

USULAN TUGAS AKHIR

**DETEKSI SMS SPAM BERBAHASA INDONESIA
MENGUNAKAN METODE PEMBOBOTAN FITUR
TF-RF DAN SUPPORT VECTOR MACHINE
CLASSIFIER**



Oleh:

**Muhammad Syulhan Al Ghofany
F1D017064**

PROGRAM STUDI TEKNIK INFORMATIKA

**FAKULTAS TEKNIK
UNIVERSITAS MATARAM
2021**

HALAMAN PENGESAHAN

Bagian ini ditimpa dengan lembar pengesahan yang dihasilkan dari system
<https://ta.if.unram.ac.id/>

TUGAS AKHIR

DETEKSI SMS SPAM BERBAHASA INDONESIA MENGGUNAKAN METODE PEMBOBOTAN FITUR TF-RF DAN SUPPORT VECTOR MACHINE CLASSIFIER

1. Pembimbing Utama


Tanggal: 2021



Ramaditia Dwiyanaputra, S.T., M.Eng.
NIP. -

2. Pembimbing Pendamping

Tanggal: 2021



Fitri Bimantoro, S.T., M.Kom.
NIP. 19860622 201504 1 002

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Teknik
Universitas Mataram

Prof. Dr.Eng. I GP Suta Wijaya, ST., MT.
NIP. 19731130 200003 1 001

ABSTRAK

SMS Spam adalah pesan teks yang tidak diminta atau tidak diinginkan oleh pengguna yang dikirim ke perangkat seluler. Pada saat ini, semakin banyak terjadi dan luasnya tindakan kejahatan yang dapat mengganggu penerimanya dengan menyebarkan SMS spam yang tidak diminta atau tidak diinginkan, diantaranya promosi, penipuan, pesan porno, dan lain sebagainya. Oleh karena itu, klasifikasi pada SMS perlu dikembangkan agar dapat membantu dalam pengkategorian SMS. Pada penelitian yang telah ada untuk mencoba mengatasi permasalahan tersebut diterapkan fitur term frequency inverse document frequency (TF-IDF), namun metode ini memiliki kekurangan yaitu menghilangkan informasi kategori pada tiap dokumen, sehingga pada penelitian ini akan dilakukan perbandingan dengan metode fitur Supervised Term Weighting (STW), yaitu salah satunya term frequency relevance frequency (TF-RF) dengan menggunakan Support Vector Machine, k-Nearest Neighbour, dan Multinomial Naïve Bayes. Pada penelitian ini, total data yang digunakan adalah 500 SMS dengan perbandingan 325 SMS non-spam dan 175 SMS spam. Hasil yang didapatkan pada penelitian ini, model SVM Kernel Sigmoid memiliki nilai rata-rata presisi, recall, dan accuracy tertinggi dibandingkan dengan SVM Kernel Linear dan RBF, k-Nearest Neighbour, dan Multinomial Naïve Bayes.

Kata kunci – SMS Spam, Klasifikasi Teks, Supervised Term Weighting, TF-IDF, TF-RF, Support Vector Machine

DAFTAR ISI

HALAMAN PENGESAHAN.....	ii
TUGAS AKHIR.....	iii
ABSTRAK.....	iv
DAFTAR ISI.....	v
DAFTAR GAMBAR.....	vi
DAFTAR TABEL.....	vii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	4
1.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Penelitian Terkait.....	5
2.2 Teori Penunjang.....	10
2.2.1 SMS Spam.....	10
2.2.2 Text Mining.....	10
2.2.3 Text Preprocessing.....	11
2.2.4 Klasifikasi Teks.....	12
2.2.5 Term Weighting.....	12
2.2.6 Support Vector Machine Classifier.....	14
2.2.7 Evaluasi.....	17
BAB III METODOLOGI PENELITIAN.....	19
3.1 Alat dan Bahan.....	19
3.2 Alur Penelitian.....	19
3.3 Perancangan Sistem.....	21
3.4 Cara Analisis.....	21
3.4.1 Input Training dan Testing SMS.....	22
3.4.2 Text Preprocessing.....	23
3.4.3 Pattern Discovery.....	27
3.4.4 Klasifikasi Support Vector Machine.....	31
3.5 Pengujian.....	34
BAB IV HASIL DAN PEMBAHASAN.....	38
4.1 Pengumpulan Data.....	38
4.2 Text Preprocessing.....	38
4.3 Pattern Discovery.....	40
4.4 Pengujian.....	41
4.5 Hasil Pengujian.....	42
BAB V KESIMPULAN DAN SARAN.....	54
5.1 Kesimpulan.....	54
5.2 Saran.....	54
DAFTAR PUSTAKA.....	56

DAFTAR GAMBAR

Gambar 2. 1 Support Vector Machine	16
Gambar 3. 1 Diagram alur pembuatan sistem	20
Gambar 3. 2 Rancangan sistem	21
Gambar 3. 3 Dataset dalam diagram kartesius	32
Gambar 3. 4 Contoh dataset dengan hyperplane	34
Gambar 3. 5 Ilustrasi cross validation 10 fold	37
Gambar 4. 1 Skema proses preprocessing	38
Gambar 4. 2 Proses case folding dataset SMS	39
Gambar 4. 3 Proses tokenizing dataset SMS	39
Gambar 4. 4 Proses stopword filtering dataset SMS	39
Gambar 4. 5 Proses stemming dataset SMS	40
Gambar 4. 6 Skema proses pattern discovery	40
Gambar 4. 7 Diagram perbandingan nilai precision	50
Gambar 4. 8 Diagram perbandingan nilai recall	51
Gambar 4. 9 Diagram perbandingan nilai accuracy	51

DAFTAR TABEL

Tabel 2. 1 State of art	8
Tabel 2. 2 Confusion Matrix	17
Tabel 3. 1 Contoh training dan testing SMS	22
Tabel 3. 2 Contoh case folding SMS	23
Tabel 3. 3 Contoh tokenizing SMS	24
Tabel 3. 4 Contoh stopword filtering SMS	25
Tabel 3. 5 Contoh stemming SMS	26
Tabel 3. 6 Nilai TF	28
Tabel 3. 7 Nilai DF dan IDF pada beberapa kata	29
Tabel 3. 8 Nilai TF-IDF pada beberapa kata	30
Tabel 3. 9 Nilai RF pada beberapa kata	31
Tabel 3. 10 Nilai TF-RF pada beberapa kata	31
Tabel 3. 11 Contoh nilai fitur dengan 2 kelas	32
Tabel 3. 12 Confusion Matrix untuk 2 kelas	35
Tabel 3. 13 Jadwal kegiatan perancangan sistem... Error! Bookmark not defined.	
Tabel 4. 1 Pengujian dengan SVM Kernel RBF yang di-stemming	43
Tabel 4. 2 Pengujian dengan SVM Kernel RBF tanpa stemming	43
Tabel 4. 3 Pengujian dengan SVM Kernel Linear yang di-stemming	44
Tabel 4. 4 Pengujian dengan SVM Kernel Linear tanpa stemming	45
Tabel 4. 5 Pengujian dengan SVM Kernel Sigmoid yang di-stemming	46
Tabel 4. 6 Pengujian dengan SVM Kernel Sigmoid tanpa stemming	46
Tabel 4. 7 Pengujian dengan k-Nearest Neighbour yang di-stemming	47
Tabel 4. 8 Pengujian dengan k-Nearest Neighbour tanpa stemming	48
Tabel 4. 9 Pengujian dengan Multinomial Naïve Bayes yang di-stemming	49
Tabel 4. 10 Pengujian dengan Multinomial Naïve Bayes tanpa stemming	49

BAB I

PENDAHULUAN

1.1 Latar Belakang

Semakin luasnya pengguna SMS pada masyarakat banyak dimanfaatkan dengan disalahgunakan oleh pihak yang tidak bertanggung jawab untuk melakukan tindak kejahatan dengan menyebarkan SMS spam yang tidak diminta dan tidak diinginkan, seperti promosi, penipuan, pesan porno, dan lain sebagainya. SMS *spam* adalah pesan teks yang tidak diinginkan atau tidak diminta dan tidak diketahui siapa pengirimnya [1]. Biasanya, pesan tersebut berisi penawaran sesuatu, atau bahkan bentuk dari modus penipuan. Menurut FCC, mengirim pesan komersial tanpa izin ke perangkat nirkabel adalah melanggar hukum, termasuk ponsel dan pager, kecuali jika pengirim mendapatkan izin Anda terlebih dahulu [2].

Terdapat beberapa upaya yang dilakukan oleh pemerintah agar masalah ini dapat teratasi, seperti revisi Peraturan Menteri Koinfo No.1/2009 yang mengatur tentang jasa pesan premium akan dipertegas aturan terkait SMS advertising. Akan tetapi SMS advertising dengan SMS spam tidak jauh berbeda karena keduanya sama-sama mengirimkan pesan singkat ke banyak pengguna, walaupun yang membedakan adalah SMS advertising harus sesuai dengan persetujuan pengguna, walaupun begitu masih meresahkan karena pesan yang dikirimkan masih tidak terkontrol dari segi jumlah dan waktu pengiriman [3]. Upaya yang dilakukan juga pelaksanaan registrasi kartu sim prabayar yang dilanjutkan dengan adanya aturan validasi IMEI, akan tetapi SMS spam ini masih terus bermunculan tidak terkendali [4].

Guna mengatasi masalah tersebut, maka teknik klasifikasi akan diterapkan pada teks SMS pesan spam untuk membedakan pesan yang berisi spam dan pesan yang tidak berisi spam. Klasifikasi merupakan proses untuk menemukan sekumpulan model atau fungsi yang mendeskripsikan dan membedakan kelas-kelas data dengan tujuan untuk memprediksikan kelas dari objek yang belum diketahui kelasnya (supervised learning) dengan karakteristik tipe data bersifat katagorik. Metode yang dapat digunakan dalam klasifikasi teks adalah metode Support Vector

Machine (SVM). Metode klasifikasi SVM adalah salah satu metode diskriminatif yang paling tepat yang digunakan dalam klasifikasi. Metode SVM terbukti dapat memberikan nilai akurasi yang tinggi dan stabil [5].

Penelitian terkait dengan klasifikasi SMS spam ini telah cukup banyak dilakukan, namun sebagian besar dari penelitian tersebut menggunakan metode pembobotan fitur tradisional atau tidak terawasi (unsupervised term weighting) yang saat ini paling populer yaitu Term Frequency Inverse Document Frequency (TF-IDF) [6]–[9]. Term Frequency adalah kemunculan frekuensi munculnya kata sama pada dokumen. Inverse Document Frequency banyaknya koleksi dokumen yang bersangkutan mengandung kata tertentu. TF-IDF memberikan bobot tinggi pada term yang jarang muncul pada seluruh dokumen. TF-IDF ini memiliki kekurangan yaitu menghilangkan informasi kategori pada tiap dokumen, TF-IDF hanya bergantung pada frekuensi term dalam dokumen dan jumlah (kebalikan) dari dokumen pelatihan yang memuat istilah ini.

Selain adanya penelitian dengan metode tradisional, juga terdapat metode Supervised term weighting (STW), atau pembobotan fitur yang terawasi dimana metode ini memanfaatkan informasi yang diketahui tentang keanggotaan dokumen pelatihan ke dalam kategori, sehingga term yang sangat diskriminatif pada kategori tertentu sangat membantu kemunculannya pada proses pengkategorian. Salah satu metode modern yang ada adalah Term Frequency Relevance Frequency atau TF-RF. Dibandingkan dengan metode tradisional yang hanya didasarkan pada distribusi istilah/kata di seluruh dokumen atau lebih menyukai istilah/kata yang jarang, metode TF-RF adalah salah satu metode yang memperhatikan istilah/kata yang sering muncul pada tiap dokumen di masing-masing kategori. Dalam beberapa penelitian yang telah dilakukan sebelumnya, metode TF-RF memiliki performa yang cukup baik untuk klasifikasi teks dan bahkan lebih baik dari metode tradisional yang telah sering dan umum digunakan yaitu TF-IDF [10]–[13].

Berdasarkan hal – hal yang telah dipaparkan, deteksi SMS spam dengan Bahasa Indonesia dapat diterapkan dengan salah satu metode STW yaitu TF-RF dan menggunakan metode klasifikasi Support Vector Machine. Dengan penerapan

metode tersebut diharapkan dapat menghasilkan performa yang baik dan sesuai dengan yang diharapkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan sebelumnya, rumusan masalah yang ingin dijawab pada penelitian ini adalah sebagai berikut.

1. Bagaimana performa metode TF-RF dengan metode klasifikasi SVM untuk deteksi SMS spam berbahasa Indonesia dalam hal akurasi, presisi dan recall?
2. Apakah metode TF-RF memiliki performa yang lebih baik dibandingkan dengan metode pembobotan fitur tradisional TF-IDF dengan metode klasifikasi SVM untuk digunakan dalam mendeteksi SMS spam berbahasa Indonesia?

1.3 Batasan Masalah

Penelitian ini memiliki beberapa batasan masalah guna memberikan lingkup penelitian sehingga lebih terfokus pada saat pengerjaannya. Adapun batasan masalah yang diberikan sebagai berikut:

1. Data latih dan data uji teks SMS yang digunakan dalam sistem adalah teks SMS berbahasa Indonesia.
2. Jenis-jenis SMS yang diproses berupa SMS promo, penawaran, pribadi, penipuan, judi, pinjaman online, dan operator.
3. Bagian dari SMS yang akan diolah pada saat preprocessing hingga klasifikasi hanya bagian subject/isi pesan.

1.4 Tujuan Penelitian

Adapun tujuan yang diharapkan dari penelitian ini adalah sebagai berikut

1. Mengetahui performa metode TF-RF untuk deteksi SMS spam berbahasa Indonesia.
2. Mengetahui perbandingan performa antara metode TF-RF dan pembobotan fitur tradisional TF-IDF untuk deteksi SMS spam berbahasa Indonesia.
3. Mengetahui performa metode Support Vector Machine untuk deteksi SMS spam berbahasa Indonesia.

1.5 Manfaat Penelitian

Manfaat yang bisa didapatkan dari penelitian ini adalah sebagai berikut

1. Mendapatkan update ilmu/teknologi yang dapat diterapkan pada permasalahan pengenalan SMS spam berbahasa Indonesia
2. Mengukur kinerja deteksi SMS spam berbahasa Indonesia menggunakan metode TF-RF.
3. Mengukur kinerja Support Vector Machine Classifier untuk deteksi SMS spam berbahasa Indonesia

1.6 Sistematika Penulisan

Sistematika untuk penulisan pada penelitian ini akan disajikan dalam beberapa bab antara lain sebagai berikut

1. Bab I Pendahuluan

Bab ini menjelaskan dasar-dasar dari penulisan laporan tugas akhir, yang terdiri dari latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat serta sistematika penulisan laporan tugas akhir.

2. Bab II Tinjauan Pustaka dan Dasar Teori

Bab ini membahas tentang penelitian-penelitian terdahulu yang mengimplementasikan metode tradisional dan supervised serta teori-teori sebagai referensi penulis ketika melakukan penelitian.

3. Bab III Metodologi Penelitian

Bab ini membahas tentang metodologi yang digunakan untuk pemrosesan data SMS, pembuatan model dalam pembobotan fitur dan proses klasifikasi.

4. Bab IV Hasil dan Pembahasan

Bab ini memuat tentang hasil dan pembahasan yang diperoleh berdasarkan hasil perhitungan dan membahas tentang metodologi yang digunakan dalam penelitian.

5. Bab V Kesimpulan dan Saran

Bab ini memaparkan hasil kesimpulan yang telah didapatkan dari hasil penelitian dan saran-saran yang diberikan untuk menyempurnakan hasil penelitian kedepannya.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Dilakukan penelitian dengan melakukan penambahan Genetic Algorithm dalam proses pemilihan atribut yang akan digunakan dalam proses klasifikasi pesan SMS dengan algoritma Naïve Bayes yang bertujuan untuk menyaring dan memisahkan SMS spam dan SMS nonspam. Dalam penelitian ini digunakan lima tahapan untuk preprocessing data yakni transform case, tokenisasi, filter stopwords, stemming, dan N-gram. Penelitian ini menggunakan teknik pembobotan TF-IDF. Tingkat keberhasilan klasifikasi pesan SMS menggunakan Naïve Bayes dengan GA menghasilkan tingkat akurasi yang lebih baik yaitu 89.73% dengan peningkatan sebesar 0.34% yang sebelumnya menggunakan Naïve Bayes sebesar 89.39% serta nilai AUC yang sebesar 0.654 [6].

Telah dilakukan penelitian dengan mengembangkan sistem untuk metode Support Vector Machine, dimana data SMS yang didapat dari database Kaggle diolah terlebih dahulu dengan menggunakan teknik tokenizing, normalisasi kata, filtering, dan stemming. Selanjutnya digunakan cross validation untuk menguji data training yang nantinya digunakan dalam proses klasifikasi. Algoritma SVM mampu mengklasifikasi spam dalam SMS dengan akurasi sebesar 96.72% dibanding naive bayes [7].

Penelitian menggunakan metode TF IDF ini untuk kasus sentiment analisis pada komentar instagram akan lebih baik jika menggunakan Support Vector Machine (SVM) dengan TF IDF. Dengan komposisi data terbaik untuk melakukan pengujian adalah 80% : 20% (data train : data tes) mendapatkan hasil nilai akurasi 87.45%, precision 87.72%, recall 91.74% dan F1-Score 89.69% pada Decision Tree dengan TF-IDF, sedangkan untuk Support Vector Machine dengan TF-IDF komposisi data terbaik untuk melakukan pengujian adalah 80% : 20% (data train : data tes) dengan mendapatkan hasil nilai akurasi 94.36%, precision 96.78%, recall 94.30% dan F1-Score 95.53%. Melihat dari hasil Support Vector Machine (SVM) dengan TF-IDF lebih baik dibandingkan Decision Tree dengan TF-IDF, tetapi hasil

dari kedua algoritma tersebut sudah sangat baik dikarenakan memiliki akurasi diatas 80% [8].

Dilakukan penelitian klasifikasi artikel berdasarkan tingkatan umur pembaca dengan menerapkan fitur term frequency dan inverse document frequency (TF-IDF) serta algoritma Multinomial Naive Bayes Classifier. Pada penelitian ini, data artikel yang digunakan bersumber dari 3 situs yaitu, bobo.grid.id yang merupakan situs dengan terget pembaca anak usia SD, untuk kategori remaja usia 15 - 24 tahun diperoleh dari situs hai.grid.id, sedangkan untuk kategori kelompok usia dewasa diperoleh dari situs www.detik.com. Hasil yang didapatkan pada penelitian ini yaitu nilai accuracy sebesar 93%, precision sebesar 94% dan recall sebesar 93% [9].

Telah dilakukan penelitian mengenai studi tentang term weighting untuk kategorisasi teks dengan mencoba varian supervised dari TF IDF yaitu TF RF dan yang lainnya. Dilakukan studi eksperimental ekstensif pada dua dataset, yaitu korpus Reuters dengan 10 atau 52 kategori dan 20 Newsgroup, dan tiga metode klasifikasi yang berbeda, yaitu pengklasifikasi SVM dengan fungsi kernel linear dan RBF serta RandomForest. Hasil yang diperoleh menunjukkan bahwa metode pembobotan TF RF mendapatkan hasil yang lebih baik pada semua dataset dan dengan semua pengklasifikasi dibandingkan dengan TF IDF, seperti pada Reuters-10, berurutan dengan kernel linear, RBF, dan RandomForest didapatkan akurasi 89%, 90%, dan 85% dibandingkan TF IDF didapatkan akurasi 87%, 80%, dan 84%. Melalui uji signifikansi statistik, ditunjukkan bahwa skema yang diusulkan selalu mencapai efektivitas lebih tinggi dan tidak pernah lebih buruk daripada metode TF IDF. Metode pembobotan TF RF memberikan hasil yang sangat baik dengan sedikit fitur dan menunjukkan beberapa penurunan (kurang dari 4%) ketika jumlah fitur meningkat [10].

Terdapat penelitian yang mengusulkan sebuah perbandingan beberapa metode term weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis, yaitu Term Frequency Inverse Document Frequency (TF-IDF), Term Frequency Inverse Document Frequency Inverse Class Frequency (TF-IDF-ICF), Term Frequency Inverse Document Frequency Inverse Class Space Density

Frequency (TF-IDFICS₈F), dan Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space Density Frequency (TF-IDF-ICS₈F-IHS₈F). Penelitian ini melakukan perbandingan hasil term weighting terhadap dataset Terjemahan 9 Kitab Hadis yang diterapkan pada mesin klasifikasi Naive Bayes dan SVM. Hasil uji coba menunjukkan bahwa hasil klasifikasi menggunakan metode term weighting TF-IDF-ICS₈FIHS₈F mengungguli term weighting lainnya, yaitu mendapatkan Precision sebesar 90%, Recall sebesar 93%, F1- Score sebesar 92%, dan Accuracy sebesar 83% [11].

Dilakukan penelitian perbandingan beberapa metode pembobotan untuk klasifikasi topik berita menggunakan decision tree. Dataset yang digunakan adalah artikel berita Bahasa Indonesia yang terdiri dari 12 kategori berita seperti politik, budaya, kesehatan, pendidikan, dan lain-lain yang terdiri dari 360 dokumen. Pengujian ini bertujuan untuk mengetahui Teknik pembobotan yang paling baik diantara TF-ABS, TF-CHI², TF-RF, dan TF-IDF. Berdasarkan pengujian diperoleh hasil bahwa TF-ABS menghasilkan akurasi yang paling tinggi sebesar 82,22% berselisih sedikit jika dibandingkan TF-CHI² yang akurasinya 80.83%, kedua Teknik lainnya memiliki akurasi yang lebih rendah yaitu TF-RF dengan nilai akurasi 65,56% dan yang paling rendah TF-IDF dengan nilai 50% [12].

Telah dilakukan penelitian analisis perbandingan metode pembobotan kata TF-IDF dan TF-RF pada trending topic di Twitter dan menggunakan metode pengklasifikasian dari data mining dimana metode yang digunakan adalah metode pengklasifikasian K-Nearest Neighbor. Jumlah data yang digunakan sebanyak 77793 data tweet didapat dari media Twitter yang dilabelkan secara manual dibagi kedalam 12 kategori yaitu ekonomi, hiburan, hukum, kesehatan, olahraga, otomotif, pendidikan, politik, seni budaya, sosial, teknologi, umum. Berdasarkan hasil pengujian implementasi pembobotan TF-IDF dan TF-RF terhadap klasifikasi K-Nearest neighbor mendapatkan hasil akurasi tertinggi menggunakan $k = 1$ dengan skenario (90-10) dan hasil akurasi didapat adalah 63,12% dengan precision 0,633 dan recall 0,633. Dalam hal ini kinerja perbandingan antara TF-IDF dan TF-RF dengan menggunakan klasifikasi K-Nearest Neighbor bahwasannya TF-IDF lebih baik dalam confusion matrix tersebut [13].

Berdasarkan berbagai penelitian yang telah dijelaskan sebelumnya, dapat disimpulkan bahwa metode pembobotan TF- IDF memiliki hasil yang cukup bagus, namun masih memiliki kekurangan sehingga dilakukan penelitian dengan metode yang lain dengan upaya perbaikan metode sebelumnya yang salah satunya TF-RF dan mendapatkan hasil yang bagus, bahkan di beberapa penelitian mendapat hasil yang lebih baik dari TF-IDF. Metode klasifikasi Support Vector Machine juga memiliki hasil yang baik ketika digunakan untuk pengelompokkan teks. Oleh karena itu, penelitian untuk deteksi SMS spam berbahasa Indonesia ini akan dilakukan dengan menggunakan metode TF-RF dan Support Vector Classifier, dimana pada pengujiannya akan membandingkan TF-RF dengan TF-IDF.

Berikut ini adalah state of art dari beberapa penelitian ini yang digunakan pada penelitian ini pada Tabel 2.1.

Tabel 2. 1 State of art

No	Judul	Tahun	SMS	TF-IDF	TF-RF	SVM
1	Klasifikasi Pesan Sms Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Genetic Algorithm	2018	√	√	×	×
2	Klasifikasi Sms Spam Menggunakan Support Vector Machine	2019	√	×	×	√
3	Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI	2020	×	√	×	√
4	Klasifikasi Artikel Berdasarkan Tingkatan Umur Pembaca	2019	×	√	×	×

	menggunakan Metode Multinomial Naive Bayes Classifier					
5	A study on term weighting for text categorization: A novel supervised variant of tf.idf	2015	×	√	√	√
6	Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis	2020	×	√	×	√
7	Perbandingan Pembobotan untuk Klasifikasi Topik Berita menggunakan Decision Tree	2019	×	√	√	×
8	Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada Trending Topic di Twitter dengan Menggunakan Klasifikasi K-Nearest Neighbor	2020	×	√	√	×
9	Deteksi SMS Spam Berbahasa Indonesia Menggunakan Metode Pembobotan Fitur TF-RF dan Support Vector Machine Classifier	2021	√	√	√	√

2.2 Teori Penunjang

Teori-teori penunjang yang digunakan dalam penelitian ini adalah sebagai berikut :

2.2.1 SMS Spam

Spamming sebagai perbuatan penyebaran pesan elektronik yang tidak diinginkan/diminta (unsolicited electronic messages) dan tanpa persetujuan penerimanya adalah fenomena sebagai akibat dari berkembangnya teknologi komunikasi dan informasi dan telah menimbulkan permasalahan hukum baru. Pesan elektronik meliputi email, pesan instan, SMS dan pesan lainnya seperti instant messaging (IM). Sebuah SMS dapat dikatakan unsolicited jika penerima tidak meminta untuk menerima pesan, tidak menyetujui untuk menerima pesan, dan tidak membutuhkan pesan tersebut. SMS Spam ini didefinisikan sebagai pengiriman informasi dan komunikasi elektronik untuk menampilkan berita iklan dan keperluan lainnya yang mengakibatkan ketidaknyamanan bagi para pengguna. Pada penelitian ini jenis-jenis SMS yang akan diolah berupa SMS promo, penawaran, penipuan, judi, pinjaman online, dan operator. Apapun jenis pesannya dari yang telah disebutkan, jika penerima tidak dikenal atau pengiriman tidak di waktu yang diinginkan dan pesan tidak dibutuhkan maka termasuk kedalam Spam karena mengakibatkan ketidaknyamanan. Spam ini biasanya datang tanpa diminta dan sering kali tidak dikehendaki oleh penerimanya, sehingga dapat mengganggu privasi (nuisance). Spamming melanggar privasi karena mengirimkan informasi (komunikasi) yang mengganggu privasi, berupa informasi yang tidak dikehendaki dan juga melanggar property. Hak atas privasi terjadi ketika informasi pribadi (property) seseorang diungkapkan, terdapat suatu keuntungan ekonomi yang diharapkan oleh pihak yang melakukan publikasi. Hak atas privasi adalah merupakan bagian dari Hak Asasi Manusia dan dilindungi oleh hukum, yang dalam Konstitusi Indonesia diatur dalam Pasal 28G UUD 1945 [14].

2.2.2 Text Mining

Text mining adalah proses menambang data berupa teks dengan sumber data biasanya dari dokumen dan tujuannya adalah mencari kata - kata yang mewakili dalam dokumen sehingga dapat dilakukan analisa keterhubungan dalam

dokumen. Data teks akan diproses menjadi data numerik agar dapat dilakukan proses lebih lanjut. Sehingga dalam text mining ada istilah preprocessing data, yaitu proses pendahulu yang diterapkan terhadap data teks yang bertujuan untuk menghasilkan data numerik [12].

2.2.3 Text Preprocessing

Text preprocessing meliputi semua proses yang dibutuhkan untuk mempersiapkan data yang akan diolah dalam text mining untuk mendapatkan informasi dan pengetahuan yang tersembunyi. Pada tahap ini teks dari berbagai sumber akan diubah ke dalam format yang sesuai untuk *text mining* [15]. Text preprocessing merupakan proses untuk mentransformasikan teks ke dalam kumpulan kata. Teks merupakan data yang tidak terstruktur, yang mana cukup sulit untuk diproses dengan komputer. Operasi numerik pun tidak dapat diaplikasikan pada data teks. Oleh karena itu, perlu dilakukan preprocessing pada teks untuk mendapatkan data yang dapat diolah menggunakan komputer [16].

2.2.3.1 Case Folding

Case Folding adalah proses mengubah seluruh huruf pada kata/teks menjadi lowercase seperti pada kalimat “Lokasi sangat strategis di tengah kota berseberangan dengan Ranch Market” menjadi “lokasi sangat strategis di tengah kota berseberangan dengan ranch market”.

2.2.3.2 Tokenizing

Tokenizing merupakan proses yang dilakukan untuk memisahkan kata dalam teks yang dipisahkan oleh spasi atau whitespace dan menjadikannya ke dalam bentuk *array* atau susunan kata, seperti “lokasi sangat strategis” menjadi ['lokasi', 'sangat', 'strategis'].

2.2.3.3 Stopword Removal

Stopword removal atau *filtering* merupakan proses yang bertujuan untuk menghilangkan kata-kata yang termasuk kedalam kategori *stop-word*, dimana *stop-word* sendiri adalah kata yang termasuk kata umum yang tidak memiliki arti spesifik yang dapat mempengaruhi nilai pada suatu kalimat, seperti “saya”, “dan”, “akan”.

2.2.3.4 Stemming

Stemming adalah proses mentranslasi kata ke dalam bentuk dasar atau kata dasarnya seperti “pelayanan sangat baik” menjadi “layanan sangat baik”.

2.2.4 Klasifikasi Teks

Klasifikasi teks atau kategorisasi teks merupakan salah satu teknik pada *Data Mining* yang bertujuan untuk secara otomatis mengelompokkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut. Hal tersebut bertujuan untuk memetakan sekumpulan data teks yang tidak beraturan menjadi beberapakelompok yang menggambarkan isi dari teks tersebut. Dengan pengklasifikasian yang terotomasi akan mempercepat proses pemetaan data yang berjumlah besar, serta dapat menghindari kesalahan pemetaan data yang masih dilakukan secara manual [15].

2.2.5 Term Weighting

Term Weighting adalah prosedur yang dilakukan saat pemberian index kata yang bertujuan untuk memberi nilai pada setiap *term* atau kata pada suatu dokumen. Secara sederhana *Term-Weighting* merupakan pembobotan dokumen menggunakan representasi vector space model dari kumpulan dataset. Dokumen dalam vector space model direpresentasikan dalam matriks yang berisi bobot kata pada dokumen. Bobot tersebut menyatakan kontribusi atau kepentingan kata pada kumpulan dokumen dan dokumen itu sendiri. Kepentingan kata dalam sebuah dokumen dapat dilihat dari tingkat kemunculannya terhadap dokumen. Kata yang berbeda memiliki tingkat kemunculan yang berbeda pula. Dibawah ini merupakan metode perhitungan *Term Frequency* (TF), *Inverse Document Frequency* (IDF), dan *Relevance Frequency* (RF) yang kemudian dikombinasikan untuk menghasilkan nilai *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Term Frequency-Relevance Frequency* (TF-RF) [15].

2.2.5.1 Term Frequency (TF)

TF adalah salah satu metode pembobotan term yang paling sederhana, caranya adalah dengan menghitung jumlah kata atau term yang muncul dalam satu dokumen. Setiap term t diasumsikan memiliki kepentingan yang proporsional terhadap jumlah kemunculan term pada dokumen d [15]. Dengan metode ini, nilai

kontribusi (bobot) suatu term pada suatu dokumen adalah sama dengan jumlah munculnya term tersebut pada dokumen [12].

$$TF(d, t) = f(d, t) \quad (2.1)$$

Dimana $f(d, t)$ adalah frekuensi kemunculan term t pada dokumen d .

2.2.5.2 Term Frequency – Inverse Document Frequency (TF-IDF)

Inverse Document Frequency (IDF) adalah metode perhitungan bobot yang mirip dengan TF hanya saja pada IDF mencari kemunculan *term* pada kumpulan dokumen, berbeda dengan TF yang hanya memerhatikan kemunculan *term* pada dokumen tersebut. Dalam kata lain IDF memperhatikan jumlah dokumen d yang memiliki kata atau term t . Pembobotan ini dilakukan untuk memberikan nilai yang tinggi pada *term t* yang jarang muncul pada kumpulan dokumen d karena *term t* sangat bernilai. Kepentingan tiap term t diasumsikan memiliki proporsi yang berkebalikan dengan jumlah dokumen d yang mengandung *term t*. Perhitungan nilai *Inverse Document* [15].

$$IDF(t) = \log \frac{n}{df(t)} \quad (2.2)$$

Setelah menemukan nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF), maka akan dilakukan pembobotan TF-IDF dengan mengalikan nilai TF dengan nilai IDF. Perhitungan nilai TF-IDF dapat dilihat pada Persamaan 2.3.

$$TF.IDF = TF(d, t) * IDF(t) \quad (2.3)$$

Keterangan variabel :

$TF(d, t)$	= bobot <i>term t</i> pada setiap dokumen d
$f(d, t)$	= frekuensi kemunculan <i>term t</i> pada dokumen d
n	= jumlah seluruh dokumen
$df(t)$	= banyak dokumen yang mengandung term t
$IDF(t)$	= bobot <i>Inverse Document Frequency</i> (IDF)
$TF.IDF$	= nilai pembobotan TF-IDF

2.2.5.3 Term Frequency – Relevance Frequency (TF-RF)

Relevance frequency merupakan metode yang muncul sebagai upaya perbaikan terhadap metode-metode yang sudah ada. Sebagai contoh, metode IDF

hanya akan menilai term berdasarkan kemunculan (ada atau tidaknya saja) term pada suatu dokumen. Berbeda dengan metode RF yang diusulkan oleh Man Lan, metode ini mempertimbangkan relevansi dokumen dilihat dari frekuensi kemunculan term di kategori yang berkaitan [17]. Persamaan untuk menghitung RF dapat dilihat di Persamaan 2.4 [12].

$$RF(t_j, c_i) = \log \left(2 + \frac{n_{ij}}{\max(1, n_{\sim ij})} \right) \quad (2.4)$$

Setelah menemukan nilai *Term Frequency* (TF) dan *Relevance Frequency* (RF), maka akan dilakukan pembobotan TF-RF dengan mengalikan nilai TF dengan nilai RF. Perhitungan nilai TF-RF dapat dilihat pada Persamaan 2.5.

$$TF.RF = TF(d, t) * RF(t_j, c_i) \quad (2.5)$$

Keterangan variabel :

$TF(d, t)$ = bobot *term t* pada setiap dokumen *d*

n_{ij} = jumlah dokumen dalam kategori *cj* yang mengandung term *tj*

$n_{\sim ij}$ = jumlah dokumen tidak dalam kategori *cj* yang mengandung term *tj*

$RF(t_j, c_i)$ = bobot *Relevance Frequency* (RF)

$TF.RF$ = nilai pembobotan TF-RF

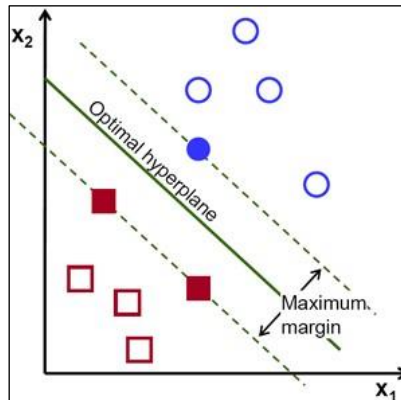
2.2.6 Support Vector Machine Classifier

Pada penelitian ini, pada proses klasifikasi akan digunakan metode klasifikasi Support Vector Machine. Karakteristik yang dimiliki oleh SMS yaitu memiliki panjang maksimum 160 karakter karena terbatasnya kemampuan saluran yang digunakan sehingga hanya ada sedikit yang bisa digunakan sebagai fitur pada tiap pesannya untuk proses klasifikasi. Karena panjang SMS yang tersedia, pengguna SMS menggunakan subset bahasa istimewa dengan singkatan, kontraksi fonetik, tanda baca yang buruk, emotikon, dll., yang berbeda dengan bahasa tertulis yang lebih tradisional yang biasanya digunakan dalam email, dimana jika dibandingkan dengan email bahwa penyaringan email spam dapat ditingkatkan dengan memasukkan informasi kontekstual yang ditemukan di header email, tetapi SMS berisi informasi yang jauh lebih sedikit di header, yang menawarkan lebih sedikit konteks untuk dikerjakan [18]. Sehingga karena alasan itulah juga mengapa bagian SMS yang diolah dari preprocessing hingga klasifikasi hanya bagian

subject/isi pesannya saja. Berdasarkan karakteristik yang dipaparkan, rekomendasi classifier yang baik untuk dapat digunakan yaitu Support Vector Machine [19], [20]. Terlebih lagi prinsip kerja SVM pada dasarnya classifier yang baik untuk menangani klasifikasi 2 class dan jumlah dataset yang akan diolah tidak berskala besar, maka metode ini dirasa akan mendapat hasil yang baik dalam proses klasifikasi nantinya [21].

Secara sederhana konsep dari SVM adalah proses untuk mencari *hyperplane* terbaik yang dimana dapat memisahkan dua buah kelas pada input space. Proses Klasifikasi dapat diartikan sebagai proses menemukan garis Batasan (*hyperplane*) yang membedakan atau memisahkan kedua kelas. *Support Vector Machine* adalah suatu metode klasifikasi yang bertujuan untuk menemukan *Maximum Marginal Hyperplane* atau MMH yang merupakan batas pemisah terbaik atau pemisah maksimal untuk semua kelas. Support Vector Machine (SVM) biasanya digunakan dalam kasus klasifikasi maupun regresi sebagai salah satu teknik prediksi. Tidak sama dengan strategi neural network yang dimana hanya berusaha mencari *hyperplane* pemisah antarclass, SVM akan menemukan *hyperplane* terbaik diantara *hyperplane* lain pada input space.

Untuk mendapatkan *hyperplane* pemisah terbaik diantara kedua class adalah dengan menghitung nilai margin *hyperplane* dan mencari titik dengan nilai tertinggi. Margin merupakan jarak antara *hyperplane* dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai *Support Vector*. Jika dilihat pada Gambar 2.4, garis yang terletak diantara kedua kelas adalah *hyperplane* terbaik, dan persegi merah dan lingkaran biru yang *Support Vector*. Semakin besar margin maka semakin tinggi akurasi. Proses untuk menemukan lokasi *hyperplane* ini adalah inti dari proses pembelajaran pada SVM [15].



Gambar 2. 1 Support Vector Machine

Cara kerja algoritma SVM adalah dengan menggambarkan data set kedalam bentuk grafik (X_i, y_i) dimana X_i adalah kumpulan *tuples* dengan label kelas pada y_i . Setiap kelas dapat memilih salah satu dari dua nilai yaitu antara +1 atau -1 ($y_i \in \{+1, -1\}$). Untuk mendapatkan *hyperplane* pemisah dapat dilakukan dengan persamaan dibawah:

$$W \cdot X + b = 0 \quad (2.6)$$

Keterangan variabel :

W = bobot vector $\{w_1, w_2, \dots, w_n\}$

n = jumlah atribut

b = nilai scalar (bias)

X = potongan data latih (*training tuples*)

Jika b adalah bobot tambahan maka nilai tersebut dapat didefinisikan sebagai w_0 , seperti pada persamaan dibawah ini:

$$w_0 + w_1x_1 + w_2x_2 = 0 \quad (2.7)$$

Dengan demikian, nilai yang berada diatas *hyperplane* pemisah akan memenuhi persamaan berikut:

$$w_0 + w_1x_1 + w_2x_2 > 0 \quad (2.8)$$

Nilai yang berada dibawah *hyperplane* pemisah akan memenuhi persamaanberikut:

$$w_0 + w_1x_1 + w_2x_2 < 0 \quad (2.9)$$

Bobot dari setiap persamaan dapat disesuaikan sehingga *hyperplane* pemisah untuk setiap sisinya dapat dituliskan seperti persamaan dibawah ini:

$$H_1 : w_0 + w_1x_1 + w_2x_2 \geq +1 \text{ untuk } y_i = +1 \quad (2.10)$$

$$H_2 : w_0 + w_1x_1 + w_2x_2 \leq -1 \text{ untuk } y_i = -1 \quad (2.11)$$

Dari kedua pertidaksamaan diatas dapat dilihat bahwa setiap *tuple* atau nilai yang berada tepat atau diatas H_1 termasuk kedalam kelas +1 dan setiap *tuple* atau nilai yang berada tepat atau dibawah H_2 termasuk kedalam kelas -1.

2.2.7 Evaluasi

Hasil klasifikasi kemudian dievaluasi untuk mendapatkan nilai akurasi yang akan dianalisis apakah model klasifikasi yang dibuat layak digunakan [22]. Teknik yang digunakan untuk melakukan evaluasi dalam klasifikasi pada penelitian ini adalah dengan menghitung recall, precision, dan f-measure. Teknik ini menggunakan confusion matrix sebagai acuan perhitungan [16]. Confusion matrix adalah metode perhitungan akurasi yang biasanya digunakan pada konsep *Data Mining* atau Sistem Pendukung Keputusan. Pada pengukuran kinerja menggunakan Confusion Matrix, terdapat 4 (empat) yang merepresentasikan hasil proses klasifikasi seperti yang ditunjukkan pada Tabel 2.2 . Keempat istilah tersebut adalah True Positive (TP), True Negative (TN), False Positive (FP) dan False Negative (FN). Nilai True Negative (TN) adalah jumlah datanegatif yang terdeteksi dengan benar, sedangkan False Positive (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, True Positive (TP) merupakan data positif yang terdeteksi benar. False Negative (FN) merupakankebalikan dari True Positive, sehingga data positif, namun terdeteksi sebagai data negatif.

Tabel 2. 2 Confusion Matrix

		Aktual	
		Positive	Negative
Prediksi	Positive	TP	FP
	Negative	FN	TN

Confusion matrix menunjukkan tingkat akurasi dari model klasifikasi yang sudah dilakukan sebelumnya. *Accuracy* menunjukkan proporsi jumlah prediksi benar. Berikut merupakan rumus untuk menguji akurasi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.12)$$

Recall atau *true positif rate* (TP) adalah proporsi dari kasus positif yang telah diidentifikasi dengan benar, berikut rumus untuk mencari *Recall*.

$$Recall = \frac{TP}{TP + FN} \quad (2.13)$$

Precision adalah tingkat ketepatan proporsi kasus positif yang telah di prediksi dengan benar (TP) dengan keseluruhan kasus yang telah diprediksi, berikut rumus untuk mencari *Precision*.

$$Precision = \frac{TP}{TP + FP} \quad (2.14)$$

BAB III

METODOLOGI PENELITIAN

3.1 Alat dan Bahan

Alat - alat yang diperlukan dalam penelitian tugas akhir ini dibagi menjadi dua yakni perangkat keras dan perangkat lunak antara lain sebagai berikut:

1. Perangkat Keras

Perangkat keras yang digunakan dalam penelitian ini adalah laptop dengan spesifikasi sebagai berikut :

- a. Processor AMD Ryzen 5
- b. Memori RAM 8 GB

2. Perangkat Lunak

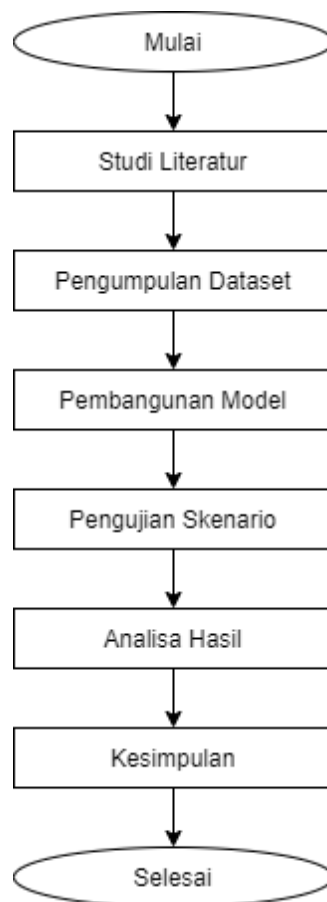
Perangkat lunak yang digunakan dalam penelitian ini yaitu :

- a. Sistem operasi Windows
- b. Jupyter Lab
- c. Bahasa pemrograman python versi 3.10.1
- d. Web browser

Bahan penelitian yang digunakan pada tugas akhir ini adalah data SMS yang dikumpulkan dari ponsel pribadi dan beberapa ponsel orang lain dengan adanya persetujuan pemilik, ponsel terlebih dahulu mendownload aplikasi melalui playstore bernama SMS Backup and Restore. SMS yang diambil yaitu SMS yang diterima sejak awal tahun 2021 hingga yang terbaru saat ini masih diterima. Jumlah data SMS yang dikumpulkan dari beberapa ponsel tersebut adalah 300 SMS. Selanjutnya SMS yang telah dikumpulkan diberi label (spam dan non-spam) sesuai dengan sumber pengirim SMS, waktu pengiriman SMS dan isi dari SMS tersebut.

3.2 Alur Penelitian

Pada penelitian ini, alur penelitian dapat digambarkan menggunakan diagram alur pada Gambar 3.1.

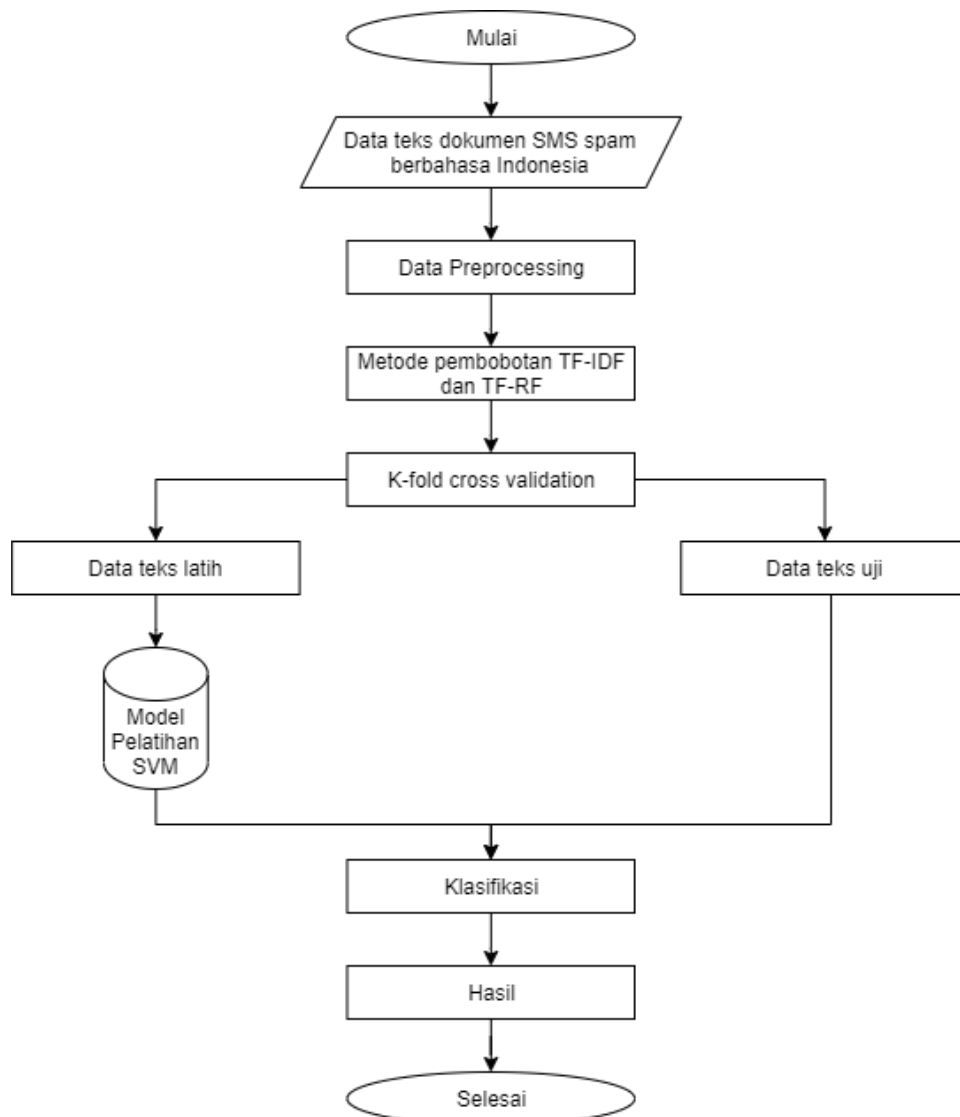


Gambar 3. 1 Diagram alur pembuatan sistem

Pada penelitian ini, tahap awal yang dilakukan adalah studi literatur untuk mempelajari tentang metode yang digunakan dalam penelitian dan perancangan sistem pengumpulan sampel SMS. Tahapan selanjutnya adalah pengumpulan dataset, dataset SMS tersebut dikumpulkan dari beberapa ponsel sejumlah 300 SMS. Selanjutnya SMS yang telah dikumpulkan diberi label spam dan non-spam. Setelah pengumpulan dataset, tahap selanjutnya adalah pengembangan model sesuai dengan literatur yang telah dipelajari. Kemudian dilakukan pengujian terhadap model yang telah dibangun untuk mengetahui apakah model telah mendapatkan hasil yang sesuai. Setelah dilakukan beberapa skenario, dilakukan analisa terhadap hasil yang didapatkan dari berbagai macam skenario yang telah direncanakan. Terakhir diambil kesimpulan dari penelitian yang telah dijalankan dari awal hingga hasil yang didapatkan.

3.3 Perancangan Sistem

Rancangan dari sistem deteksi SMS spam berbahasa Indonesia dengan metode TF-RF menggunakan klasifikasi support vector machine terdiri dari beberapa tahapan, yang dapat dilihat pada Gambar 3.2 Rancangan sistem.



Gambar 3. 2 Rancangan sistem

3.4 Cara Analisis

Langkah-langkah dalam melakukan analisis berdasarkan tahapan dalam perancangan sistem dapat dilihat seperti berikut:

3.4.1 Input Training dan Testing SMS

Pada tahap ini SMS yang telah dikumpulkan akan diproses oleh sistem dalam dua fase yaitu training SMS dan testing SMS.

1. Input Training

Pada tahap ini, SMS yang telah diberi label dari berbagai ponsel sebagai SMS dengan kategori spam dan non-spam akan dimasukkan kedalam sistem untuk diproses pada tahap selanjutnya, hasil preprocessing berupa gabungan tiap kata pada tiap SMS akan digunakan sebagai fitur untuk diproses pada pembobotan dan klasifikasi.

2. Input Testing SMS

Pada tahap ini, SMS yang tidak diketahui kelasnya akan dimasukkan kedalam sistem untuk diproses seperti pada proses training SMS, hasil preprocessing berupa gabungan tiap kata pada tiap SMS akan digunakan sebagai fitur untuk diproses pada pembobotan dan klasifikasi.

Berikut ini beberapa contoh training serta testing SMS yang diperoleh dari beberapa ponsel dapat dilihat pada Tabel 3.1.

Tabel 3. 1 Contoh training dan testing SMS

Kategori Dokumen	Isi Dokumen
Spam	Maaf Mengganggu Kmi Ingin Menawarkan Pinjaman Online Bunga Rendah Mulai 5jt sampai 500jt Info Whatshapp : 0823-3432-7577
Spam	maaf anda perlu uang untuk usaha dll info chat wa.085283033802
Non-Spam	Plg yth, Paket Extra Pulsa Rp10000 Anda telah berakhir. Pemakaian data selanjutnya akan dikenakan tariff normal. Dapatkan paket lainnya di *363# atau MyTelkomsel.
Non-Spam	Penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. Masa aktif kartu

Kategori Dokumen	Isi Dokumen
	anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021
Testing SMS	Saat libur Idul Fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. Nekat mudik, korbakan keluarga! covid19.go.id #tidakmudik

3.4.2 Text Preprocessing

Pada tahap ini, terdapat beberapa hal yang dilakukan agar data yang diolah pada tahap selanjutnya dapat diproses dengan baik. Tahap - tahap tersebut meliputi:

a. Case Folding

Tahap ini berfungsi untuk mengubah seluruh kata pada dokumen atau korpus menjadi huruf kecil agar tidak terjadi ambiguitas saat membandingkan kata yang diawali dengan huruf besar dan huruf kecil pada term atau kata yang sama. Contoh case folding pada dokumen SMS seperti pada Tabel 3.2.

Tabel 3. 2 Contoh case folding SMS

Input Dokumen	Dokumen setelah Case Folding
Maaf Mengganggu Kmi Ingin Menawarkan Pinjaman Online Bunga Rendah Mulai 5jt sampai 500jt Info Whatshapp : 0823-3432-7577	maaf mengganggu kmi ingin menawarkan pinjaman online bunga rendah mulai 5jt sampai 500jt info whatshapp : 0823-3432-7577
maaf anda perlu uang untuk usaha dll info chat wa.085283033802	maaf anda perlu uang untuk usaha dll info chat wa.085283033802
Plg yth, Paket Extra Pulsa Rp10000 Anda telah berakhir. Pemakaian data selanjutnya akan dikenakan tariff normal. Dapatkan paket lainnya di *363# atau MyTelkomsel.	plg yth, paket extra pulsa rp10000 anda telah berakhir. pemakaian data selanjutnya akan dikenakan tariff normal. dapatkan paket lainnya di *363# atau mytelkomsel.

Input Dokumen	Dokumen setelah Case Folding
Penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. Masa aktif kartu anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021	penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. masa aktif kartu anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021
Saat libur Idul Fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. Nekat mudik, korbakan keluarga! covid19.go.id #tidakmudik	saat libur idul fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. nekat mudik, korbakan keluarga! covid19.go.id #tidakmudik

b. Tokenizing

Proses ini berfungsi untuk mengubah kumpulan kalimat pada teks menjadi satuan kata atau token. Contoh tokenizing pada dokumen SMS seperti pada Tabel 3.3.

Tabel 3. 3 Contoh tokenizing SMS

Input Dokumen	Dokumen setelah Tokenizing
Maaf Mengganggu Kmi Ingin Menawarkan Pinjaman Online Bunga Rendah Mulai 5jt sampai 500jt Info Whatshapp : 0823-3432-7577	maaf, mengganggu, kmi ingin, menawarkan, pinjaman, online, bunga, rendah, mulai, sampai, info, whatshapp
maaf anda perlu uang untuk usaha dll info chat wa.085283033802	maaf, anda, perlu, uang, untuk, usaha, dll, info, chat, wa
Plg yth, Paket Extra Pulsa Rp10000 Anda telah berakhir. Pemakaian data selanjutnya akan dikenakan tariff normal. Dapatkan paket lainnya di *363# atau MyTelkomsel.	plg, yth, paket, extra, pulsa, anda, telah, berakhir, pemakaian, data, selanjutnya, akan, dikenakan, tariff, normal, dapatkan, paket, lainnya, di, atau, mytelkomsel

Input Dokumen	Dokumen setelah Tokenizing
Penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. Masa aktif kartu anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021	penggunaan, terakhir, anda, adalah, dan, telah, memenuhi, pemakaian, pulsa, minimum, masa, aktif, kartu, anda, diperpanjang, selama, hari, dan, akan, aktif, sampai
Saat libur Idul Fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. Nekat mudik, korbakan keluarga! covid19.go.id #tidakmudik	saat, libur, idul, fitri, terjadi, kenaikan, kasus, harian, dan, tingkat, kematian, mingguan, nekat, mudik, korbakan, keluarga, covid19.go.id

c. Stopword Filtering

Tahap ini berfungsi untuk menghilangkan kata - kata yang tidak relevan yang terdapat pada korpus atau dokumen. Penghilangan kata - kata yang tidak relevan pada dokumen dilakukan agar kata yang tidak mewakili ciri dari suatu dokumen tidak diproses sehingga dapat mempercepat waktu komputasi. Stopword list yang digunakan pada penelitian ini diperoleh dari daftar konjungsi dalam Bahasa Indonesia. Contoh dokumen SMS yang telah melalui proses stopwords filtering dapat dilihat pada Tabel 3.4.

Tabel 3. 4 Contoh stopwords filtering SMS

Input Dokumen	Dokumen setelah Stopword Filtering
Maaf Mengganggu Kmi Ingin Menawarkan Pinjaman Online Bunga Rendah Mulai 5jt sampai 500jt Info Whatshapp : 0823-3432-7577	maaf, mengganggu, kmi ingin, menawarkan, pinjaman, online, bunga, rendah, mulai, sampai, info, whatshapp
maaf anda perlu uang untuk usaha dll info chat wa.085283033802	maaf, anda, perlu, uang, usaha, info, chat, wa
Plg yth, Paket Extra Pulsa Rp10000 Anda telah berakhir. Pemakaian data	plg, yth, paket, extra, pulsa, anda, berakhir, pemakaian, data,

Input Dokumen	Dokumen setelah Stopward Filtering
selanjutnya akan dikenakan tariff normal. Dapatkan paket lainnya di *363# atau MyTelkomsel.	selanjutnya, dikenakan, tariff, normal, dapatkan, paket, lainnya, mytelkomsel
Penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. Masa aktif kartu anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021	penggunaan, terakhir, anda, telah, memenuhi, pemakaian, pulsa, minimum, masa, aktif, kartu, anda, diperpanjang, selama, hari, aktif, sampai
Saat libur Idul Fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. Nekat mudik, korbakan keluarga! covid19.go.id #tidakmudik	libur, idul, fitri, terjadi, kenaikan, kasus, harian, tingkat, kematian, mingguan, nekat, mudik, korbakan, keluarga

d. Stemming

Proses ini berfungsi untuk mengubah kata bentukan atau kata dengan imbuhan menjadi kata dasarnya dengan menghilangkan awalan dan akhiran kata. Proses stemming diperlukan agar setiap kata yang memiliki imbuhan yang berbeda-beda tetap dianggap sebagai kata yang sama sesuai dengan kata dasarnya. Algoritma stemming yang digunakan dalam penelitian ini adalah Algoritma Nazief & Adriani yang diperoleh dari pustaka Sastrawi. Contoh dokumen SMS setelah melalui proses stemming dapat dilihat pada Tabel 3.5.

Tabel 3. 5 Contoh stemming SMS

Input Dokumen	Dokumen setelah Stemming
Maaf Mengganggu Kmi Ingin Menawarkan Pinjaman Online Bunga Rendah Mulai 5jt sampai 500jt Info Whatshapp : 0823-3432-7577	maaf, ganggu, kmi ingin, tawar, pinjaman, online, bunga, rendah, mulai, sampai, info, whatshapp

Input Dokumen	Dokumen setelah Stemming
maaf anda perlu uang untuk usaha dll info chat wa.085283033802	maaf, anda, perlu, uang, usaha, info, chat, wa
Plg yth, Paket Extra Pulsa Rp10000 Anda telah berakhir. Pemakaian data selanjutnya akan dikenakan tariff normal. Dapatkan paket lainnya di *363# atau MyTelkomsel.	plg, yth, paket, extra, pulsa, anda, akhir, pakai, data, lanjut, kena, tarif, normal, dapat, paket, lain, mytelkomsel
Penggunaan terakhir anda adalah 170003.00 dan telah memenuhi pemakaian pulsa minimum. Masa aktif kartu anda diperpanjang selama 30 hari dan akan aktif sampai 30/08/2021	guna, terakhir, anda, telah, penuh, pakai, pulsa, minimum, masa, aktif, kartu, anda, panjang, lama, hari, aktif, sampai
Saat libur Idul Fitri 2020, terjadi kenaikan kasus harian 93% dan tingkat kematian mingguan 66%. Nekat mudik, korbankan keluarga! covid19.go.id #tidakmudik	libur, idul, fitri, terjadi, naik, kasus, hari, tingkat, mati, minggu, nekat, mudik, korban, keluarga

3.4.3 Pattern Discovery

Tahap ini bertujuan untuk menentukan pola kata pada suatu SMS seperti menentukan ciri atau fitur pada masing - masing kategori SMS melalui pembobotan pada term atau kata pada SMS. Pada proses ini juga dilakukan training dataset yang akan menghasilkan model training. Berikut ini beberapa tahap yang dilakukan pada tahap pattern discovery :

3.4.3.1 Term Frequency – Inverse Document Frequency

a. Term Frequency (TF)

Pada tahap ini, dibentuk vector space model berdasarkan term atau kata yang terdapat pada seluruh teks. Seluruh kata yang terdapat pada dokumen akan dijadikan sebagai feature pada masing – masing dokumen training dan test. Kemudian setiap dokumen training dan testing akan ditransformasi menjadi

bentuk vector space model dan mengikuti urutan kata seperti yang sudah dibuat sebelumnya berdasarkan kata yang terdapat pada dokumen. Nilai numerik akan diberikan pada vector dokumen training dan test berdasarkan jumlah kata yang muncul pada dokumen tersebut sesuai dengan kata acuannya. Nilai – nilai tersebut akan menjadi bobot term frequency sesuai dengan Persamaan (2.1).

Berikut ini contoh beberapa data yang telah melalui proses case folding, stopword filtering serta proses tokenisasi yang dibentuk menjadi feature dalam bentuk vector pada dokumen training:

[maaf, ganggu, kami, ingin, tawar, pinjaman, online, bunga, rendah, mulai, sampai, info, whatsapp, anda, perlu, uang, usaha, chat, wa, plg, yth, paket, extra, pulsa, akhir, pakai, data, lanjut, kena, tarif, normal, dapat, mytelkomsel, guna, terakhir, telah, penuh, minimum, masa, aktif, kartu, panjang, lama, hari]

Tahap selanjutnya yaitu mencari frekuensi kemunculan kata-kata tersebut pada tiap dokumen di masing-masing kelas. Hasil yang didapatkan dapat dilihat pada Tabel 3.6 Nilai TF pada beberapa kata.

Tabel 3. 6 Nilai TF

Term	TF			
	D1	D2	D3	D4
maaf	1	1	0	0
ganggu	1	0	0	0
kami	1	0	0	0
ingin	1	0	0	0
tawar	1	0	0	0
...

b. Inverse Document Frequency (IDF)

Setelah mendapatkan jumlah kemunculan kata (term frequency) pada masing-masing dokumen, selanjutnya dicari nilai inverse document frequency pada setiap kata. Nilai inverse document frequency mewakili seberapa sering suatu kata muncul pada suatu dokumen atau kelas. Apabila terdapat kata yang

memiliki term frequency tinggi pada suatu dokumen, namun term frequency kata tersebut juga tinggi pada kelas dokumen lainnya, maka nilai dari IDF kata tersebut menjadi rendah yang berarti kata tersebut tidak dapat mewakili ciri spesifik dari suatu dokumen. Berikut ini contoh perhitungan serta hasil dari pencarian nilai IDF menggunakan Persamaan (2.2) pada masing-masing kata:

Contoh nilai IDF pada kata “akhir” :

$$IDF_{(akhir)} = \log \frac{4}{1} = 0.60206$$

Sehingga nilai IDF pada beberapa kata dapat dilihat pada Tabel 3.7 berikut :

Tabel 3. 7 Nilai DF dan IDF pada beberapa kata

Term	DF	IDF $\log \frac{n}{df(t)}$
maaf	2	0,30103
ganggu	1	0,60206
kami	1	0,60206
ingin	1	0,60206
tawar	1	0,60206
...

c. Term Frequency – Inverse Document Frequency (TF-IDF)

Tahap terakhir untuk memberi bobot pada kata berdasarkan metode TF-IDF adalah mengalikan nilai term frequency dengan nilai inverse document frequency seperti pada Persamaan (2.3). Nilai akhir TF-IDF pada beberapa kata dapat dilihat pada Tabel 3.8.

Tabel 3. 8 Nilai TF-IDF pada beberapa kata

Term	$TF.IDF = TF(d,t) * IDF(t)$			
	D1	D2	D3	D4
maaf	0,30103	0,30103	0	0
ganggu	0,60206	0	0	0
kami	0,60206	0	0	0
ingin	0,60206	0	0	0
tawar	0,60206	0	0	0
...

3.4.3.2 Term Frequency – Relevance Frequency

a. Relevance Frequency (RF)

Setelah mendapatkan jumlah kemunculan kata (term frequency) pada masing-masing dokumen, selanjutnya dicari nilai relevance frequency pada setiap kata. Nilai relevance frequency mempertimbangkan relevansi dokumen dilihat dari frekuensi kemunculan term di kategori yang berkaitan. Kebalikan dari IDF, apabila terdapat kata yang memiliki term frequency tinggi pada suatu dokumen, dan term frequency kata tersebut juga tinggi pada kelas dokumen lainnya, maka nilai dari RF kata tersebut menjadi tinggi yang berarti kata tersebut dapat mewakili ciri spesifik dari suatu dokumen. Berikut ini contoh perhitungan serta hasil dari pencarian nilai RF menggunakan Persamaan (2.4) pada masing-masing kata:

Contoh nilai RF pada kata “akhir” :

$$RF_{(akhir)} = \log \left(2 + \frac{1}{\max(1,0)} \right) = 0.47712$$

Sehingga nilai RF pada beberapa kata dapat dilihat pada Tabel 3.9 berikut:

Tabel 3. 9 Nilai RF pada beberapa kata

Term	n_{ij}	$n_{\sim ij}$	RF $\log \left(2 + \frac{n_{ij}}{\max(1, n_{\sim ij})} \right)$
maaf	2	0	0,60206
ganggu	1	0	0,47712
kami	1	0	0,47712
ingin	1	0	0,47712
tawar	1	0	0,47712
...

b. Term Frequency – Relevance Frequency (TF-RF)

Tahap terakhir untuk memberi bobot pada kata berdasarkan metode TF-RF adalah mengalikan nilai term frequency dengan nilai relevance frequency seperti pada Persamaan (2.5). Nilai akhir TF-RF pada beberapa kata dapat dilihat pada Tabel 3.10.

Tabel 3. 10 Nilai TF-RF pada beberapa kata

Term	$TF.RF = TF(d, t) * RF(t)$			
	D1	D2	D3	D4
maaf	0,60206	0,60206	0	0
ganggu	0,477121	0	0	0
kami	0,477121	0	0	0
ingin	0,477121	0	0	0
tawar	0,477121	0	0	0
...

3.4.4 Klasifikasi Support Vector Machine

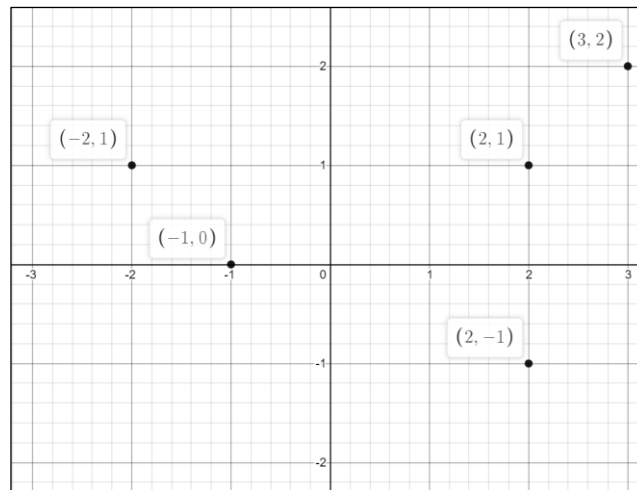
Setelah mendapatkan semua feature TF-IDF dan TF-RF seluruh kata pada data training, maka tahap selanjutnya adalah klasifikasi SMS berdasarkan model training data sebelumnya. Model training tersebut akan digunakan oleh dokumen uji untuk menentukan suatu SMS termasuk pada kategori SMS spam atau tidak. Algoritma klasifikasi data yang digunakan pada penelitian ini adalah support vector machine classifier.

Perhitungan ini akan dilakukan menggunakan nilai dari perhitungan TF-IDF, karena disini hanya sebagai contoh perhitungan manualnya, fitur yang digunakan hanya dua agar dapat mudah dipahami, bentuk nilai yang akan diolah dapat dilihat pada Tabel 3.11.

Tabel 3. 11 Contoh nilai fitur dengan 2 kelas

	X_1	X_2	Kelas
Dokumen 1	-2	1	-1
Dokumen 2	-1	0	-1
Dokumen 3	2	1	+1
Dokumen 4	2	-1	+1
Dokumen 5	3	2	+1

Titik-titik di atas dimasukkan ke dalam diagram kartesius dan membentuk diagram scatter plot seperti Gambar 3.3 Dataset dalam diagram kartesius.



Gambar 3. 3 Dataset dalam diagram kartesius

Dari nilai-nilai pada table diambil 3 buah data support vector yaitu $S_1=(-1, 0)$, $S_2=(2, 1)$ dan $S_3=(2, -1)$. Selanjutnya semua data ini digunakan untuk mencari persamaan hyperplane. Adapun proses pencarian persamaan hyperplane dijabarkan sebagai berikut.

$$\begin{aligned} \alpha_1 \phi(S_1) \cdot \phi(S_1) + \alpha_2 \phi(S_2) \cdot \phi(S_1) + \alpha_1 \phi(S_3) \cdot \phi(S_1) &= -1 \\ \alpha_1 \phi(S_1) \cdot \phi(S_2) + \alpha_2 \phi(S_2) \cdot \phi(S_2) + \alpha_1 \phi(S_3) \cdot \phi(S_2) &= +1 \\ \alpha_1 \phi(S_1) \cdot \phi(S_3) + \alpha_2 \phi(S_2) \cdot \phi(S_3) + \alpha_1 \phi(S_3) \cdot \phi(S_3) &= +1 \end{aligned} \quad (3.1)$$

$$\phi(S_1) \cdot \phi(S_1) = (a|b|c)(d|e|f) = ((a*d) + (b*e) + (c*f)) \quad (3.2)$$

$$\phi(S_1) \cdot \phi(S_1) = w \cdot b + b, b = 1 \quad (3.3)$$

Ketiga data yang telah dipilih sebagai support vector kemudian disubstitusikan ke dalam Persamaan (3.1). Proses perhitungannya dijabarkan sebagai berikut.

$$\alpha_1 \phi(S_1) \cdot \phi(S_1) + \alpha_2 \phi(S_2) \cdot \phi(S_1) + \alpha_1 \phi(S_3) \cdot \phi(S_1) = (-1|0|1) (-1|0|1) + (2|1|1) (-1|0|1) + (2|-1|1) (-1|0|1)$$

$$\alpha_1 \phi(S_1) \cdot \phi(S_1) + \alpha_2 \phi(S_2) \cdot \phi(S_1) + \alpha_1 \phi(S_3) \cdot \phi(S_1) \Rightarrow 2\alpha_1 - \alpha_2 - \alpha_3 = -1$$

$$\alpha_1 \phi(S_1) \cdot \phi(S_2) + \alpha_2 \phi(S_2) \cdot \phi(S_2) + \alpha_1 \phi(S_3) \cdot \phi(S_2) = (-1|0|1) (2|1|1) + (2|1|1) (2|1|1) + (2|-1|1) (2|1|1)$$

$$\alpha_1 \phi(S_1) \cdot \phi(S_2) + \alpha_2 \phi(S_2) \cdot \phi(S_2) + \alpha_1 \phi(S_3) \cdot \phi(S_2) \Rightarrow -\alpha_1 + 6\alpha_2 + 4\alpha_3 = +1$$

$$\alpha_1 \phi(S_1) \cdot \phi(S_3) + \alpha_2 \phi(S_2) \cdot \phi(S_3) + \alpha_1 \phi(S_3) \cdot \phi(S_3) = (-1|0|1) (2|-1|1) + (2|1|1) (2|-1|1) + (2|-1|1) (2|-1|1)$$

$$\alpha_1 \phi(S_1) \cdot \phi(S_3) + \alpha_2 \phi(S_2) \cdot \phi(S_3) + \alpha_1 \phi(S_3) \cdot \phi(S_3) \Rightarrow -\alpha_1 + 4\alpha_2 + 6\alpha_3 = +1$$

Hasil perhitungan menggunakan Persamaan (3.2) kemudian disubstitusikan ke dalam Persamaan (3.1)

$$2\alpha_1 - \alpha_2 - \alpha_3 = -1$$

$$-\alpha_1 + 6\alpha_2 + 4\alpha_3 = +1$$

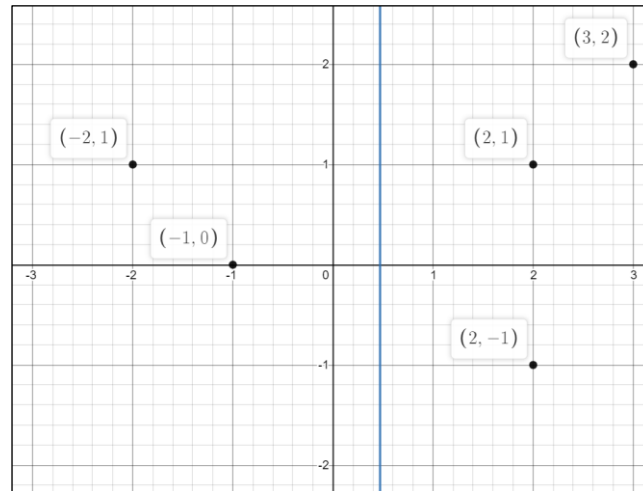
$$-\alpha_1 + 4\alpha_2 + 6\alpha_3 = +1 \quad (3.4)$$

Dengan menggunakan metode substitusi, didapatkan nilai $\alpha_1 = -0.44$, $\alpha_2 = 0.06$ dan $\alpha_3 = 0.06$. Langkah selanjutnya yaitu menghitung offset dan bobot hyperplane dengan menggunakan persamaan (3.5).

$$w = \sum \alpha_i S_i \quad (3.5)$$

$$w = -0,44 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + 0,06 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} - 0,06 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$$

Sehingga diperoleh offset dan bobot hyperplane $y = wx + b$ dengan $w = \begin{pmatrix} 0,68 \\ 0 \end{pmatrix}$ dan $b = -0,32$. Gambar 3.4 menunjukkan dataset yang dipisahkan oleh garis hyperplane.



Gambar 3. 4 Contoh dataset dengan hyperplane

3.5 Pengujian

Dataset SMS yang digunakan dalam penelitian ini diperoleh langsung dari gawai pribadi dan beberapa orang lainnya dengan cara diambil melalui web message. Adapun SMS yang diambil dapat berupa iklan komersil, penawaran dari provider, penipuan, info resmi dari pemerintah, dan lain-lain yang bertujuan untuk mendapatkan data yang lebih beragam. Dataset SMS inilah yang akan diambil teksnya untuk menjadi dataset pada penelitian ini. Dataset pada penelitian ini dibagi menjadi 70% data latih dan 30% data uji.

Pada penelitian ini digunakan kelas yang berbeda. Kategori spam atau tidaknya SMS ditentukan berdasarkan hukum privasi dan karakteristik yang didapatkan berdasarkan penelitian-penelitian mengenai SMS itu sendiri. Terdapat beberapa parameter yang diuji dalam penelitian ini antara lain:

1. Pengaruh metode pembobotan TF-RF terhadap akurasi

Seperti yang telah dipaparkan sebelumnya, pada penelitian digunakan dua jenis pembobotan yaitu secara tradisional dengan TF-IDF dan secara modern dengan TF-RF. Dilakukan perbandingan metode yang mana memiliki performa lebih baik antara keduanya.

2. Pengaruh stemming terhadap akurasi

Pada beberapa penelitian yang ada [9] [17], stemming justru menurunkan akurasi karena dalam prosesnya merubah kata- kata yang ternyata menjadi ciri dari dokumen terkait.

3. Pengujian akurasi berdasarkan jenis-jenis kernel SVM yaitu linear, RBF, dan sigmoid.
4. Pengujian akurasi menggunakan metode klasifikasi k-Nearest Neighbor dengan nilai k adalah 1 hingga 10.
5. Pengujian akurasi menggunakan metode klasifikasi Multinomial Naïve Bayes.

Perhitungan evaluasi dapat menggunakan berbagai cara salah satunya yaitu menggunakan confusion matrix, hasil klasifikasi sistem secara keseluruhan dapat digambarkan antara prediksi dan kelas aktual, serta dapat dicari nilai akurasinya. Contoh data seperti yang tertera Tabel 3.12.

Tabel 3. 12 Confusion Matrix untuk 2 kelas

		Aktual	
		Spam	Non-Spam
Prediksi	Spam	TP(18)	FP(6)
	Non-Spam	FN(9)	TN(27)

Pada tabel confusion matrix diatas, TP (true positive) adalah jumlah kejadian kelas spam diklasifikasikan sebagai kelas- spam, kelas- non spam diklasifikasikan sebagai kelas-non spam. Sedangkan FP (false positive) adalah jumlah kejadian suatu kelas diprediksi sebagai kelas yang tidak sebenarnya seperti kelas- spam diklasifikasikan sebagai kelas- non spam atau kelas- non spam diklasifikasikan sebagai kelas- spam.

Berdasarkan confusion matrix tersebut bisa didapatkan nilai presisi, akurasi dan recall. Nilai akurasi pada klasifikasi mewakili presentasi ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian. Nilai presisi mewakili proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Recall atau sensitivity adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar [9]. Perhitungan nilai akurasi,

presisi dan recall dapat dilihat pada persamaan berikut mengikuti Persamaan (2.12, 2.13, dan 2.14) :

1. Presisi

$$Precision = \frac{18}{18 + 6} = 0.75$$

2. Recall

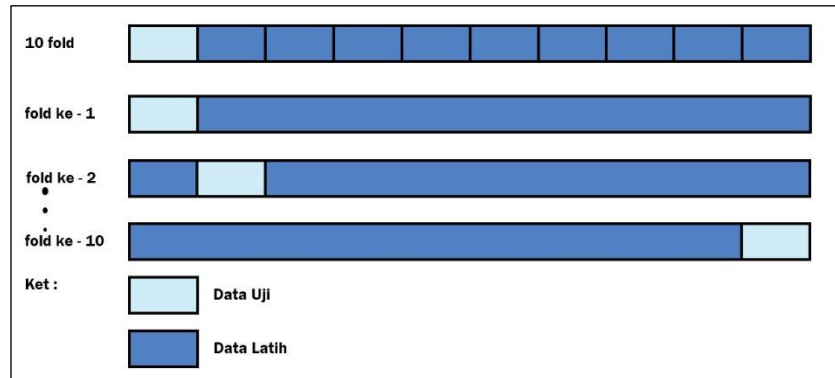
$$Recall = \frac{18}{18 + 9} = 0.67$$

3. Akurasi

$$Accuracy = \frac{18 + 27}{18 + 27 + 6 + 9} = 0.75$$

Evaluasi hasil klasifikasi juga dilakukan dengan menggunakan metode K-fold cross validation. K-fold cross validation merupakan metode untuk mengevaluasi kinerja classifier, metode ini dapat digunakan apabila memiliki jumlah data yang terbatas (jumlah instance tidak banyak). K-fold cross validation adalah suatu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan reduksi dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. K-fold cross validation diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan ukuran yang sama variable Data ke 1, variabel Data ke 2, variabel Data ke 3 .. Dn selanjutnya proses uji dan latih dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data uji dan sisanya akan menjadi data latih. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model [24].

Gambar 3.5 Ilustrasi cross validation 10 fold.



Gambar 3. 5 Ilustrasi cross validation 10 fold
Kinerja dari K-fold cross validation yaitu:

1. Total instance dibagi menjadi N bagian
2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi atau kesamaan atau kedekatan suatu hasil pengukuran dengan angka atau data yang sebenarnya berdasarkan porsi data tersebut. Perhitungan akurasi tersebut menggunakan persamaan sebagai berikut.

$$Akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} \times 100 \quad (3.6)$$

3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai fold ke-k. Hitung rata-rata akurasi dari k buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

BAB IV

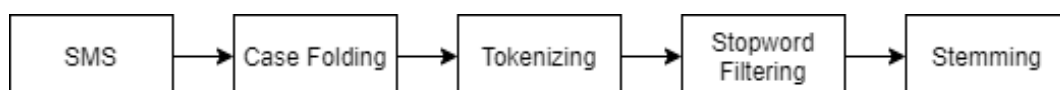
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Pada penelitian ini dilakukan pengumpulan dataset yang bersumber dari 5 orang yang telah bersedia. Dataset tersebut terdiri dari 500 SMS berupa gabungan antara SMS Spam dan non-spam. Pengumpulan SMS diambil dengan cara mem-backup SMS yang ada di smartphone menggunakan aplikasi yang diunduh melalui playstore bernama SMS Backup and Restore, output dari hasil backup tersebut adalah file berekstensi xml, selanjutnya dari file tersebut diambil dari tiap xml beberapa SMS yang dibutuhkan dan disimpan pada file berekstensi csv. Jumlah SMS yang diambil dari tiap xml masing-masing 50 hingga 100 SMS, sehingga total SMS yang digunakan sebagai dataset sejumlah 500 SMS. Jumlah persentase dari SMS Spam sendiri yaitu 175 dan non-spam yaitu 325, ini karena SMS yang diterima lebih banyak dari sisi non-spam, dimana SMS Spam rata-rata kemunculannya 1 hingga 2 SMS per hari tiap bulannya, sehingga persentase SMS yang diterima lebih besar untuk penggunaan secara semestinya. Karakteristik dari SMS Spam ini sendiri untuk kontennya sudah jelas polanya, hanya karakter penulisannya yang beragam, sehingga cukup memfokuskan ke SMS yang lebih sering diterima dan tujuan penggunaan dari SMS itu sendiri yaitu untuk keperluan pribadi.

4.2 Text Preprocessing

Pada penelitian ini, tahapan preprocessing yang dilakukan terhadap seluruh dataset meliputi beberapa tahapan yang terdapat pada Gambar 4.1 Berikut



Gambar 4. 1 Skema proses preprocessing

Tahap pertama yang dilakukan pada SMS merupakan case folding yaitu mengubah seluruh teks pada artikel menjadi huruf kecil agar tidak terjadi

ambiguitas saat membandingkan kata yang diawali dengan huruf besar dan huruf kecil pada term atau kata yang sama, sehingga teks yang dihasilkan seperti pada Gambar.

	teks	jenis		teks	jenis
0	WOW SELAMAT! Ga hanya kuota GRATIS dr pemerint...	non	→	wow selamat! ga hanya kuota gratis dr pemerint...	non
1	10 Desember 2021: Selamat Merayakan Hari Hak A...	non		10 desember 2021: selamat merayakan hari hak a...	non
2	HOT PROMO, Isi ulang skr: Min 50rb cashback 10...	non		hot promo, isi ulang skr: min 50rb cashback 10...	non
3	HotPromo super deal, 25rb dpt 12.12GB, yuk bel...	non		hotpromo super deal, 25rb dpt 12.12gb, yuk bel...	non
4	Hakim ganteng KIM MIN KYU naksir CHORONG APINK...	non		hakim ganteng kim min kyu naksir chorong apink...	non

Before **After**

Gambar 4. 2 Proses case folding dataset SMS

Kemudian dilakukan proses tokenizing, yaitu memecah teks artikel yang terdiri dari kumpulan paragraf serta kalimat menjadi satuan kata atau token, sehingga teks yang dihasilkan seperti pada Gambar.

	teks	jenis		teks	jenis
0	WOW SELAMAT! Ga hanya kuota GRATIS dr pemerint...	non	→	[wow, selamat, ga, kuota, gratis, dr, perintah...	non
1	10 Desember 2021: Selamat Merayakan Hari Hak A...	non		[10, desember, 2021, selamat, raya, hari, hak,...	non
2	HOT PROMO, Isi ulang skr: Min 50rb cashback 10...	non		[hot, promo, isi, ulang, skr, min, 50rb, cashb...	non
3	HotPromo super deal, 25rb dpt 12.12GB, yuk bel...	non		[hotpromo, super, deal, 25rb, dpt, 12, 12gb, y...	non
4	Hakim ganteng KIM MIN KYU naksir CHORONG APINK...	non		[hakim, ganteng, kim, min, kyu, naksir, choron...	non

Before **After**

Gambar 4. 3 Proses tokenizing dataset SMS

Selanjutnya setiap kata yang dihasilkan pada proses tokenizing akan melalui proses stopword filtering untuk menghilangkan kata - kata yang dianggap tidak relevan dan tidak mewakili ciri dari kategori suatu dokumen, hasilnya seperti pada Gambar.

	teks	jenis		teks	jenis
0	WOW SELAMAT! Ga hanya kuota GRATIS dr pemerint...	non	→	wow selamat ga kuota gratis dr perintah bisa a...	non
1	10 Desember 2021: Selamat Merayakan Hari Hak A...	non		10 desember 2021 selamat raya hari hak asasi m...	non
2	HOT PROMO, Isi ulang skr: Min 50rb cashback 10...	non		hot promo isi ulang skr min 50rb cashback 100 ...	non
3	HotPromo super deal, 25rb dpt 12.12GB, yuk bel...	non		hotpromo super deal 25rb dpt 12 12gb yuk beli ...	non
4	Hakim ganteng KIM MIN KYU naksir CHORONG APINK...	non		hakim ganteng kim min kyu naksir chorong apink...	non

Before **After**

Gambar 4. 4 Proses stopword filtering dataset SMS

Tahap terakhir yaitu stemming yang berguna untuk mengubah kata yang berisi imbuhan menjadi kata dasarnya, hasil yang didapatkan pada Gambar.

	teks	jenis		teks	jenis
0	WOW SELAMAT! Ga hanya kuota GRATIS dr pemerint...	non	→	wow selamat ga hanya kuota gratis dr perintah ...	non
1	10 Desember 2021: Selamat Merayakan Hari Hak A...	non		10 desember 2021 selamat raya hari hak asasi m...	non
2	HOT PROMO, Isi ulang skr: Min 50rb cashback 10...	non		hot promo isi ulang skr min 50rb cashback 100 ...	non
3	HotPromo super deal, 25rb dpt 12.12GB, yuk bel...	non		hotpromo super deal 25rb dpt 12 12gb yuk beli ...	non
4	Hakim ganteng KIM MIN KYU naksir CHORONG APINK...	non		hakim ganteng kim min kyu naksir chorong apink...	non

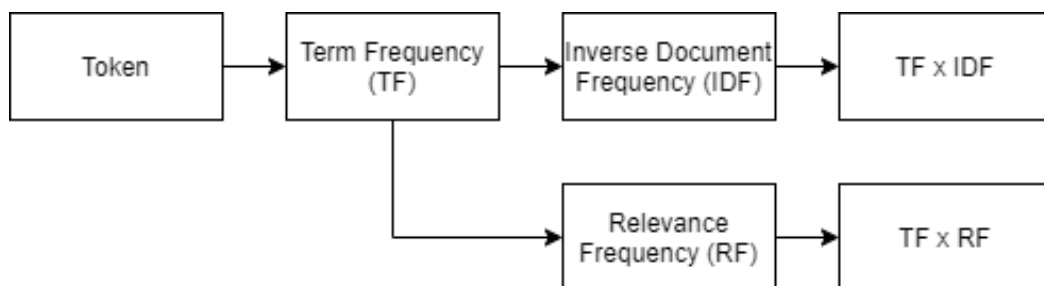
Before **After**

Gambar 4. 5 Proses stemming dataset SMS

Selanjutnya, setiap kata yang dihasilkan melalui semua proses text preprocessing akan disimpan dalam bentuk vector. Vector token atau kata yang telah dibentuk nantinya akan digunakan dalam proses training data.

4.3 Pattern Discovery

Tahap ini bertujuan untuk menentukan pola kata pada suatu SMS seperti menentukan ciri atau fitur pada masing - masing kategori SMS melalui pembobotan pada term atau kata pada SMS. Pada proses ini juga dilakukan training dataset yang akan menghasilkan model training. Berikut ini beberapa tahap yang dilakukan di tahap pattern discovery pada Gambar 4.6.



Gambar 4. 6 Skema proses pattern discovery

Masukan dari pada proses training atau pelatihan ini merupakan vector token atau kata yang dihasilkan pada proses text preprocessing. Vector hasil proses text preprocessing tersebut akan melalui proses pattern discovery, yaitu memberikan bobot setiap kata pada vector. Pada proses pattern discovery, setiap kata pada vector tersebut kemudian disimpan bersama dengan jumlah kemunculan kata tersebut pada setiap dokumen pada masing - masing kelas, nilai tersebut yang kemudian mewakili nilai dari term frequency. Selanjutnya setiap kata pada masing-masing kelas dicari nilai keunikan kata tersebut pada dokumen lain dan nilai term frequency yang tinggi pada suatu dokumen dan dokumen di kelas lainnya yang

merupakan nilai dari inverse document frequency dan nilai dari relevance frequency. Kemudian masing-masing nilai dari inverse document frequency dan relevance frequency pada setiap kata dikalikan dengan nilai term frequency pada masing - masing kata, nilai tersebut akan menghasilkan bobot dari Term Frequency Inverse Document Frequency (TF-IDF) dan Term Frequency Relevance Frequency (TF-RF). Nilai dari TF-IDF dan TF-RF inilah yang akan digunakan sebagai bobot atau fitur pada vector token atau kata yang telah dibentuk, kemudian vector tersebut nantinya akan disimpan sebagai model training. Model training yang telah disimpan pada proses sebelumnya akan digunakan pada proses klasifikasi terhadap data uji yang digunakan.

4.4 Pengujian

Pada penelitian ini terdapat beberapa skenario pengujian yang dilakukan, antara lain sebagai berikut.

1. Pengaruh metode pembobotan TF-RF terhadap akurasi

Seperti yang telah dipaparkan sebelumnya, pada penelitian digunakan dua jenis pembobotan yaitu secara tradisional dengan TF-IDF dan secara modern dengan TF-RF. Dilakukan perbandingan metode yang mana memiliki performa lebih baik antara keduanya.

2. Pengaruh stemming terhadap akurasi

Pada beberapa penelitian yang ada [9] [17], stemming justru menurunkan akurasi karena dalam prosesnya merubah kata- kata yang ternyata menjadi ciri dari dokumen terkait.

3. Pengujian akurasi berdasarkan jenis-jenis kernel SVM yaitu linear, RBF, dan sigmoid.

4. Pengujian akurasi menggunakan metode klasifikasi k-Nearest Neighbor dengan nilai k adalah 1 hingga 10.

5. Pengujian akurasi menggunakan metode klasifikasi Multinomial Naïve Bayes.

Skenario pengujian dijalankan dari skenario pertama hingga skenario terakhir dengan menggunakan Jupyter Notebook, dimana scenario pertama dilakukan untuk melihat akurasi klasifikasi yang dihasilkan jika menggunakan

model training data dari metode pembobotan TF-RF, apakah lebih baik dari TF-IDF atau bahkan lebih buruk. Skenario kedua melihat pengaruh dari proses preprocessing stemming ketika digunakan dan tidak, apakah merubah hasil dari akurasi klasifikasi atau tidak. Skenario ketiga melihat akurasi dari metode SVM jika menggunakan kernel yang berbeda. Lalu scenario keempat dan kelima melihat hasil dari akurasi k-Nearest Neighbour dan Multinomial Naïve Bayes, jika dibandingkan dengan metode Support Vector Machine apakah mendapat hasil yang lebih baik atau sebaliknya, yang manakah diantara ketiganya yang memiliki hasil akurasi paling tinggi.

Untuk menghindari bias dalam pembagian data, maka pada setiap pengujian skenario digunakan K-Fold Cross Validation, dimana pada skenario pertama hingga skenario ketiga digunakan nilai K atau nilai fold sebanyak 10, pembagian dilakukan dengan membagi data train:test dengan perbandingan 9:1. Pada penelitian ini library untuk klasifikasi yang digunakan adalah library SKLearn pada Python.

4.5 Hasil Pengujian

Berdasarkan skenario-skenario yang telah dipaparkan, proses pengujian akan dijelaskan menjadi 2 bagian yaitu pengujian dengan dataset yang melalui proses stemming dan dataset yang tidak melalui proses stemming.

Uji coba ini dilakukan untuk mengetahui jika dengan proses stemming, bagaimana hasil akurasi yang didapatkan dari tiap classifier untuk nantinya akan dibandingkan dengan tanpa proses stemming, yaitu dengan Support Vector Machine dengan 3 jenis kernel (Linear, RBF, dan Sigmoid), k-Nearest Neighbour, dan Multinomial Naïve Bayes, baik menggunakan metode pembobotan TF-IDF ataupun TF-RF. Berdasarkan pengujian menggunakan dataset yang telah dikumpulkan pada penelitian ini, kemudian didapatkan hasil yaitu nilai accuracy, precision, dan recall yang disajikan pada tabel pada masing-masing metode klasifikasi dan perbandingan tiap metode baik metode pembobotan maupun metode klasifikasi.

1. Pengujian dengan Support Vector Machine

Tabel 4. 1 Pengujian dengan SVM Kernel RBF yang di-stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	87.96%	95.60%		98.92%	100%	99.60%
2	100%	88.19%	95.80%		98.59%	100%	99.50%
3	100%	88.56%	95.93%		98.53%	100%	99.47%
4	100%	88.40%	95.90%		98.58%	100%	99.50%
5	100%	88.35%	95.84%		98.12%	100%	99.36%
6	100%	88.08%	95.73%		98.52%	100%	99.40%
7	100%	88.15%	95.77%		98.13%	100%	99.34%
8	100%	88.26%	95.80%		98.16%	100%	99.35%
9	100%	88.17%	95.78%		98.13%	100%	99.36%
10	100%	88.19%	95.80%		97.96%	100%	99.30%
11	100%	88.25%	95.82%		97.99%	100%	99.31%
12	100%	88.17%	95.78%		98.04%	100%	99.32%
13	100%	88.13%	95.77%		98.05%	100%	99.32%
14	100%	88.09%	95.76%		98.08%	100%	99.33%
15	100%	87.96%	95.60%		98.92%	100%	99.60%
Rata-Rata (%)					Rata-Rata (%)		
	100%	88.21%	95.79%		98.27%	100%	99.39%

Tabel 4. 2 Pengujian dengan SVM Kernel RBF tanpa stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	88.38%	95.80%		98.92%	100%	99.60%
2	100%	87.76%	95.60%		98.86%	100%	99.60%
3	100%	87.58%	95.53%		98.96%	100%	99.60%
4	100%	87.34%	95.50%		98.88%	100%	99.60%
5	100%	87.34%	95.48%		98.46%	100%	99.48%
6	100%	87.05%	95.37%		98.54%	100%	99.50%
7	100%	87.00%	95.34%		98.59%	100%	99.51%
8	100%	87.20%	95.40%		98.65%	100%	99.53%
9	100%	87.23%	95.42%		98.63%	100%	99.53%
10	100%	87.18%	95.42%		98.45%	100%	99.48%
11	100%	87.09%	95.38%		98.50%	100%	99.49%
12	100%	87.05%	95.37%		98.53%	100%	99.50%
13	100%	87.03%	95.35%		98.56%	100%	99.51%
14	100%	87.01%	95.36%		98.60%	100%	99.51%
15	100%	88.38%	95.80%		98.92%	100%	99.60%
Rata-Rata (%)					Rata-Rata (%)		
	100%	87.30%	95.45%		98.65%	100%	99.53%

Tabel 4.1 merupakan hasil pengujian dari dataset menggunakan kernel RBF tanpa stemming. Pengujian dilakukan melalui skema 10 fold cross validation yang dilakukan sebanyak 15 kali iterasi, dimana terdapat 2 kolom masing-masing untuk metode pembobotan TF-IDF dan TF-RF. Setelah melalui proses iterasi, didapatkan nilai rata - rata recall, precision serta accuracy pada setiap 10 fold cross validation. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik. Pada Tabel 4.2, proses yang sama juga dilakukan, namun pada dataset tersebut tidak digunakan proses stemming pada preprocessing. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik.

Tabel 4. 3 Pengujian dengan SVM Kernel Linear yang di-stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	98.91%	99.60%		100%	100%	100%
2	100%	98.84%	99.60%		100%	100%	100%
3	100%	98.71%	99.53%		100%	100%	100%
4	100%	98.76%	99.55%		100%	100%	100%
5	100%	98.65%	99.52%		100%	100%	100%
6	100%	98.60%	99.50%		100%	100%	100%
7	100%	98.59%	99.49%		100%	100%	100%
8	100%	98.62%	99.50%		100%	100%	100%
9	100%	98.59%	99.49%		100%	100%	100%
10	100%	98.63%	99.50%		100%	100%	100%
11	100%	98.61%	99.49%		100%	100%	100%
12	100%	98.64%	99.50%		100%	100%	100%
13	100%	98.65%	99.51%		100%	100%	100%
14	100%	98.67%	99.51%		100%	100%	100%
15	100%	98.91%	99.60%		100%	100%	100%
Rata-Rata (%)					Rata-Rata (%)		
	100%	98.68%	99.52%		100%	100%	100%

Tabel 4. 4 Pengujian dengan SVM Kernel Linear tanpa stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	97.75%	99.20%		100%	100%	100%
2	100%	97.90%	99.30%		100%	100%	100%
3	100%	97.92%	99.27%		100%	100%	100%
4	100%	98.00%	99.30%		100%	100%	100%
5	100%	97.97%	99.28%		100%	100%	100%
6	100%	98.14%	99.33%		100%	100%	100%
7	100%	98.20%	99.34%		100%	100%	100%
8	100%	98.22%	99.35%		100%	100%	100%
9	100%	98.17%	99.33%		100%	100%	100%
10	100%	98.20%	99.34%		100%	100%	100%
11	100%	98.15%	99.33%		100%	100%	100%
12	100%	98.13%	99.32%		100%	100%	100%
13	100%	98.19%	99.34%		100%	100%	100%
14	100%	98.12%	99.31%		100%	100%	100%
15	100%	97.75%	99.20%		100%	100%	100%
Rata-Rata (%)					Rata-Rata (%)		
	100%	98.08%	99.31%		100%	100%	100%

Tabel 4.3 merupakan hasil pengujian dari dataset menggunakan kernel Linear tanpa stemming. Pengujian dilakukan melalui skema 10 fold cross validation yang dilakukan sebanyak 15 kali iterasi, dimana terdapat 2 kolom masing-masing untuk metode pembobotan TF-IDF dan TF-RF. Setelah melalui proses iterasi, didapatkan nilai rata - rata recall, precision serta accuracy pada setiap 10 fold cross validation. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik. Pada Tabel 4.4, proses yang sama juga dilakukan, namun pada dataset tersebut tidak digunakan proses stemming pada preprocessing. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik.

Tabel 4. 5 Pengujian dengan SVM Kernel Sigmoid yang di-stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	98.91%	99.60%		100%	100%	100%
2	100%	98.94%	99.60%		100%	100%	100%
3	100%	98.89%	99.60%		100%	100%	100%
4	100%	99.05%	99.65%		100%	100%	100%
5	100%	98.92%	99.60%		100%	100%	100%
6	100%	98.93%	99.60%		99.91%	100%	99.97%
7	100%	98.96%	99.60%		99.92%	100%	99.97%
8	100%	99.01%	99.63%		99.85%	100%	99.95%
9	100%	98.93%	99.60%		99.87%	100%	99.96%
10	100%	98.93%	99.60%		99.84%	100%	99.94%
11	100%	98.88%	99.58%		99.80%	100%	99.93%
12	100%	98.89%	99.58%		99.82%	100%	99.93%
13	100%	98.88%	99.58%		99.78%	100%	99.92%
14	100%	98.89%	99.59%		99.80%	100%	99.93%
15	100%	98.91%	99.60%		100%	100%	100%
Rata-Rata (%)					Rata-Rata (%)		
	100%	98.93%	99.60%		99.90%	100%	99.96%

Tabel 4. 6 Pengujian dengan SVM Kernel Sigmoid tanpa stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100%	98.91%	99.60%		100%	100%	100%
2	100%	99.18%	99.70%		100%	100%	100%
3	100%	99.28%	99.73%		100%	100%	100%
4	100%	99.01%	99.65%		100%	100%	100%
5	100%	98.88%	99.60%		100%	100%	100%
6	100%	98.91%	99.60%		100%	100%	100%
7	100%	98.93%	99.60%		100%	100%	100%
8	100%	98.94%	99.60%		100%	100%	100%
9	100%	98.86%	99.58%		100%	100%	100%
10	100%	98.92%	99.60%		100%	100%	100%
11	100%	98.87%	99.58%		100%	100%	100%
12	100%	98.88%	99.58%		100%	100%	100%
13	100%	98.93%	99.60%		100%	100%	100%
14	100%	98.97%	99.61%		100%	100%	100%
15	100%	98.91%	99.60%		100%	100%	100%
Rata-Rata (%)					Rata-Rata (%)		
	100%	98.96%	99.62%		100%	100%	100%

Tabel 4.5 merupakan hasil pengujian dari dataset menggunakan kernel Sigmoid tanpa stemming. Pengujian dilakukan melalui skema 10 fold cross

validation yang dilakukan sebanyak 15 kali iterasi, dimana terdapat 2 kolom masing-masing untuk metode pembobotan TF-IDF dan TF-RF. Setelah melalui proses iterasi, didapatkan nilai rata - rata recall, precision serta accuracy pada setiap 10 fold cross validation. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik. Pada Tabel 4.6, proses yang sama juga dilakukan, namun pada dataset tersebut tidak digunakan proses stemming pada preprocessing. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik.

2. Pengujian dengan k-Nearest Neighbour

Tabel 4. 7 Pengujian dengan k-Nearest Neighbour yang di-stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100.00%	39.71%	78.60%		100%	77.22%	91.80%
2	100.00%	38.26%	78.30%		100%	76.87%	91.90%
3	100.00%	38.63%	78.47%		100%	76.54%	91.87%
4	100.00%	38.17%	78.40%		100%	77.02%	92.00%
5	100.00%	38.37%	78.44%		100%	77.14%	92.08%
6	100.00%	38.83%	78.53%		100%	77.09%	92.10%
7	100.00%	38.71%	78.51%		100%	77.16%	92.11%
8	100.00%	38.51%	78.48%		100%	76.82%	91.97%
9	100.00%	38.02%	78.33%		100%	77.07%	92.00%
10	100.00%	38.27%	78.40%		100%	77.26%	92.02%
11	100.00%	38.30%	78.42%		100%	77.28%	92.00%
12	100.00%	38.46%	78.43%		100%	77.36%	92.02%
13	100.00%	38.58%	78.46%		100%	77.38%	92.02%
14	100.00%	38.58%	78.46%		100%	77.26%	92.00%
15	100.00%	39.71%	78.60%		100%	77.22%	91.80%
Rata-Rata (%)					Rata-Rata (%)		
	100.00%	38.53%	78.45%		100%	77.11%	91.99%

Tabel 4. 8 Pengujian dengan k-Nearest Neighbour tanpa stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	100.00%	35.56%	77.20%		100%	32.30%	90.80%
2	100.00%	36.05%	77.50%		100%	32.18%	91.20%
3	100.00%	36.52%	77.67%		100%	31.57%	90.87%
4	100.00%	36.10%	77.60%		100%	31.76%	91.05%
5	100.00%	36.06%	77.56%		100%	32.29%	91.04%
6	100.00%	36.51%	77.67%		100%	31.96%	91.03%
7	100.00%	36.53%	77.69%		100%	31.89%	91.00%
8	100.00%	36.41%	77.68%		100%	32.06%	90.90%
9	100.00%	36.12%	77.62%		100%	31.92%	90.82%
10	100.00%	36.53%	77.76%		100%	31.82%	90.86%
11	100.00%	36.56%	77.78%		100%	31.92%	90.84%
12	100.00%	36.52%	77.75%		100%	31.72%	90.88%
13	100.00%	36.66%	77.78%		100%	31.68%	90.86%
14	100.00%	36.65%	77.79%		100%	31.71%	90.86%
15	100.00%	35.56%	77.20%		100%	32.30%	90.80%
Rata-Rata (%)					Rata-Rata (%)		
	100.00%	36.34%	77.65%		100%	31.91%	90.93%

Tabel 4.7 merupakan hasil pengujian dari dataset menggunakan k-Nearest Neighbour tanpa stemming. Pengujian dilakukan melalui skema 10 fold cross validation yang dilakukan sebanyak 15 kali iterasi, dimana terdapat 2 kolom masing-masing untuk metode pembobotan TF-IDF dan TF-RF. Setelah melalui proses iterasi, didapatkan nilai rata - rata recall, precision serta accuracy pada setiap 10 fold cross validation. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik. Pada Tabel 4.8, proses yang sama juga dilakukan, namun pada dataset tersebut tidak digunakan proses stemming pada preprocessing. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik.

3. Pengujian dengan Multinomial Naïve Bayes

Tabel 4. 9 Pengujian dengan Multinomial Naïve Bayes yang di-stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	90.18%	98.41%	95.40%		91.85%	98.92%	96.80%
2	90.67%	98.05%	95.60%		92.52%	98.96%	97.00%
3	90.67%	98.02%	95.73%		92.72%	98.89%	97.00%
4	90.46%	97.91%	95.55%		92.65%	98.88%	97.00%
5	90.60%	97.87%	95.60%		92.80%	98.88%	97.04%
6	90.42%	97.87%	95.50%		92.82%	98.90%	97.03%
7	90.33%	97.87%	95.49%		92.84%	98.88%	97.00%
8	89.55%	97.87%	95.33%		92.87%	98.87%	96.98%
9	89.63%	97.85%	95.20%		92.92%	98.87%	97.00%
10	89.61%	97.86%	95.18%		92.99%	98.89%	97.00%
11	89.54%	97.88%	95.16%		93.03%	98.85%	97.00%
12	89.64%	97.93%	95.22%		93.00%	98.84%	97.00%
13	89.74%	97.92%	95.26%		93.08%	98.85%	97.03%
14	89.74%	97.92%	95.26%		93.10%	98.83%	97.01%
15	90.18%	98.41%	95.40%		91.85%	98.92%	96.80%
Rata-Rata (%)					Rata-Rata (%)		
	90.06%	97.95%	95.39%		92.80%	98.88%	96.99%

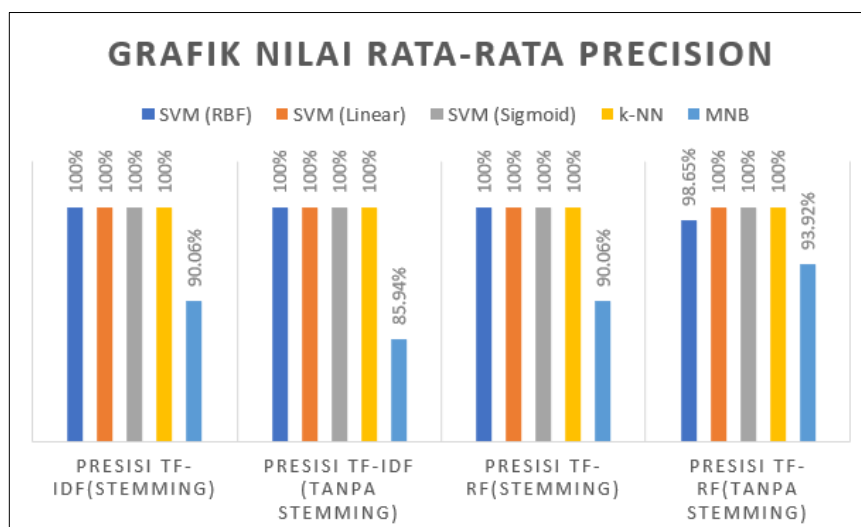
Tabel 4. 10 Pengujian dengan Multinomial Naïve Bayes tanpa stemming

No.	TF-IDF				TF-RF		
	Presisi	Recall	Akurasi		Presisi	Recall	Akurasi
1	84.84%	98.91%	93.20%		93.38%	98.36%	97.00%
2	85.92%	99.18%	93.90%		93.65%	98.64%	97.20%
3	86.36%	99.11%	94.00%		94.03%	98.84%	97.27%
4	86.32%	99.09%	93.95%		93.97%	98.41%	72.50%
5	86.47%	99.05%	94.08%		93.85%	98.50%	97.24%
6	86.30%	99.05%	94.00%		93.83%	98.49%	97.20%
7	86.22%	99.12%	94.00%		93.92%	98.52%	97.23%
8	86.05%	99.16%	93.92%		93.98%	98.55%	97.25%
9	85.72%	99.19%	93.80%		94.01%	98.52%	97.27%
10	85.75%	99.21%	93.82%		94.07%	98.57%	97.30%
11	85.63%	99.23%	93.78%		94.06%	98.56%	97.29%
12	85.67%	99.19%	93.80%		94.01%	98.53%	97.28%
13	85.90%	99.22%	93.92%		94.01%	98.52%	97.28%
14	86.04%	99.23%	93.97%		94.07%	98.52%	97.30%
15	84.84%	98.91%	93.20%		93.38%	98.36%	97.00%
Rata-Rata (%)					Rata-Rata (%)		
	85.94%	99.14%	93.87%		93.92%	98.54%	95.47%

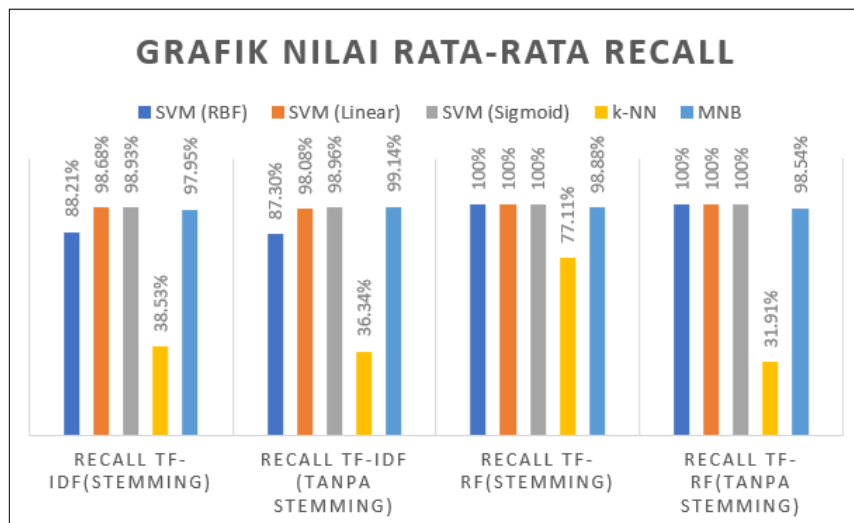
Tabel 4.9 merupakan hasil pengujian dari dataset menggunakan kernel Linear tanpa stemming. Pengujian dilakukan melalui skema 10 fold cross validation yang dilakukan sebanyak 15 kali iterasi, dimana terdapat 2 kolom masing-masing untuk metode pembobotan TF-IDF dan TF-RF. Setelah melalui proses iterasi, didapatkan nilai rata - rata recall, precision serta accuracy pada setiap 10 fold cross validation. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik. Pada Tabel 4.10, proses yang sama juga dilakukan, namun pada dataset tersebut tidak digunakan proses stemming pada preprocessing. Berdasarkan pengujian tersebut, dengan rata-rata nilai presisi, recall, dan accuracy yang didapatkan antara metode TF-IDF dengan metode TF-RF, bahwa metode TF-RF memiliki hasil yang lebih baik.

4. Analisis hasil pengujian tiap metode pembobotan dan klasifikasi

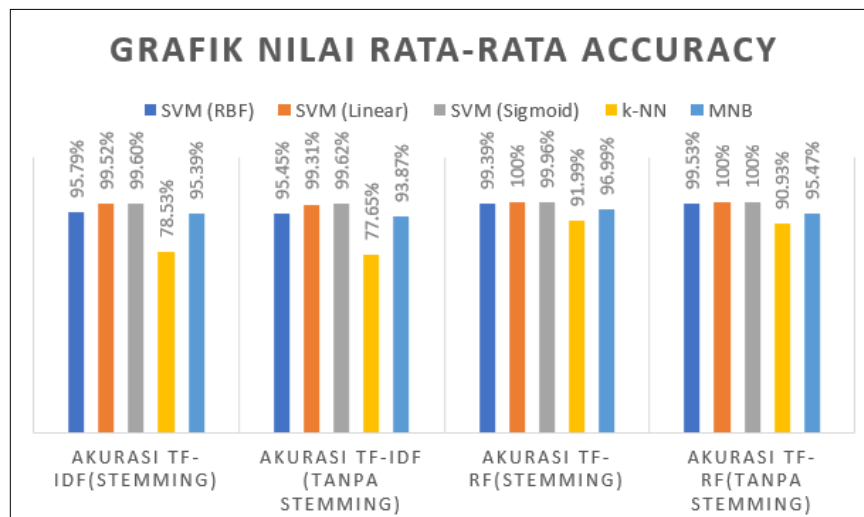
Berdasarkan hasil pengujian menggunakan beberapa skema pengujian yang menggunakan metode pembobotan dan klasifikasi yang berbeda-beda serta penggunaan stemming dan tanpa stemming, didapatkan beberapa nilai rata-rata dari precision, recall serta accuracy dari masing-masing percobaan. Berikut ini ditampilkan diagram perbandingan nilai rata-rata hasil pengujian 10 fold cross validation yang dilakukan pada masing-masing skema pengujian dengan stemming dan tanpa stemming.



Gambar 4. 7 Diagram perbandingan nilai precision



Gambar 4. 8 Diagram perbandingan nilai recall



Gambar 4. 9 Diagram perbandingan nilai accuracy

Berdasarkan diagram tersebut, diketahui bahwa nilai precision, recall, dan accuracy memiliki kecenderungan lebih tinggi didapatkan dengan melalui proses stemming. Berdasarkan hasil akurasi pada masing-masing metode diatas, diketahui untuk metode SVM pada masing-masing kernelnya jika dibandingkan bahwa SVM dengan kernel linear dan sigmoid memiliki nilai akurasi tertinggi dengan nilai akurasi, presisi, dan recall yang sama. Kernel RBF merupakan fungsi kernel yang digunakan ketika data tidak terpisah secara linear, sedangkan Kernel Linear merupakan fungsi kernel yang baik digunakan ketika data sudah terpisah secara linear, sehingga bisa disimpulkan dataset telah terbagi secara linear sehingga akurasi yang didapatkan lebih tinggi dengan Kernel Linear. Kedua kernel tersebut

juga secara keseluruhan memiliki akurasi paling tinggi diantara metode klasifikasi yang lainnya yaitu k-Nearest Neighbour dan Multinomial Naïve Bayes. Namun secara rata-rata nilai precision, recall, dan accuracy pada kedua metode pembobotan serta stemming dan tanpa stemming, Kernel Sigmoid menjadi fungsi kernel terbaik dengan perbedaan nilai presisi sebesar 0% dengan Linear dan 0.34% dengan RBF, nilai recall sebesar 0.28% dengan Linear dan 5.59% dengan RBF, dan nilai accuracy sebesar 0.09% dengan Linear dan 2.26% dengan RBF.

Begitu juga dibandingkan dengan metode k-Nearest Neighbour, nilai k terbaik yang digunakan yaitu k=1. Hasil akurasi yang didapatkan sekalipun dengan nilai k terbaik masih tidak lebih baik jika dibandingkan dengan SVM Kernel Sigmoid dan kernel yang lainnya, walaupun terdapat peningkatan ketika menggunakan metode pembobotan TF-RF. Menurut penelitian-penelitian terkait, ini bisa dikarenakan metode ini hanya dapat bekerja lebih baik pada dataset yang non-linear dan data yang tidak sedikit. Secara rata-rata nilai precision, recall, dan accuracy pada kedua metode pembobotan serta stemming dan tanpa stemming, dibandingkan dengan SVM Kernel Sigmoid, perbedaan nilai presisi sebesar 0%, nilai recall sebesar 53.5%, dan nilai accuracy sebesar 15.02%. Berbeda dengan metode Multinomial Naïve Bayes, hasil akurasi yang didapatkan sangat tinggi, bersaing dengan SVM tetapi masih lebih baik SVM Kernel Sigmoid untuk di masing-masing metode pembobotan. Menurut penelitian-penelitian terkait, dikarenakan metode ini tidak membutuhkan data latih yang banyak dan dilakukan proses perhitungan nilai probabilitas pada setiap kata, dimana proses ini akan menghasilkan sebuah kata pada setiap dokumen yang mengkarakteristikan dokumen pada suatu kategori tertentu, sehingga proses training data dilakukan secara optimal sebelum klasifikasi seperti SVM. Secara rata-rata nilai precision, recall, dan accuracy pada kedua metode pembobotan serta stemming dan tanpa stemming, dibandingkan dengan SVM Kernel Sigmoid, perbedaan nilai presisi sebesar 10.01%, nilai recall sebesar 0.84%, dan nilai accuracy sebesar 4.37%

Metode pembobotan yang lebih baik diantara TF-IDF dengan TF-RF berdasarkan hasil akurasi tiap metode yaitu TF-RF. Rata-rata TF-RF meningkatkan performa dalam hal presisi sebesar 0.66%, nilai recall sebesar 6.43%, dan nilai

accuracy sebesar 3.96% dibandingkan dengan menggunakan TF-IDF. Walaupun keduanya sama-sama fokus kepada kemunculan kata, namun TF-RF juga mempertimbangkan kemunculan kata berdasarkan kategori tertentu tidak seperti TF-IDF yang tanpa melihat dokumen tersebut kategorinya apa sehingga nilai akurasi untuk TF-RF selalu lebih tinggi daripada TF-IDF.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang sudah didapatkan, dapat disimpulkan bahwa :

1. Metode term frequency relevance frequency (TF-RF) memiliki hasil rata-rata precision, recall, dan accuracy yang lebih tinggi dibandingkan dengan term frequency inverse document frequency (TF-IDF) yang dikombinasikan dengan tiap metode klasifikasi yang digunakan pada penelitian ini, dengan perbedaan rata-rata precision sebesar 0.66%, recall sebesar 6.43%, dan accuracy sebesar 3.96%.
2. Metode klasifikasi Support Vector Machine Kernel Sigmoid memiliki hasil rata-rata akurasi tertinggi pada tiap metode pembobotan dengan stemming dan tanpa stemming dibandingkan dengan tiap metode klasifikasi yang digunakan pada penelitian ini, dimana perbedaan rata-rata akurasi dengan Kernel RBF sebesar 2.26%, Kernel Linear sebesar 0.09%, k-Nearest Neighbour sebesar 15.02%, dan Multinomial Naïve Bayes 4.37%.
3. Skema pengujian yang melalui tahapan stemming cenderung mendapatkan hasil yang lebih tinggi dibandingkan dengan pengujian yang tidak melalui tahapan stemming. Hal tersebut disebabkan karena imbuhan pada setiap suku kata yang sama justru menambah variasi fitur saat pemrosesan pembobotan sehingga memberikan nilai bobot yang berbeda-beda.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan, berikut beberapa saran perbaikan ataupun pengembangan yang dapat dilakukan pada penelitian kedepannya :

1. Untuk mendapatkan model yang lebih baik lagi, jumlah data yang digunakan tentunya berpengaruh, disarankan untuk menambah dataset agar data lebih bervariasi sehingga sistem dapat mempelajari lebih banyak variasi SMS dan

tidak menimbulkan bias terhadap salah satu jenis SMS.

2. Mencoba membandingkan metode ekstraksi fitur selain dari TF-IDF dan TF-RF dikarenakan kedua metode tersebut hanya focus pada kemunculan kata tertentu saja tanpa mempertimbangkan ketidakmunculan kata tersebut.

DAFTAR PUSTAKA

- [1] N. Zakiah, "Cara Menyingkirkan SMS Spam, supaya Gak Merasa Terganggu Lagi," <https://www.idntimes.com/tech/trend/nena-zakiah-1/cara-stop-sms-spam/4>, 2020.
- [2] Okezone, "Apa Itu SMS Spam?," <https://techno.okezone.com/read/2020/01/25/207/2158113/apa-itu-sms-spam>, 2020.
- [3] Kominfo, "SMS Spam Dilarang, SMS Iklan Buka Peluang," https://kominfo.go.id/content/detail/1825/sms-spam-dilarang-sms-iklan-buka-peluang/0/sorotan_media, 2012.
- [4] CNN, "Banyak SMS Sampah, Ombudsman Kritik Registrasi Prabayar," <https://www.cnnindonesia.com/teknologi/20190815135828-185-421609/banyak-sms-sampah-ombudsman-kritik-registrasi-prabayar>, 2019.
- [5] M. . Imelda A.Muis & Muhammad Affandes, "Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet," *Sains, Teknol. dan Ind. Sultan Syarif Kasim Riau*, vol. 12, no. 2, pp. 189–197, 2015.
- [6] I. Munitasri, S. Santosa, and C. Supriyanto, "Klasifikasi Pesan Sms Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Genetic Algorithm," *J. Teknol. Inf.*, vol. 14, no. 1, 2018, [Online]. Available: <http://research>.
- [7] G. A. Sandag, R. J. Sambur, and J. Bororing, "Klasifikasi Sms Spam Menggunakan Support Vector Machine," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, 2019, doi: 10.33480/pilar.v15i2.693.
- [8] M. F. Asshiddiqi and K. M. Lhaksmana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," *Univ. Telkom*, vol. 5, no. 3, pp. 177–178, 2020.
- [9] I. D. Putra, "Klasifikasi Artikel Berdasarkan Tingkatan Umur Pembaca menggunakan Metode Multinomial Naive Bayes Classifier," no. November, 2019.
- [10] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A study on term weighting for text categorization: A novel supervised variant of tf.idf," *DATA 2015 - 4th Int. Conf. Data Manag. Technol. Appl. Proc.*, pp. 26–37, 2015, doi: 10.5220/0005511900260037.
- [11] A. T. Ni'mah and A. Z. Arifin, "Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis," *Rekayasa*, vol. 13, no. 2, pp. 172–180, 2020, doi: 10.21107/rekayasa.v13i2.6412.
- [12] H. Tantyoko, Adiwijaya, and U. N. Wisesty, "Perbandingan Pembobotan untuk Klasifikasi Topik Berita menggunakan Decision Tree," *J. Teknol. APERTI BUMN*, vol. 2, pp. 97–113, 2019.
- [13] A. N. Assidyk *et al.*, "Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada Trending Topic di Twitter dengan Menggunakan Klasifikasi K-Nearest Neighbor," *Univ. Telkom*, vol. 7, no. 2, pp. 7773–7781, 2020.
- [14] T. Rahmatullah, "Perlindungan Hukum Terhadap Privacy dari Spamming Berdasarkan Undang-Undang No. 11 Tahun 2008 Tentang Informasi dan Transaksi Elektronik," *Media Justitia Nusantara*, vol. 1, no. 11, pp. 102–123, 2015.
- [15] N. S. Dhuha, "Klasifikasi Teks Pengaduan Sambat Online Menggunakan Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, 2020.
- [16] A. Sabrani, I. G. Putu, W. Wedashwara, and F. Bimantoro, "METODE MULTINOMIAL NAÏVE BAYES UNTUK KLASIFIKASI ARTIKEL ONLINE

- TENTANG GEMPA DI INDONESIA (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia),” vol. 2, no. 1, pp. 89–100, 2020.
- [17] M. Lan, “A New Term Weighting Method for Text Categorization,” *Natl. Univ. Singapore*, 2006.
 - [18] S. J. Delany, M. Buckley, and D. Greene, “SMS spam filtering: Methods and data,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9899–9908, 2012, doi: 10.1016/j.eswa.2012.02.053.
 - [19] O. Abayomi-Alli, S. Misra, A. Abayomi-Alli, and M. Odusami, “A review of soft techniques for SMS spam classification: Methods, approaches and applications,” *Eng. Appl. Artif. Intell.*, vol. 86, no. August, pp. 197–212, 2019, doi: 10.1016/j.engappai.2019.08.024.
 - [20] P. Navaney, G. Dubey, and A. Rana, “SMS Spam Filtering Using Supervised Machine Learning Algorithms,” *Proc. 8th Int. Conf. Conflu. 2018 Cloud Comput. Data Sci. Eng. Conflu. 2018*, pp. 43–48, 2018, doi: 10.1109/CONFLUENCE.2018.8442564.
 - [21] A. I. Kadhim, “Survey on supervised machine learning techniques for automatic text classification,” *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.
 - [22] H. N. Irmanda and R. Astriratma, “Klasifikasi Jenis Pantun dengan Metode Support Vector Machines (SVM),” *RESTI*, vol. 1, no. 10, 2021.
 - [23] K. Akromunnisa and R. Hidayat, “KLASIFIKASI DOKUMEN TUGAS AKHIR (SKRIPSI) MENGGUNAKAN K-NEAREST NEIGHBOR,” *JISKa*, vol. 4, no. 1, pp. 69–75, 2019.
 - [24] S. Hulu, “ANALISIS KINERJA METODE CROSS VALIDATION DAN K-NEAREST NEIGHBOR DALAM KLASIFIKASI DATA,” *Talent. Publ.*, p. 95, 2020.