tanpa k-fold	0,904	0,890
3	0,977	0,972
5	0,967	0,959
7	0,969	0,956
10	0,967	0,952

Hasil pengujian dari Multinomial Naïve Bayes menunjukkan hasil yang baik dengan menggunakan cross validation di mana rata-rata akurasi mencapai 97% dan f1-score di atas 95%. Hasil tertinggi didapat pada cross validation dengan nilai k=3 di mana hasil akurasinya adalah 97,7%. Dari hasil tersebut diperoleh confusion matrix seperti pada gambar 5 sebagai berikut.



Gambar 5 Confusion Matrix Multinominal Fold 3

Selain itu dilakukan juga pengujian untuk Bernoulli Naïve Bayes dan Gaussian Naïve Bayes apakah dapat mengklasifikasikan data multiclass atau yang bersifat kontinu, berikut adalah hasil pengujiannya dengan menggunakan cross validation dengan nilai k=3.

Tabel 3 Klasifikasi Bernoulli dan Gaussian

K-Fold	Akurasi	F1-Score
Bernoulli	0,964	0,957
Gaussian	0,990	0,989

Dari tabel 3 di atas Bernoulli dan Gaussian berhasil mengklasifikasikan data SMS spam yang memiliki lebih dari dua kelas dan tidak bersifat kontinu dengan hasil akurasi yang baik. Untuk menguji apakah hasil tersebut hanya berupa angka, selanjutnya akan dilakukan pengujian prediksi menggunakan model dari ketiga tipe Naïve Bayes Classfier tersebut. Data yang akan digunakan untuk melakukan pengujian tersebut tersaji pada tabel 4 sebagai berikut.

Tabel 4 Data pengujian

Data	SMS
D1	Ada PROMO di bln JUNI! Isi PULSA & bli Paket INTERNET OMG di myTELKOMSEL & byr pakai Shopeepay/Gopay/OVO dpt cashback hgg 40%. Info: tsel.me/cashbacknampol D2
D2	kirim BANK BNI ini aja no rekeningnya 123-293-4632 a/n VIKTOR
D3	Spesial untukmu! Beli paket Rp13Rb/7hr dpt - Kuota 4G 3GB - Telp 120min&200SMS Tsel Balas WOW13 ke SMS ini atau*363*20# & tsel.me/HO Promo sd 22 Agt
D4	DISKON 50% Biaya Pasang Baru INDIHOME up to 100 Mbps Rp250Rb + CASHBACK LinkAja Rp150Rb, Bonus (Earphones + Finger Sleeves/Sepatu League/Sarung Sobat Gurun)* myads.id/45m8f
D5	Bsk kelas kosong kan ?
D6	Gak punya Kuota tapi pengen beli Diamond? Ayo top up Diamond Gamesmu di *900#! Mudah topup nya,gampang prosesnya,pilihan games beragam! Lgsg aja ke *900#. SKB
D7	Slmt 4nda dpt 125jt dri 5_h_o_p_3_3 K0d3 4nda (25477BLW) Kl1k. s.id/layanan-program56 WA 082211485887
D8	Pernah nyobain Kuota Yang Dicap Dua Jempol ngga? 50 GB 30 hari 120 ribu, auto rebahan sepanjang hari! Klik di sini: go.byu.id/belikuota50GB
D9	Ass!!!...Kami Dri Aplikasi Tik Tok Slmt Anda Mendptkan Udian 75.jt Dri ID anda (TIK21K77) Info lbhi lengkap chat WA:082268136279
D10	kemaren ada penipu ketangkep

Untuk hasil pengujian tersaji pada tabel 5 sebagai berikut.

Tabel 5 Hasil pengujian

Data	Label	Multonimial	Bernoulli	Gaussion
D1	2	2	2	2
D2	1	1	1	1
D3	2	2	2	2
D4	2	2	2	1
D5	0	0	0	0
D6	2	2	2	2
D7	1	1	0	0
D8	2	2	2	0
D9	1	1	1	0
D10	0	0	0	1
Total Benar		10	9	5

Berdasarkan hasil pengujian pada tabel 5, dapat disimpulkan bahwa Multinomial dapat memprediksi data uji dengan benar semua, sedangkan Bernoulli dapat memprediksi dengan benar 9 dari 10 data, dan Gaussian hanya dapat memprediksi setengah dari jumlah data yaitu hanya 5. Dengan demikian angka akurasi yang tinggi pada model Gaussian hanya angka saja tetapi tidak dengan keakuratan dalam memprediksi data dengan benar. Berdasarkan hasil dari pengujian sebelumnya, maka model yang akan digunakan pada sistem adalah Multinomial Naïve Bayes dengan nilai k=3.

4. KESIMPULAN

Klasifikasi dilakukan dengan algoritma Naïve Bayes Classifier di mana hasil yang diperoleh menunjukkan bahwa Multinomial Naïve Bayes dengan nilai k-fold 3 mendapatkan nilai akurasi yang tinggi dan dapat melakukan pengujian prediksi dengan benar. Hasil pengujian juga menunjukkan bahwa Bernoulli Naïve Bayes dan Gaussian Naïve Bayes dapat melakukan klasifikasi terhadap data multiclass dan tidak bersifat kontinu, meskipun tidak sepenuhnya dapat memprediksi atau mengklasifikasikan semua data uji dengan benar. Ini mungkin terjadi karena menggunakan library yang sama yaitu sklearn.naive_bayes di mana variabel inputnya sama yaitu X dan Y namun berbeda perhitungan

DAFTAR PUSTAKA

[1] Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research (IJSR), 5(1), 1842–1845. https://doi.org/10.21275/v5i1.nov153131

[2] Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. International Journal of Recent Contributions from Engineering, Science & IT (IJES), 7(2), 91.

[3] Le Bodic, G. (2005). Mobile Messaging Technologies and Services: SMS, EMS and MMS: Second Edition. In Mobile Messaging Technologies and Services: SMS, EMS and MMS: Second Edition.

[4] Mahmoud, T. M., Mahfouz, A. M., & Minia, E. (2012). SMS Spam Filtering Technique Based on Artificial Immune System. International Journal of Computer Science Issues, 9(2), 589–597.
[5] Vijayarani, S., Ilamathi, M. J., Nithya, M., Professor, A., & Research Scholar, M. P. (n.d.). (2015). Preprocessing Techniques for Text Mining -An Overview. 5(1), 7–16.
[6] L.Sumathy, K., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues An Overview. International Journal of Computer Applications, 80(4), 29–32. https://doi.org/10.5120/13851- 1685
[7] Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. International Journal of Technology Enhancements and Emerging Engineering Research, 1, 131-134.
[8] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102–4107. https://doi.org/10.3923/jeasci.2017.4102.4107
[9] C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), 2018, pp. 218-222.
[10] Patil, T. R., Sherekar, M. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88–91.
[11] Moreno-Torres, J. G., Saez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1304–1312