Teknik Bagging Pada Algoritma Klasifikasi Decision Tree dan SVM Untuk Klasifikasi SMS Berbahasa Indonesia

M. Ibnu Umar Rosyidi[1], Naim Rochmawati[2]

[1,2] Jurusan Teknik Informatika, Fakultas Teknik, Universitas Negeri Surabaya
[1]m.ibnu.19048@mhs.unesa.ac.id
[2]naimrochmawati@unesa.ac.id

***Abstrak*— Perkembangan teknologi di berbagai sektor yang sangat cepat selama satu decade ini, membuat kita semakin dimudahkan dalam melakukan aktivitas sehari-hari. Dari bertukar surat melalui burung merpati dengan jangkauan terbatas hingga dapat melakukan panggilan video di seluruh dunia dengan selisih waktu yang hampir secara *realtime*. Pada tahun 2000-an adalah masa dimana mulai ada SMS (Short Message Service). SMS dengan cepat menjadi sarana komunikasi tidak langsung yang populer, situasi ini dimanfaatkan orang tidak bertanggung jawab untuk melakukan kegiatan melanggar hukum seperti penipuan. Untuk mengurangi korban penipuan SMS, perlu untuk menerapkan filter SMS agar tidak semua SMS masuk ke pengguna, salah satu caranya adalah dengan melakukan klasifikasi dan prediksi dari SMS yang masuk apakah SMS tersebut mengandung penipuan atau tidak. Teknik yang biasa dipakai dalam klasifikasi adalah *decision tree* dan SVM, pada penelitian ini juga akan digunakan teknik *ensemble* yang dapat meningkatkan kinerja dari algoritma yang digunakan yaitu teknik bagging, data yang digunakan adalah dataset SMS dari penelitan Rahmi dan Wibisono[1]. Penggunaan teknik bagging memiliki pengaruh yang signifikan pada peningkatan nilai akurasi algoritma *decision tree* dan SVM, *decision tree* mengalami kenaikan nilai akurasi sebanyak 5% dari 86% menjadi 91% menggunakan data unigram tanpa TF-IDF pada uji 5-fold, algoritma SVM tidak mengalami peningkatan nilai akurasi yang signifikan saat diterapkan teknik bagging.**

***Kata Kunci*— Bagging, Klasifikasi, SMS, *Decision Tree*, SVM, TF-IDF, N-Gram**

I. PENDAHULUAN

Perkembangan teknologi diberbagai sektor yang sangat cepat selama satu dekade ini, membuat kita semakin dimudahkan dalam melakukan aktivitas harian. Dari bertukar surat melalui burung merpati dengan jangkauan terbatas hingga dapat melakukan panggilan video di seluruh dunia dengan selisih waktu yang hampir realtime.

Tahun 1990-an telepon koin masih sangat digemari oleh masyarakat, selain harganya murah dan mudah dijumpai hampir setiap tempat bahkan ada warung telepon yang memang memiliki banyak unit telepon yang siap digunakan bagi para pelanggan. Telepon koin ini cepat sekali menggantikan surat menyurat pada kala itu, meskipun surat menyurat masih dilakukan jika memang kita ingin mengundang atau memberi suatu informasi secara resmi, tetapi melalui telepon kita bisa mendengarkan suara orang yang kita hubungi secara langsung.

Tahun 2000-an adalah masa dimana mulai bermunculan telepon genggam. Masyarakat sudah memiliki telepon masing masing, sehingga menjadi lebih mudah untuk berkomunikasi. Telepon genggam juga memiliki fitur lain selain komunikasi secara langsung menggunakan suara, yaitu melalui pesan text atau yang biasa disebut SMS (Short Message Service). SMS diperkenalkan di eropa pada tahun 1992, SMS terintegrasi dengan GSM yang setelah itu berubah menjadi CDMA[2]. SMS dengan cepat menjadi sarana berkomunikasi tidak langsung yang populer, karena harganya yang lebih murah daripada telepon, SMS juga bisa membuat penerima pesan tidak langsung membalas pesan tersebut jika dalam keadaan yang sibuk, belum lagi pengirim sms ada di dalam area yang memang tidak boleh berisik seperti rumah sakit dan tempat lainnya. Namun, dengan tarif SMS yang murah dan ditambah lagi banyaknya bonus yang diberikan dalam penggunaan SMS pada situasi tertentu menyebabkan adanya pihak yang tidak bertanggung jawab yang dapat melakukan tindakan kriminal penipuan melalui pesan SMS.

SMS spam tidak semuanya mengandung penipuan terhadap penerima SMS, layanan jasa atau produk mengadakan promo untuk beberapa pelanggan dengan kriteria tertentu. Tentunya kita tidak akan mau melewatkan promo seperti itu, karena akan lebih menguntungkan jika bisa mendapatkan promo. Salah satu bentuk penipuan melalui SMS yaitu pelaku akan memberi pesan bahwa korban mendapatkan hadiah dan pelaku akan meminta sejumlah uang kepada korban agar hadiah dapat dimiliki, setelah mengirim sejumlah uang kepada pelaku dan melakukan panggilan telepon biasanya tidak dapat dihubungi kembali[3]. Karena itu, masyarakat pengguna layanan SMS mengalami kerugian. Salah satu cara untuk mengatasi SMS spam adalah dengan mengkategorikannya menjadi SMS normal, promosi, dan fraud (penipuan) sehingga tidak semua pesan masuk ke pengguna. Klasifikasi ini dilakukan untuk membedakan pesan menurut kelasnya, apakah pesan tersebut masuk kedalam kategori normal, promo atau fraud. Sehingga pengguna layanan SMS tidak dirugikan dengan adanya SMS spam.

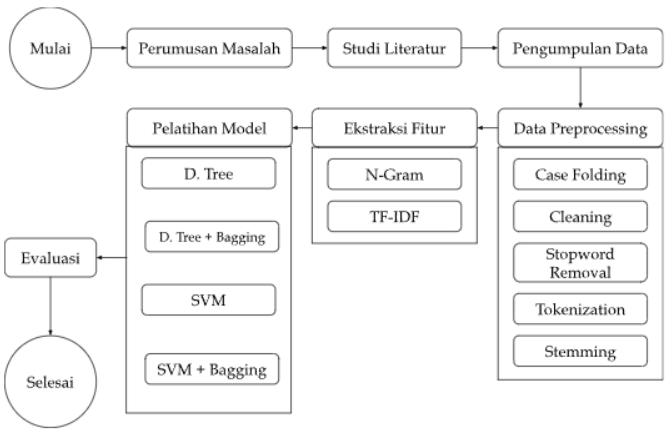
Klasifikasi adalah teknik untuk mengkategorikan jenis label atau class berdasarkan fitur atau atribut yang dimiliki oleh data, teknik klasifikasi bertujuan untuk menemukan suatu fungsi keputusan yang secara tepat memprediksi kelas dari data uji yang berasal dari fungsi distribusi, sama halnya dengan data latih[4]. Banyak algoritma yang dapat digunakan untuk masalah klasifikasi diantaranya Support Vector Machine (SVM) dan Decision Tree. Selain metode klasifikasi yang telah disebutkan tadi, juga ada metode ensemble, yaitu metode gabungan dari beberapa model untuk menghasilkan model dengan performa yang lebih baik diantara metode ensemble

yang sering dipakai adalah bagging. Bagging adalah teknik untuk meningkatkan performa dari model klasifikasi dan mengatasi overfitting. Kekurangan dari bagging adalah mahalnya biaya komputasi dikarenakan model dilatih secara paralel dan juga bisa terjadi under-fitting jika model tidak dibuat dengan baik.

Beberapa penelitian sudah dilakukan untuk mengatasi masalah klasifikasi SMS yaitu penelitian oleh Devi[5], menggunakan 4 algoritma yang akan diuji mengklasifikasikan SMS, algoritma yang digunakan diantaranya Support Vector Machine (SVM), Naive Bayes, Random Forest dan Bagging Classifier. Penelitian yang dilakukan oleh Devi[5] belum mencantumkan parameter apa yang dipakai dalam metode bagging, jika tidak dicantumkan maka parameter default akan digunakan yaitu decision tree. Hasilnya Bagging Classifier menjadi algoritma dengan nilai akurasi tertinggi sebesar 97.4%. Dari penelitian yang dilakukan oleh Edi[6], performa dari algoritma tersebut masih dapat ditingkatkan menggunakan ensemble learning. Ensemble learning adalah gabungan dari beberapa pembelajaran mesin untuk meningkatkan performa sistem secara keseluruhan[7]. Studi empiris telah menunjukkan bahwa masalah klasifikasi dan regresi menggunakan ensemble learning seringkali menghasilkan performa model yang akurat daripada model base individual[8, 9, 10]. Salah satu metode boosting yang sering digunakan adalah bagging. Teknik bagging, boosting dan random subspace method (RSM) di desain dan biasanya diaplikasikan pada Decision Tree[11]. Selain dari pemaparan tersebut mengapa penelitian ini menggunakan data sms, karena data sms tersedia dan berlisensi creative common. Perbedaan penelitian ini dengan yang sebelumnya adalah digunakannya 2 parameter *base model* yang berbeda yaitu *decision tree* dan SVM untuk diterapkan Teknik bagging sehingga akan didapatkan 4 model akan dihasilkan dari penelition ini.

II. METODE PENELITIAN

Dalam melakukan penelitian, diperlukan pedoman sebagai acuan agar penelitian sesuai dengan tujuan yang diharapkan sehingga dapat menjawab rumusan masalah yang sudah ditentukan.



Gbr. 1 Gambaran Umum Penelitian

Berdasarkan Gambar 1. Gambaran umum penelitian, dapat diketahui langkah langkah yang akan diterapkan dalam penelitian ini, yaitu:

A. *Pengumpulan Data*

Dataset yang digunakan dalam penelitian ini adalah dataset sekunder yang didapat dari penelitian Rahmi dan Wibisono[1]. Dataset berjumlah 1143 pesan SMS dengan 3 label kelas dengan rincian label terdapat pada table berikut.

TABEL I
KETERANGAN LABEL DATASET

Label	Keterangan
0	SMS Normal
1	SMS *Fraud* atau Penipuan
2	SMS Promo

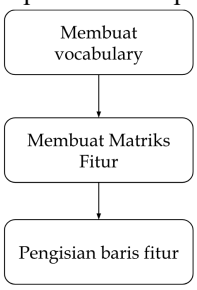
B. *Pra Pemrosesan Data*

Sebelum data dibagi menjadi data latih dan uji, data akan melalui tahap preprocessing data yang bertujuan untuk memudahkan dalam pelatihan model. Teknik yang digunakan dalam penelitian ini, yaitu:

1. *Case Folding*
   Tahap merubah semua huruf pada kalimat menjadi huruf kecil atau *lowercase*.
2. *Cleaning*
   Tahap membersihkan data dari simbol, tanda baca dan angka yang tidak diperlukan dalam klasifikasi.
3. *Remove Stopwords*
   Tahap menghapus stopwords atau kata yang kurang memiliki makna saat data digunakan sebagai data latih pada machine learning.
4. *Tokenization*
   Tahap memisahkan teks kalimat menjadi token perkata sebagai langkah mempermudah saat menggunakan teknik stemming.
5. *Stemming*
   Tahap mengembalikan kata menjadi kata dasar, manfaat lain dari teknik stemming adalah mengurangi dimensi cospus atau fitur data.

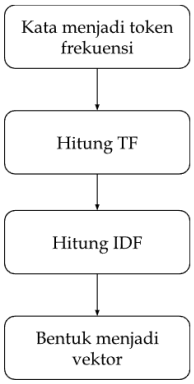
C. Ektraksi Fitur

Ada dua teknik ekstraksi fitur yang akan digunakan pada penelitian ini, teknik pertama adalah n-gram. Teknik n-gram menampilkan kata sejumlah n, pada penelitian ini akan digunakan unigram dan bigram. Berikut adalah bagaimana n-gram direpresentasikan pada gambar dibawah.



Gbr. 2 Proses n-gram

Langkah pertama adalah membuat vocabulary atau kamus berisi kata unik dari data yang digunakan, langkah kedua adalah membuat matriks fitur dengan memasukkan tiap kata pada kolom yang berbeda, langkah terakhir adalah mengisi nilai 1 pada kata yang muncul dan 0 pada kata yang tidak muncul. Teknik ekstraksi fitur kedua adalah tf-idf, berikut adalah bagaimana tf-idf direpresentasikan pada gambar dibawah.

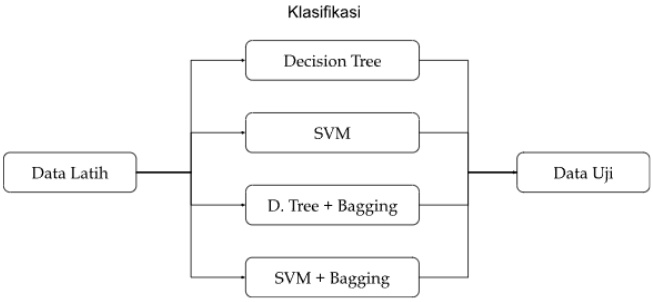


Gbr. 3 Proses tf-idf

Langkah pertama dalam implementasi tf-idf adalah membuat kata menjadi token frekuensi, artinya tiap kata unik akan dihitung berapa kali kata tersebut muncul pada dokumen, langkah kedua menghitung nilai tf dengan membag jumlah kata pada dokumen, langkah ketiga adalah menghitung idf. Terakhir data akan direpresentasikan pada bentuk table dan nilai tf-idf didapat dari perkalian tf dan idf.

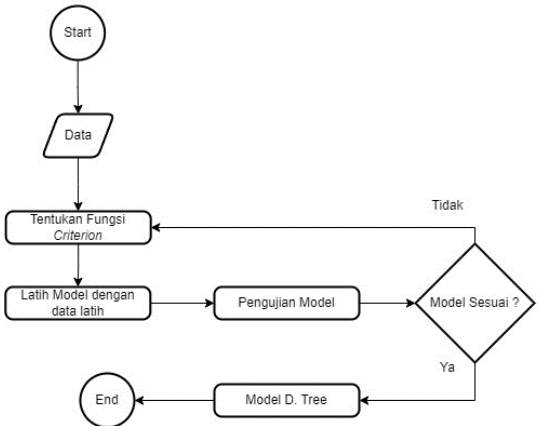
D. Pelatihan Model

Setelah melalui tahap pra pemrosesan data dan ekstraksi fitur, data siap digunakan sebagai data latih dan uji dari model machine learning. Penelitian ini akan menggunakan teknik bagging dengan dua jenis base model algoritma decision tree dan SVM.



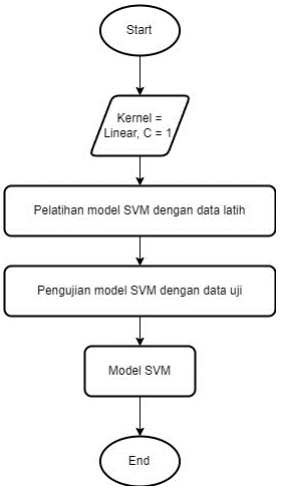
Gbr. 4 Proses Pelatihan Model

Metode pertama yang akan diuji adalah algoritma decision tree, gambar dibawah adalah diagram bagaimana melakukan metode decision tree pada penelitian ini.



Gbr. 5 *Decision tree*

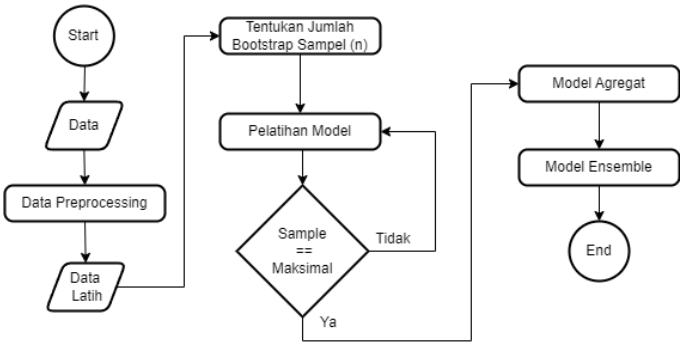
Decision tree perlu untuk membagi data menjadi data latih dan uji. Setelah dibagi, tentukan fungsi *criterion* dan jika model sudah sesuai maka proses selesai. Metode kedua adalah SVM, bagaimana implementasi dari algoritma SVM pada penelitian ini akan direpresentasikan pada diagram dibawah.



Gbr. 6 SVM

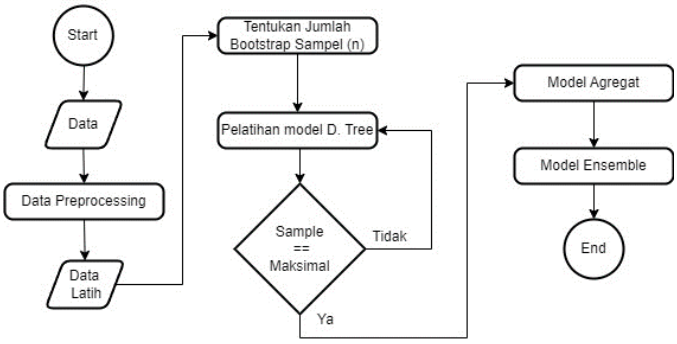
Parameter algoritma SVM yang digunakan dalam penelitian ini adalah kernel linear dan nilai C = 1, setelah

model melalui proses pelatihan dengan data latih, model akan digunakan untuk pengujian. Metode ketiga atau terakhir yang digunakan pada penelitian ini adalah teknik bagging, teknik ensemble yang sering digunakan untuk meningkatkan kinerja dari algoritma *machine learning*. Bagaimana bagging diimplementasikan pada penelitian ini akan direpresentasikan pada diagram dibawah.



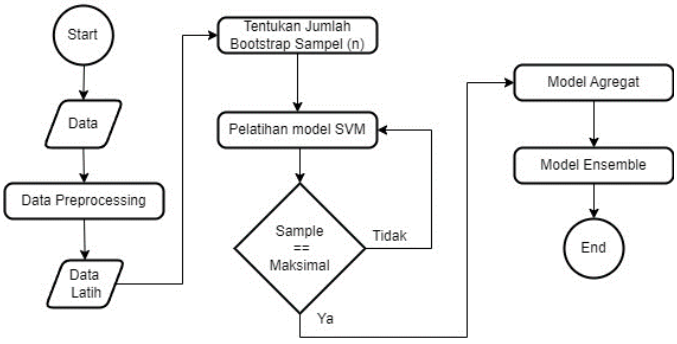
Gbr. 7 Bagging

Gambar 7 diatas adalah bagaimana metode bagging dilakukan, langkah pertama adalah menentukan jumlah bootstrap sampel, diteruskan dengan pelatihan model hingga boostrap sampel maksimal.



Gbr. 8 Bagging + D. Tree

Pada gambar 8, adalah posisi dari algoritma decision tree saat diterapkan teknik bagging. Tidak jauh berbeda dengan SVM pada gambar 9, posisi algoritma saat diterapkan teknik bagging juga sama, hanya saja base model atau bergantung pada basis dari teknik bagging yang digunakan.



Gbr. 9 Bagging + SVM

E. *Evaluasi*

Proses evaluasi dilakukan penghitungan nilai akurasi dari skenario pembagian jumlah dataset latih dan uji 70:30, 80:20 dan 90:10. Dalam penelitian ini juga akan menggunakan k-fold cross validation dalam evaluasi dan validasi model. Nilai K yang digunakan adalah 5 dan 10.

III. *HASIL DAN PEMBAHASAN*

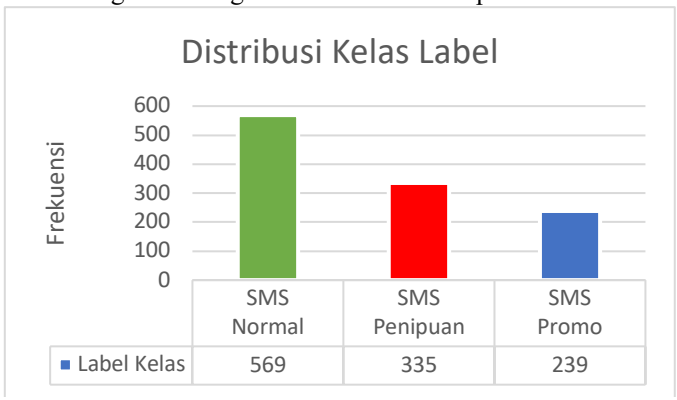
A. *Analisis Eksploratif Data*

Dataset yang telah didapat dari penelitian Rahmi dan Wibisono[1] ditampilkan pada panel data pandas untuk mengetahui bagaimana karakteristik dari data.

TABEL II
PANEL DATA DATASET

Label	Teks
0	Ada semua di panduan pla..dishare di grup ini dan fb
1	De inih bapa pake no temen, tolong kirimin pulsa ke no inih soalnya penring banget, nanti di rumah digantiin
2	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat anda yang terpilih. Aktifkan sekarang juga di *550*907# Buruan..! SKB

Untuk lebih mengenal tentang karakteristik dataset, juga dilakukan analisa distribusi kelas label pada dataset. Berikut adalah diagram batang distribusi kelas label pada dataset.



Gbr. 10 Distribusi Kelas Label

Pada gambar 10 adalah bagaimana distribusi kelas label dataset, jumlah kelas label normal sebanyak 569, kelas penipuan 335 dan kelas promo 239.

B. *Pra Pemrosesan Data*

Pada analisis eksploratif data, terlihat data masih belum dapat digunakan sebagai data latih, jika dipaksakan menjadi data latih maka akan mempengaruhi hasil dari model pembelajaran mesin. Dalam dunia *machine learning* ada istilah "*garbage in garbage out"* yaitu jika sesuatu yang dimasukkan kurang baik maka hasil keluaran juga akan mengikuti. Teknik prapemrosesan data pertama adalah *case folding*, pada tabel III dibawah adalah hasil data SMS setelah melalui pra pemrosesan data.

TABEL III
HASIL PRA PEMROSESAN DATA

Sebelum	Sesudah
Case Folding	
4.5GB/30 hari hanya Rp 55 Ribu Spesial buat anda yang terpilih. Aktifkan sekarang juga di *550*907# Buruan..! SKB	4.5gb/30 hari hanya rp 55 ribu spesial buat anda yang terpilih. aktifkan sekarang juga di *550*907# buruan..! skb
Cleaning	
4.5GB/30 hari hanya Rp 55 Ribu Spesial buat anda yang terpilih. Aktifkan sekarang juga di *550*907# Buruan..! SKB	hari hanya rp ribu spesial buat anda yang terpilih aktifkan sekarang juga di buruan skb
Remove Stopwords	
hari hanya rp ribu spesial buat anda yang terpilih aktifkan sekarang juga di buruan skb	hari hanya rp ribu special buat terpilih aktifkan sekarang juga buruan skb
Tokenization	
hari hanya rp ribu special buat terpilih aktifkan sekarang juga buruan skb	{"hari", "hanya", "rp", "ribu", "special", "buat", "terpilih", "aktifkan", "sekarang", "juga", "buruan", "skb"}
Stemming	
{"hari", "hanya", "rp", "ribu", "special", "buat", "terpilih", "aktifkan", "sekarang", "juga", "buruan", "skb"}	hari hanya rp ribu special buat pilih aktif sekarang juga buru skb

C. *Ektraksi Fitur*

Setelah melakukan pelatihan model data yang sudah melalui pra pemrosesan, perlu dilakukan ekstraksi fitur untuk mengenali pola dari tiap kelas SMS. Ekstraksi fitur pertama yang digunakan adalah N-Gram.

TABEL IV
TABEL FITUR UNIGRAM DAN BIGRAM

Fitur Unigram						
advice	aja	ayo	…	ball	trust	yth
Fitur Bigram						
advice text	aja data	aktif laku	…	trust in	wow aktif	yth rekening

Pada tabel IV diatas adalah fitur unigram dan bigram yang didapat dari dataset dengan menyeleksi kata unik menjadi pasangan kata jika ingin menggunakan lebih dari unigram atau 2 pasangan kata bahkan lebih.

TABEL V
PENGISIAN TABEL FITUR UNIGRAM DAN BIGRAM

Fitur Unigram						
advice	aja	ayo	…	ball	trust	yth
0	0	0	…	1	0	0
0	1	0	…	0	0	0
1	0	0	…	0	1	0
…	…	…	…	…	…	…
0	0	1	…	0	0	1
0	0	0	…	0	0	0
0	0	0	…	0	0	0
Fitur Bigram						
advice text	aja data	aktif laku	…	trust in	wow aktif	yth rekening
0	0	0	…	0	0	0
0	1	0	…	0	0	0
1	0	0	…	1	0	0
…	…	…	…	…	…	…
0	0	1	…	0	0	1
0	0	0	…	0	1	0
0	0	0	…	0	0	0

Pada table V diatas adalah hasil dari pengisian baris fitur dari unigram dan bigram, jika kalimat mengandung kata dari salah satu fitur maka akan diberi nilai 1 dan jika tidak maka diberi nilai 0. Proses ekstraksi fitur kedua yang akan digunakan pada penelitian ini adalah tf-idf, setelah diterapkan teknik n-gram data melalui pembobotan tf-idf.

diskon	kacamata	maaf	promo	bayar	gratis
0.217670	0.556839	0.000000	0.157912	0.00000	0.000000
0.000000	0.000000	0.271116	0.000000	0.31051	0.000000
0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
0.000000	0.000000	0.295538	0.000000	0.00000	0.000000
0.228875	0.000000	0.000000	0.166041	0.00000	0.000000
0.000000	0.000000	0.000000	0.000000	0.00000	0.252293
0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
0.000000	0.000000	0.000000	0.000000	0.41398	0.000000
0.000000	0.000000	0.000000	0.000000	0.00000	0.188515
0.000000	0.000000	0.000000	0.000000	0.00000	0.000000

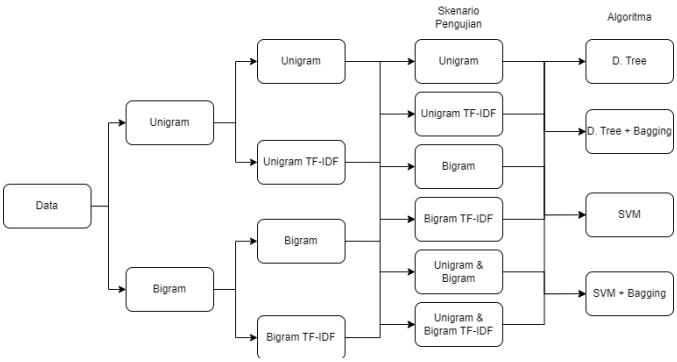
Gbr. 11 Hasil Implementasi Unigram tf-idf

abadi big	advice text	aja data	aktif laku	aman klik	appreciation trust	ayo coba	ball pool	banget loh	banget main
0.00000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.27735	0.27735	0.27735
0.00000	0.000000	0.57735	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000
0.00000	0.218218	0.00000	0.000000	0.000000	0.218218	0.000000	0.00000	0.00000	0.00000
0.00000	0.000000	0.00000	0.305386	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000
0.00000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000
0.00000	0.000000	0.00000	0.000000	0.353553	0.000000	0.353553	0.00000	0.00000	0.00000
0.26062	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000

Gbr. 12 Hasil Implementasi Bigram tf-idf

Pada gambar 11 dan 12 adalah hasil dari teknik n-gram setelah dilakukan pembobotan tf-idf. Berbeda dengan n-gram saja yang menggunakan notasi 0 dan 1, pada tf-idf menggunakan persamaan matematika sehingga didapat bobot yang berbeda antar kalimatnya.

D. *Pelatihan Model*

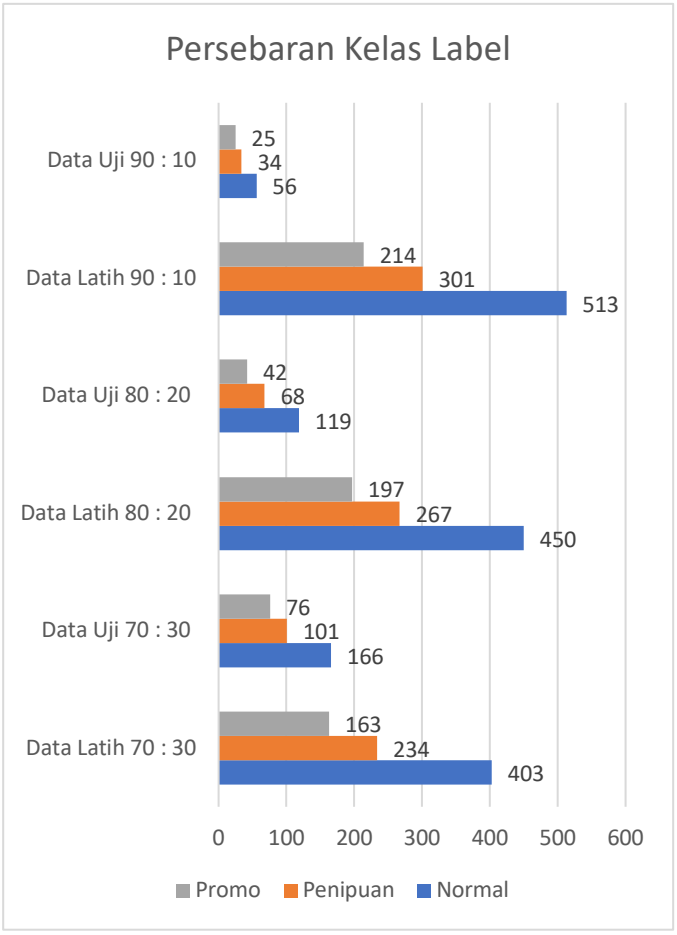


Gbr. 13 Skenario Pelatihan

Pada gambar 13 menunjukkan skenario pelatihan dimana data yang utuh akan dibagi menjadi unigram dan bigram. Setelah itu dipecah lagi menjadi 2 jenis lagi yaitu menggunakan pembobotan tf-idf dan tidak sehingga data siap digunakan pada model pembelajaran mesin.

E. *Evaluasi*

Tahap terakhir dalam pembelajaran mesin adalah melakukan evaluasi pada model yang telah di latih dan di uji, pada gambar 14 dibawah adalah grafik persebaran kelas label dataset.



Gbr. 14 Persebaran Kelas Label

Hasil nilai akurasi pengujian pada tiap skenario ditunjukkan pada tabel VI dibawah, pada skenario pertama adalah perbandingan data latih dan uji sebanyak 70:30, diikuti 80:20 dan terakhir 90:10.

TABEL VI
HASIL NILAI AKURASI BERDASARKAN SKENARIONYA

Algoritma	Perbandingan Data Latih dan Uji 70:30					
	Unigram	Unigram tf-idf	Bigram	Bigram tf-idf	Unigram & Bigram	Unigram & Bigram tf-idf
D. Tree	0.85	0.85	0.8	0.77	0.86	0.84
D. Tree + Bagging	0.91	0.9	0.68	0.66	0.9	0.88
SVM	0.92	0.93	0.75	0.88	0.91	0.94
SVM + Bagging	0.92	0.93	0.73	0.86	0.9	0.94
Algoritma	Perbandingan Data Latih dan Uji 80:20					
	Unigram	Unigram tf-idf	Bigram	Bigram tf-idf	Unigram & Bigram	Unigram & Bigram tf-idf
D. Tree	0.9	0.89	0.79	0.78	0.9	0.86
D. Tree + Bagging	0.95	0.95	0.72	0.71	0.93	0.93
SVM	0.93	0.94	0.76	0.88	0.92	0.96
SVM + Bagging	0.92	0.96	0.74	0.85	0.93	0.96
Algoritma	Perbandingan Data Latih dan Uji 90:10					
	Unigram	Unigram tf-idf	Bigram	Bigram tf-idf	Unigram & Bigram	Unigram & Bigram tf-idf
D. Tree	0.89	0.82	0.81	0.8	0.89	0.85
D. Tree + Bagging	0.9	0.92	0.68	0.7	0.92	0.89
SVM	0.91	0.91	0.76	0.84	0.9	0.91
SVM + Bagging	0.9	0.91	0.72	0.85	0.89	0.91

Pada tabel VI menunjukkan nilai akurasi berdasarkan skenario pengujian pada perbandingan data latih dan uji, pada tiap skenario nilai tertinggi ada pada SVM dan SVM + Bagging yang memiliki nilai 96% saat menggunakan data unigram & bigram tf-idf.

TABEL VII
HASIL NILAI AKURASI TERTINGGI TIAP ALGORITMA

Perbandingan Data Latih dan Uji 70:30		
Algoritma	Jenis Data Latih	Nilai Akurasi
D. Tree	Unigram dan Bigram	0.86
D, Tree + Bagging	Unigram	0.91
SVM	Unigram & Bigram tf-idf	0.94
SVM + Bagging	Unigram & Bigram tf-idf	0.94
Perbandingan Data Latih dan Uji 80:20		
Algoritma	Jenis Data Latih	Nilai Akurasi
D. Tree	Unigram dan Unigram & Bigram	0.9
D, Tree + Bagging	Unigram dan Unigram tf-idf	0.95
SVM	Unigram & Bigram tf-idf	0.96
SVM + Bagging	Unigram tf-idf dan Unigram & Bigram tf-idf	0.96
Perbandingan Data Latih dan Uji 90:10		
Algoritma	Jenis Data Latih	Nilai Akurasi
D. Tree	Unigram dan Unigram & Bigram	0.89
D, Tree + Bagging	Unigram tf-idf dan Unigram & Bigram	0.92
SVM	Unigram, Unigram tf-idf dan unigram & bigram tf-idf	0.91
SVM + Bagging	Unigram tf-idf dan Unigram & Bigram tf-idf	0.91

Pada tabel VII adalah hasil nilai akurasi tertinggi tiap algoritma, pattern atau pola yang dapat diketahui dari data tersebut adalah jenis data unigram selalu menjadi penyebab dari nilai akurasi tertinggi.

TABEL VIII
HASIL NILAI K-FOLD CROSS VALIDATION

Nilai K = 5						
Algoritma	Data					
	Unigram	Unigram tf-idf	Bigram	Bigram tf-idf	Unigram & Bigram	Unigram & Bigram tf-idf
D. Tree	0.86	0.87	0.79	0.79	0.87	0.85
D. Tree + Bagging	0.91	0.91	0.68	0.68	0.89	0.89
SVM	0.93	0.93	0.76	0.76	0.91	0.92
SVM + Bagging	0.92	0.93	0.72	0.73	0.9	0.9
Nilai K = 10						
Algoritma	Data					
	Unigram	Unigram tf-idf	Bigram	Bigram tf-idf	Unigram & Bigram	Unigram & Bigram tf-idf
D. Tree	0.88	0.88	0.78	0.79	0.88	0.87
D. Tree + Bagging	0.91	0.91	0.7	0.7	0.89	0.91
SVM	0.93	0.93	0.77	0.76	0.93	0.92
SVM + Bagging	0.92	0.92	0.74	0.73	0.91	0.91

Pengujian k-fold menggunakan k = 5, Peningkatan signifikan pada algoritma D. Tree setelah diterapkan teknik bagging terjadi saat menggunakan data unigram tanpa tf-idf dengan kenaikan akurasi dari 86% menjadi 91%.

IV. *KESIMPULAN*

Penerapan teknik bagging berpengaruh signifikan pada peningkatan akurasi algoritma decision tree. Dalam uji k-fold algoritma SVM memiliki nilai akurasi paling tinggi mencapai 93%. Peningkatan signifikan pada algoritma D. Tree setelah diterapkan teknik bagging terjadi saat menggunakan data unigram tanpa TF-IDF dengan kenaikan akurasi dari 86% menjadi 91% pada uji 5-fold. Algoritma SVM tidak ada penigkatan nilai akurasi secara signifikan atau berdampak saat diterapkan teknik bagging pada uji k-fold. Data bigram tidak memberikan hasil yang maksimal pada hasil akurasi model. Model terbaik adalah algoritma SVM menggunakan data unigram & bigram tf-idf dan SVM + Bagging menggunakan unigram tf-idf yang menunjukkan performa baik dengan memiliki nilai akurasi 96%.

UCAPAN TERIMA KASIH

Puji syukur Alhamdulillah senantiasa penulis ucapkan kepada Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan proses pembuatan jurnal ini hingga selesai. Tak lupa juga ucapan terimakasih kepada pihak-pihak yang telah membantu proses penyusunan jurnal ini, sehingga penulis dapat menyelesaikan jurnal tepat pada waktunya.

REFERENSI

[1] Rahmi F. & Wibisono Y., "Aplikasi SMS Spam Filtering pada Android menggunakan Naïve Bayes", 2021.
[2] Sunardi, H. Murti dan H. Listiyono, "Aplikasi SMS Gateway", vol. XIV, no. 1, 2009.
[3] Novanema Duha, "Short Message Services (SMS) Fraud Against Mobile Telephone Provider Consumer Review From Law Number 8 Of 1999 Concerning Consumer Protection", vol. III, no. 1, 2021.
[4] Nabila Affah Mumtazahh, "Perbandingan Hasil Metode Support Vector Machine (SVM) dengan Ensemble SMOTE Bagging dan SMOTE Boosting pada Data Kelulusan Mahasiswa UNIMUS", 2021.
[5] Devi Irawan, Eza Budi Perkasa dkk, "Perbandingan Klasifikasi SMS Berbasis Support Vector Machine, Naïve Bayes Classifier, Random Forest dan Bagging Classifier", vol. X, 2021.
[6] Edi Zuviyanto, Teguh Bharata Adji, dkk, "Perbandingan Algoritme-Algoritme Pembelajaran Mesin Pada Klasifikasi SMS Spam", 2018.
[7] Giorgio Valentini, Francesco Masulli, "Ensemble Of Learning Machines", 2002.
[8] E. Bauer dan R. Kohavi, "An Empirical Comparison of voting classification algorithms: Bagging, boosting and variants Machine Learning", 525-536, 1999.
[9] T. G. Dietterich, "An Experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization Machine Learning", 139-158, 2000.
[10] Y. Freund dan R. Schapire, "Experiments with a new boosting algorithm", 148-156, 1996.
[11] Robert Duin, "Limited Bagging, Boosting and the Random Subspace Method for Linear Classifiers, 2002.
[12] Rizku Tri Prasetio dan Pratiwi, "Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan kelas Dataset Medis", vol. II, 2015.
[13] Azmiardhy Zulkifli F., "Deteksi Surel SPAM dan Non-Spam Bahasa Indonesia Menggunakan Metode Naïve Bayes", 2021.
[14] Panji Bimo Nugroho, Dkk., "Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5", 64-71, 2020.
[15] Yogo Aryo Jatmiko, Dkk., " Analisis Perbandingan Kinerja CART Konvensional, Bagging dan Random Forest Pada Klasifikasi Objek: Hasil dari Dua Simulasi", 2019.
[16] Agung Nugroho, Yoga Religia, "Analisis Optimasi Algoritma Klasifikasi Naïve Bayes menggunakan Genetic Algorithm dan Bagging", 2021.
[17] Nur Diana Putri Saputri, "Analisis Perbandingan Performa Algortima C4.5 dengan Optimasi Menggunakan Metode Bagging dan Adaboost Pada Dataset Penyakit Stroke, 2021.
[18] Eko Puji, dkk., "Optimasi Nilai K pada Algoritma KNN untuk Klasifikasi Spam dan Ham Email", 2020.
[19] Muhammad Khafidhun Alim Muslim, "Klasifikasi Opini Pengguna Twitter Terhadap Sekolah Daring dengan Metode Naïve Bayes dan SVM", 2021.