TESIS

KOMPARASI ALGORITMA KLASIFIKASI TEKS DENGAN METODE EKSTRAKSI N-GRAM PADA HASIL MEDIASI PERKARA PERDATA DI PENGADILAN NEGERI

Disusun oleh:

Nama : **Retzi Yosia Lewu**
NIM : **22.55.2300**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

TESIS

KOMPARASI ALGORITMA KLASIFIKASI TEKS DENGAN METODE EKSTRAKSI N-GRAM PADA HASIL MEDIASI PERKARA PERDATA DI PENGADILAN NEGERI

COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION ALGORITHMS WITH N-GRAM EXTRACTION METHOD ON THE RESULT OF CIVIL CASE MEDIATION IN DISTRICT COURTS

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Retzi Yosia Lewu**
NIM : **22.55.2300**
Konsentrasi : **Digital Transformation Intelligence**

PROGRAM STUDI S2 INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

HALAMAN PENGESAHAN

KOMPARASI ALGORITMA KLASIFIKASI TEKS DENGAN METODE EKSTRAKSI N-GRAM PADA HASIL MEDIASI PERKARA PERDATA DI PENGADILAN NEGERI

COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION ALGORITHMS WITH N-GRAM EXTRACTION METHOD ON THE RESULT OF CIVIL CASE MEDIATION IN DISTRICT COURTS

Dipersiapkan dan Disusun oleh

Retzi Yosia Lewu

22.55.2300

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 1 Agustus 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 1 Agustus 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

KOMPARASI ALGORITMA KLASIFIKASI TEKS DENGAN METODE EKSTRAKSI N-GRAM PADA HASIL MEDIASI PERKARA PERDATA DI PENGADILAN NEGERI

COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION ALGORITHMS WITH N-GRAM EXTRACTION METHOD ON THE RESULT OF CIVIL CASE MEDIATION RESULT IN DISTRICT COURTS

Dipersiapkan dan Disusun oleh

Retzi Yosia Lewu

22.55.2300

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 1 Agustus 2024

Pembimbing Utama

Prof. Dr. Kusrini, M.Kom
NIK. 190302106

Anggota Tim Penguji

Dr. Ferry Wahyu Wibowo, S.Si., M.Cs
NIK. 190302235

Pembimbing Pendamping

Dhani Ariatmanto, S.Kom., M.Kom., Ph.D.
NIK. 190302197

Ainul Yaqin, M.Kom
NIK. 190302255

Prof. Dr. Kusrini, M.Kom
NIK. 190302106

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 1 Agustus 2024
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Retzi Yosia Lewu
NIM : 22.55.2300
Konsentrasi : Digital Transformation Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Komparasi Algoritma Klasifikasi Teks dengan Metode Ekstraksi N-gram pada hasil mediasi perkara perdata di Pengadilan Negeri

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom
Dosen Pembimbing Pendamping : Ainul Yaqin, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 1 Agustus 2024
Yang Menyatakan,

Retzi Yosia Lewu

HALAMAN PERSEMBAHAN

Teruntuk mereka yang menjadi sumber inspirasi penulisan Tesis ini, terima kasih yang tak terhingga.

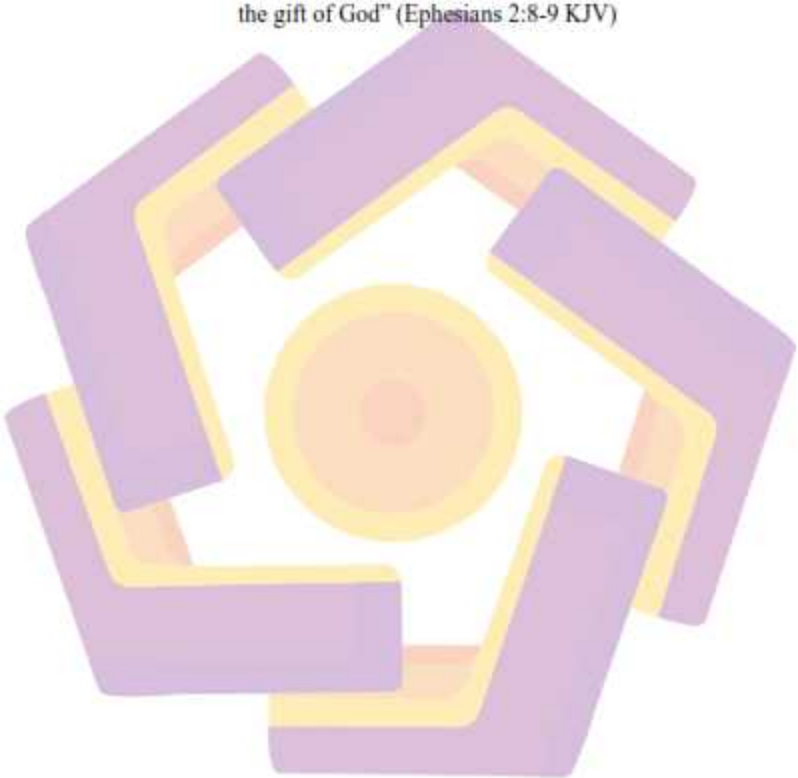
Untuk diri sendiri yang telah berani mengambil keputusan untuk mewujudkan mimpi dan cita – cita yang terucap 20 tahun yang lalu, we did it and we will make many more come true!

Untuk siapa saja yang membaca tulisan ini, semoga dapat menginspirasi dan memotivasi untuk tidak pernah berhenti belajar, tidak cepat menyerah dan tetap berusaha.

HALAMAN MOTTO

Selalu ada faktor TUHAN dalam hal apapun yang kita usahakan

"For by grace are you saved through faith; and that not of yourselves: it is the gift of God" (Ephesians 2:8-9 KJV)



KATA PENGANTAR

Sembah syukur Penulis haturkan kepada Tuhan Yang Maha Kuasa yang telah menganugerahkan kesehatan dan kemampuan untuk merampungkan Tesis ini, yang walaupun masih banyak kekurangan dan keterbatasannya telah mengantar penulis dalam pemahaman bahwasanya kesempurnaan hanyalah milik Tuhan Yang Maha Kuasa. Dalam perjalanan penyusunan Tesis ini, Penulis memperoleh insight bahwa dengan perkembangan teknologi yang sangat cepat, diperlukan kemampuan literasi digital yang mumpuni untuk mencari penelitian – penelitian terbaru sehingga penelitian yang dilakukan Penulis dapat memberi kontribusi. Kesadaran ini yang memotivasi penulis untuk terus mengupgrade pengetahuan dan kemampuan dengan banyak membaca dan berdiskusi dengan banyak pihak. Untuk itu, terima kasih yang sedalam – dalamnya Penulis haturkan kepada segenap pihak yang telah menginspirasi, memberi dukungan moril maupun materiil dalam penyusunan tesis ini. Semoga semuanya kembali menjadi pahala dan berkat dari Tuhan Yang Maha Esa.

Selesainya tesis ini tidak lepas dari bimbingan dan arahan para pembimbing yang selalu memberikan dorongan dan bimbingan tanpa mengenal lelah, meskipun hasil akhir penulisan ini sepenuhnya menjadi tanggung jawab penulis. Penulis mengucapkan terima kasih yang tak terhingga kepada Prof. Dr. Kusrini, M.Kom., sebagai Dosen Pembimbing I juga selaku Penguji 3, yang adalah salah satu inspirasi penulis dalam penyusunan tesis ini. Terima kasih untuk setiap cerita dan pengalaman yang dibagikan yang telah menginspirasi dan memotivasi untuk terus

bergerak maju, beliau yang dalam kesibukannya selalu konsisten mengingatkan tenggat waktu dalam setiap tahapan, memberikan bantuan saat publikasi, memberikan masukan – masukan yang sangat memperkaya dan koreksi yang membangun untuk seluruh tahap penulisan tesis ini, serta memberikan penulis kesempatan untuk menggunakan fasilitas server kampus untuk mengerjakan uji coba kode program. Juga untuk Pembimbing II Bapak Ainul Yaqin, S.Kom., M.Kom yang selalu memberikan masukan dalam penulisan tesis ini, terima kasih yang tak terhingga untuk setiap koreksi dan saran yang diberikan dalam setiap tahapan.

Terima kasih pula penulis haturkan kepada para Dosen Penguji dalam setiap tahapan, kepada Bapak Dr. Ferry Wahyu Wibowo S.Si., M.Cs sebagai Penguji I pada tahapan SHPT dan Ujian Tesis, seluruh masukan yang telah diterima adalah pengetahuan yang sangat bernilai bagi penulis. Juga kepada Bapak Dhani Ariatmanto, S. Kom., M. Kom., Ph.D sebagai Penguji II pada Ujian Tesis yang juga telah memberikan banyak sekali koreksi dan saran yang membangun sejak Seminar Proposal dan SHPT. Penulis ucapkan terima kasih pula kepada Bapak Alva Hendi Muhammad, M.Eng., Ph.D sebagai Dosen Penguji dalam Seminar Proposal, atas masukan dan sarannya untuk penyempurnaan tesis ini, terima kasih pula atas ilmu yang dibagikan dalam mata kuliah Cyber Scurity.

Terima kasih pula kepada Rektor Universitas AMIKOM Yogyakarta, Prof. Dr., M. Suyanto, M.M, yang juga menjadi pengajar dalam mata kuliah Technopreneurship, atas ilmu dan pengalaman yang telah diberikan, yang sangat menginspirasi dan memotivasi penulis. Juga kepada seluruh dosen yang tak bisa

disebutkan satu persatu, yang telah membagikan ilmu dan pengetahuan kepada seluruh mahasiswa PJJ Magister Informatika Universitas AMIKOM Yogyakarta Program Studi Digital Transformation Intelligence sehingga kami dapat menempuh dan menyelesaikan studi Strata dua (S2) tepat waktu dengan hasil yang sangat baik.

Terima kasih pula kepada instansi dimana Penulis mengabdikan diri, Mahkamah Agung RI dan secara khusus untuk Pengadilan Tinggi Manado, kepada para Pimpinan Pengadilan Tinggi Manado yang telah memberikan ijin untuk melanjutkan pendidikan formal ke jenjang S2, semoga tulisan ini dapat membawa manfaat untuk institusi.

Kepada teman – teman Angkatan 8 PJJ MTI Amikom yang telah banyak berbagi pengalaman, pengetahuan, suka duka selama perkuliahan, terima kasih untuk tetap solid dalam perkuliahan dan pembuatan tugas – tugas.

Akhirnya, terima kasih pula kepada keluarga, suami Juan Demmis Abram dan anak tercinta, Darrel Harry Semuel Abram, atas doa, dukungan dan pengertiannya selama penyusunan tesis ini. Semoga Tuhan yang Maha Kuasa selalu menyertai kita.

Akhir kata, penulis berharap kiranya tesis ini dapat memenuhi persyaratan yang telah ditentukan sebagai persyaratan kelulusan di Universitas AMIKOM Yogyakarta dan semoga membawa manfaat untuk ilmu pengetahuan khususnya untuk pemanfaatan Teknologi Informasi di Pengadilan.

Yogyakarta, 1 Agustus 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

Data Collection: Proses mengumpulkan data dari berbagai sumber

Data Selection: Proses memilah data yang telah dikumpulkan

Data preparation: Proses persiapan data agar dapat dipreprocess

Data Preprocessing: Proses mengubah data mentah dengan metode – metode text mining

Casefolding: Tahapan preprocessing berupa pengubahan huruf dalam dataset menjadi huruf kecil

Cleaning: Tahapan preprocessing berupa pembersihan dataset dari tag HTML, tanda baca maupun karakter dan simbol

Tokenization: Tahapan preprocessing beruba pemecahan kalimat menjadi kata – kata tunggal

Filtering: Tahapan preprocessing berupa pemilihan, penyariangan atau modifikasi dataset dengan teknik tertentu.

Stop word removal: Tahapan preprocessing berupa penghapusan kata – kata umum didasarkan pada daftar kata umum / kamus kata

Library: kumpulan kode yang dapat digunakan berulang – ulang, merupakan kumpulan fungsi atau kelas atau modul

NLTK: kependekan dari Natural Language Tool Kit yaitu salah satu library dalam Bahasa pemrograman python yang digunakan dalam tahapan preprocessing

Sastrawi: Library python khusus Bahasa Indonesia

Pandas: Library python untuk manipulasi data, terutama data tabular atau spreadsheet. Struktur utamanya adalah series dan data frame

Scikit learn: Modul Bahasa python untuk melakukan processing atau training data untuk machine learning atau data science. Kebanyakan digunakan untuk perhitungan matematis.

N-gram: potongan n-karakter dari suatu string atau kalimat. N-gram ditentukan dari nilai n. sering digunakan umtuk pemodelan prediksi kata yang diperoleh dari kata N-1

TF-IDF: Faktor pembobot yang menunjukkan seberapa penting sebuah kata atau term dalam sebuah korpus dokumen atau kalimat.

Data latih: data yang digunakan untuk melatih model dalam text mining, biasanya diambil dalam persentasi yang paling besar dari keseluruhan dataset.

Data uji: data yang digunaka untuk menguji model yang dibangun dari data latih. Biasanya jumlah lebih kecil dari data latih

Naïve bayes: Algoritma klasifikasi teks berdasarkan kelas dan probabilitas keanggotaan dalam kelas

Logistic regression: Algoritma klasifikasi teks dengan model statistik yang menggunakan fungsi logistic atau logit, dalam matematika sebagai persamaan x dan y.

Decision tree: Algoritma klasifikasi teks yang dalam implementasinya digambarkan dalam bentuk pohon untuk pemodelan hubungan antara variabel – variabel dalam pengambilan keputusan.

Support Vector Machine: Algoritma klasifikasi teks untuk supervised learning yang mebagi dataset menjadi dua kelas positiif dan negative menggunakan garis vector.

Komparasi: Perbandingan berdasarkan metriks tertentu

Algoritma: urutan operasi yang disusun secara logis dan sistematis untuk menghasilkan output tertentu

Performa Algoritma: kinerja algoritma dalam melakukan pemodelan terhadap kondisi tertentu

Akurasi: seberapa tepat dan akurat sebuah model machine learning memprediksi nilai target sesuai sasaran

Presisi: proporsi prediksi berlabel positif yang benar terhadap keseluruhan prediksi positif

Recall: sejauh mana model machine learning dapat mengidentifikasi dengan benar semua nilai positif yang ada

F1-score: gabungan keseimbangan antara presisi dan recall

Entropy: distribusi probabilitas dalam teori informasi yang digunakan untuk mengukur tingkat homogenitas distribusi kelas.

Gain: nilai keuntungan dari sebuah node dalam decision tree. Node dengan nilai gain besar akan menjadi root node

Probabilitas prior: probabilitas masing – masing kelas tanpa mempertimbangkan fitur – fitur yang ada

Probabilitas likelihood: probabilitas bahwa bukti X (yaitu, n-gram yang diamati dalam dokumen) akan muncul jika hipotesis H (yaitu, kelas dokumen) benar.

Threshold: ambang batas

Mediasi: cara penyelesaian sengketa sebelum persidangan antara pihak berperkara dengan bantuan mediator yang netral

Petitum: bagian dari surat gugatan yang berisi rincian yang dikehendaki penggugat kepada tergugat untuk diputuskan oleh pengadilan

Variabel target: variabel yang menjadi tujuan

Variabel sumber: variabel yang menjadi basis/dasar untuk melatih model machine learning

INTISARI

Penelitian ini bertujuan melakukan komparasi penggunaan algoritma klasifikasi teks dalam pemodelan untuk memprediksi status keberhasilan mediasi berdasarkan dokumen – dokumen hasil mediasi perkara perdata terdahulu dengan metode ekstraksi n-gram. Proses analisis dilakukan melalui Seleksi Data, Persiapan Data, Pra-pemrosesan, Implementasi N-gram, Pembobotan TF-IDF, Pemodelan, dan Evaluasi Model dilanjutkan dengan komparasi performa algoritma. Penelitian ini diharapkan memperoleh gambaran algoritma mana yang memiliki performa terbaik dalam mengoptimalkan prediksi hasil mediasi.

Seleksi data dilakukan terhadap 2591 data awal untuk memilah data yang dapat digunakan dan yang tidak, dilanjutkan dengan persiapan dengan hanya menggunakan 2 (dua) status hasil mediasi yaitu Y = mediasi berhasil dan T = mediasi gagal. Dalam pra-pemrosesan digunakan beberapa teknik yaitu perubahan huruf pada kolom petitum menjadi huruf kecil, pembersihan dengan menghilangkan tag HTML, tanda baca, simbol dan juga angka, dilanjutkan dengan tokenisasi untuk mendapatkan unit kata dalam kalimat. Proses selanjutnya adalah pencarian akar kata dengan memotong imbuhan menggunakan library Sastrawi yang dibuat khusus untuk Bahasa Indonesia kemudian disaring menggunakan kamus kata Sastrawi. Dataset kemudian dipisah dengan perbandingan 70:30 untuk data latih dan data uji. Terhadap data latih dilakukan implementasi n-gram yang kemudian dihitung nilai TF-IDFnya untuk melihat sejauh mana pemilihan nilai n mempengaruhi TF-IDF. Hasilnya diproses lebih lanjut untuk membuat model menggunakan algoritma Naïve Bayes, Regresi Logistik, Pohon Keputusan dan Support Vector Machine (SVM). Dengan data uji model tersebut dievaluasi performanya menggunakan metrik akurasi, presisi, pemanggilan kembali dan skor F1.

Hasil penelitian menunjukkan bahwa hasil mediasi terdahulu dapat digunakan untuk membuat model prediksi hasil mediasi dengan kolom petitum sebagai target. Penggunaan fitur ekstraksi n-gram dengan nilai n antara 1 hingga 5 mempengaruhi performa model. Implementasi Algoritma menunjukkan bahwa untuk Kelas T, Regresi Logistik dan Pohon Keputusan menunjukkan performa terbaik dengan akurasi sangat tinggi dan metrik kinerja yang seimbang. Untuk Kelas Y, Regresi Logistik dan Pohon Keputusan juga menunjukkan performa yang sangat baik pada n-gram lebih dari 1, sementara Naïve Bayes dan SVM gagal dalam klasifikasi ini pada n-gram unigram maupun n > 1. Penggunaan n > 1 sedikit meningkatkan performa terutama pada Kelas T, tetapi tidak memberikan keuntungan signifikan untuk Kelas Y di Naïve Bayes dan SVM. Logistic Regression dan Decision Tree tetap menunjukkan performa terbaik secara konsisten pada semua konfigurasi n-gram.

Kata kunci: Algoritma Klasifikasi Teks, N-gram, status hasil mediasi

ABSTRACT

This research aimed to compare the use of classification algorithms to produce a model in predicting the mediation success rate using previous mediation documents by applying the n-gram extraction method. The data analysis processes are Data Selection, Data Preparation, Preprocessing, N-gram implementation, TF-IDF weighting, Modelling using Classification Algorithms, Model evaluation then continued to compare the performances. It is expected to know one among the algorithms has the best performance in optimizing mediation result prediction.

Data selection was held to the original 2591 dataset to properly choose among them a suitable and qualified one to be used in the next step of preparation in which only two mediation statuses, which are Y for succesfull mediation and T for the failed one, are being used. By preprocessing, some techniques are applied. They are Case-folding for the petitum field, Cleaning – to remove HTML tags, punctuations, symbols and numbers, then Tokenization – to tokenize the sentences into word units. Stemming is the next process – to stem a word into its base unit using specialized library for Indonesian language called Sastrawi, then filtered using Sastrawi reserved stopwords. The result is then divided into 70:30 training data and test data. Training data are used in the implementation of n-gram and TF-IDF weigthing – by analyzing how the value of n affect the weight of TF-IDF. The result is occupied to produce the model using Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine (SVM). Test data are used to evaluate the performance using Accuracy, Precision, recall and F1 Score metrices.

The study's findings imply that previous mediation resulix can be utilised to predict the outcome of a new mediation request when the petitum column is used as the targeted label. Using the n-gram extraction feature with values ranging from 1 to 5 affects the model performance. For Class T, LR and DT outperform with extremely high accuracy and balanced performance metrics. For Class Y, LR and DT also perform well on n-grams greater than one, however NB and SVM failed in this classification on both unigrams and $n > 1$. Using $n > 1$ improves performance slightly, particularly in Class T, but has no substantial advantage over Class Y in NB or SVM.

Keyword: Text Classification Algorithms, N-gram, Mediation result prediction

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Pengadilan negeri adalah pengadilan tingkat pertama yang memiliki kewenangan untuk memeriksa, menyidangkan dan memutus perkara, baik perkara perdata maupun pidana. Penyelesaian perkara perdata di pengadilan negeri dilakukan dalam beberapa tahap sebagaimana Gambar 1.1 berikut:

Gambar 1.1. Alur proses perkara perdata pada pengadilan negeri

. Salah satu tahapan dalam alur proses tersebut adalah mediasi, dimana pengadilan negeri, dengan perantaraan Ketua Pengadilan Negeri, memiliki kewajiban untuk mengusahakan perdamaian antara kedua pihak yang bersengketa (Pasal 130 ayat 1 HIR/Pasal 154 Rbg). Tujuan mediasi adalah untuk memperoleh kesepakatan damai di antara para pihak terkait sengketa yang dihadapi sehingga perkara tidak perlu dilanjutkan pada persidangan - persidangan yang memakan waktu lama.

Mediasi dinyatakan berhasil apabila kedua pihak berperkara menyatakan sepakat untuk berdamai di hadapan mediator yang dipilihnya atau ditunjuk pengadilan. Mediasi dinyatakan gagal apabila pihak tidak beritikad baik, dengan

kriteria menurut PERMA Nomor 1 tahun 2016 tentang mediasi yaitu: Tidak hadir setelah dipanggil secara patut 2 kali berturut-turut tanpa alasan sah dan tidak menandatangani konsep kesepakatan perdamaian. Di dalam mediasi, seorang mediator berkewajiban menyampaikan pertimbangan – pertimbangan hukum kepada kedua belah pihak berperkara berdasarkan fakta – fakta hukum dan hukum yang diterapkan di Indonesia yang sesuai dengan perkara yang disengketakan.

Tidak semua perkara perdata yang diajukan ke pengadilan dapat dimediasi. Menurut Pasal 4 Ayat Peraturan Mahkamah Agung RI Tahun 2016 Tentang Mediasi, sengketa yang dikecualikan dari kewajiban mediasi adalah:

a. sengketa yang pemeriksaannya di persidangan ditentukan tenggang waktu penyelesaiannya meliputi:
   i. sengketa yang diselesaikan melalui prosedur Pengadilan Niaga;
   ii. sengketa yang diselesaikan melalui prosedur Pengadilan Hubungan Industrial;
   iii. keberatan atas putusan Komisi Pengawas Persaingan Usaha;
   iv. keberatan atas putusan Badan Penyelesaian Sengketa Konsumen;
   v. permohonan pembatalan putusan arbitrase;
   vi. keberatan atas putusan Komisi Informasi;
   vii. penyelesaian perselisihan partai politik;
   viii. sengketa yang diselesaikan melalui tata cara gugatan sederhana;
   ix. sengketa lain yang pemeriksaannya di persidangan ditentukan tenggang waktu penyelesaiannya dalam ketentuan peraturan perundang-undangan;
b. sengketa yang pemeriksaannya dilakukan tanpa hadirnya penggugat atau tergugat yang telah dipanggil secara patut;
c. gugatan balik (rekonvensi) dan masuknya pihak ketiga dalam suatu perkara (intervensi);
d. sengketa mengenai pencegahan, penolakan, pembatalan dan pengesahan perkawinan;

e. sengketa yang diajukan ke Pengadilan setelah diupayakan penyelesaian di luar Pengadilan melalui Mediasi dengan bantuan Mediator bersertifikat yang terdaftar di Pengadilan setempat tetapi dinyatakan tidak berhasil berdasarkan pernyataan yang ditandatangani oleh Para Pihak dan Mediator bersertifikat

Berikut ini adalah data perkara perdata pada pengadilan negeri (pengadilan tingkat pertama) yang berada di wilayah Pengadilan Tinggi Manado yang dimediasi pada rentang waktu tahun 2020 – 2023 (Sumber: Observasi pada Pengadilan Negeri).

Tabel 1.1. Pelaksanan mediasi tahun 2020 - 2023

No.	Satuan Kerja	Jumlah perkara Perdata yang dimediasi	Jumlah Perkara Perdata yang tidak berhasil dimediasi
1.	Pengadilan Negeri Manado	948	32
2.	Pengadilan Negeri Bitung	289	14
3.	Pengadilan Negeri Tondano	438	19
4.	Pengadilan Negeri Kotamobagu	212	9
5.	Pengadilan Negeri Amurang	132	7
6.	Pengadilan Negeri Airmadidi	365	19
7.	Pengadilan Negeri Tahuna	159	3
8.	Pengadilan Negeri Melonguane	69	6
Total		2591	109

Tabel 1.1 di atas menunjukkan rendahnya angka mediasi berhasil yaitu 109 perkara dalam 3 tahun atau 6.47 persen dari jumlah keseluruhan perkara yang wajib dimediasi. Dengan demikian, ada sebanyak 2503 perkara atau 93.53 persen dari total keseluruhan perkara yang dimediasi, yang harus ditangani oleh pengadilan melalui pemeriksaan dan persidangan. Selain itu, para pihak berperkara harus menjalani persidangan – persidangan yang memakan waktu lama, bisa berbulan – bulan bahkan bertahun – tahun.

Sebagai aturan tambahan untuk pelaksanaan mediasi, Mahkamah Agung telah menerbitkan Peraturan Mahkamah Agung RI Nomor 3 tahun 2022 tentang Mediasi elektronik di pengadilan. Peraturan ini mendorong pelaksanaan mediasi dari tatap

muka menjadi mediasi pada ruang virtual (dengan menggunakan aplikasi teleconference) serta pengiriman maupun pertukaran dokumen dilakukan secara elektronik. Tujuan utamanya adalah untuk mengurangi jumlah pertemuan tatap muka antara pihak berperkara yang seringkali berpotensi mediasi batal dilaksanakan karena para pihak tidak mau saling bertemu secara fisik.

Usulan tentang pemanfaatan ruang virtual dalam mediasi (yang selanjutnya disebut e-mediasi) di pengadilan negeri di Indonesia pernah dikemukakan oleh Lumbantoruan, dkk (Lumbantoruan et al., 2021). Penelitian tersebut menyatakan bahwa budaya, kepribadian para pihak dan bahkan tempat diadakannya mediasi dapat mempengaruhi keberhasilan mediasi perkara. Di dalamnya juga disebutkan, dengan melakukan e-mediasi para pihak dapat memperoleh akses keadilan yang lebih besar dalam menemukan penyelesaian sengketa, lebih memuaskan dan memenuhi rasa keadilan. E-mediasi, menurutnya, karena dilakukan pada tahap awal penanganan perkara dapat mengatasi masalah penumpukan perkara di pengadilan dan memaksimalkan fungsi lembaga non peradilan (mediator non hakim) untuk penyelesaian sengketa disamping proses pengadilan yang bersifat memutus atau ajudikatif.

Terdapat penelitian – penelitian lain yang berfokus pada metode penyelesaian sengketa yang mengusulkan penggunaan teknologi informasi dalam rangka menjembatani kesenjangan yang terjadi antara penyelesaian sengketa modern maupun tradisional. Hal ini meningkat pada saat dan pasca Covid-19 sebagai konsekuensi dari social distancing dan lockdowns yang memaksa banyak hal dilakukan secara elektronik, termasuk mediasi (Adrian, 2021). Hal ini

menunjukkan bahwa terdapat kesadaran yang signifikan mengenai manfaat teknologi dalam ranah hukum (Alcántara Francia et al., 2022). Di antara penelitian tersebut adalah penelitian dari (Park et al., 2021) yang mengusulkan penggunaan metode supervised-learning pada data putusan pengadilan untuk melakukan prediksi hasil persidangan. Lamanya hukuman dan jumlah denda digunakan sebagai variabel tujuan. Penerapan teknologi hukum yang meramalkan hasil persidangan menurutnya dapat membantu hakim dalam mengambil keputusan dengan lebih cepat. Selain itu, bagi para pihak berperkara metode ini digunakan untuk seberapa besar kemungkinan mereka memenangkan suatu kasus. Penelitian ini menunjukkan bahwa prediksi artikel hukum dengan supervised learning dapat menemukan artikel hukum terkait untuk kasus tertentu. Selain itu, penelitian ini mendorong penggunaan dokumen terstruktur untuk analisis data dikarenakan sulit untuk memproses dan menganalisis dokumen mentah dalam format teks. Analisis dilakukan dengan metode ekstraksi fitur untuk mengubah dokumen tersebut menjadi bentuk tabel. Fitur-fitur tersebut harus diidentifikasi dan dikumpulkan dengan menerapkan teknik nontrivial seperti text mining dan NLP.

Zeleznikow pada tahun 2021 menyajikan suatu model untuk membangun sistem cerdas pada hybrid Online Dispute Resolution (ODR) yang berpusat pada pengguna, yang terdiri dari 6 fitur yaitu (1) Case management, (2) Triaging, (3) The provision of Advisory tools, (4) Communication tools, (5) Decision Support Tools dan (6) Drafting software and Agreement Technologies (Zeleznikow, 2021). Dari keenam fitur ini, menurutnya, belum ada suatu kondisi yang memungkinkan penggunaannya secara bersama – sama.

Pada tahun 2022, Cohen, dkk (Cohen et al., 2022) membahas bagaimana *data science* dan AI dapat membantu pengacara dan pihak yang berperkara memprediksi hasil hukum, termasuk perselisihan ketenagakerjaan, pelanggan, dan asuransi. Penelitian ini juga mengeksplorasi keterbatasan model *machine learning* di bidang hukum saat ini, terutama karena terbatasnya prediktabilitas hasil hukum. Penelitian ini meninjau cara-cara potensial untuk mengatasi keterbatasan penelitian *machine learning* tradisional dan pada akhirnya meningkatkan kekuatan model prediktif berbasis data. Selain dampak transformatif pada sistem peradilan, menurut penelitian ini kemajuan dalam penerapan AI pada hukum dan negosiasi dapat secara signifikan mengubah cara memahami penyelesaian sengketa. Misalnya, membuat model kesepakatan perdamaian yang diperkuat dengan keputusan pengadilan pada perkara serupa untuk meningkatkan prediksi atau membuat model yang menunjukkan poin-poin perbedaan antara kesepakatan perdamaian dan pertimbangan hukum pada putusan pengadilan sehingga dapat merekomendasi dan mempengaruhi putusan akhir.

Selain itu terdapat penelitian lain yang mengembangkan model dengan penggunaan *machine learning* dengan mengkombinasikan penggunaan algoritma prediksi dan klasifikasi untuk melakukan text-mining. Di antaranya adalah penelitian yang dilakukan oleh (Sueno et al., 2020) tentang sejumlah teknik vektorisasi yang digunakan untuk mengubah informasi teks ke dalam format numerik. Pemrosesan data vektorisasi dengan dimensi besar memerlukan sejumlah besar fitur yang dikonversi dari data teks dari satu halaman, yang memerlukan waktu. Studi ini melakukan vektorisasi dokumen berdasarkan distribusi

probabilitas yang menunjukkan kemungkinan kategori dokumen tersebut, sehingga mengurangi jumlah dimensi dengan metode Naïve Bayes yang disempurnakan.

Khoirunnisa, dkk, pada tahun 2020 juga mengkaji pengaruh N-Gram terhadap klasifikasi dokumen berita dengan algoritma Naïve Bayes Classifier yaitu menghasilkan akurasi sebesar 32,68% dan tanpa menerapkan N-Gram sebesar 84,97%. Penurunan hasil klasifikasi disebabkan oleh banyaknya fitur yang dihasilkan dari pemecahan N-Gram yang unik atau dominan terhadap kategori lain. Dengan kata lain penerapan N-Gram dalam klasifikasi dokumen menggunakan algoritma Naïve Bayes Classifier memberikan efek penurunan kinerja klasifikasi (Khoirunnisa et al., 2020).

Penggunaan N-gram dalam text-mining telah dilakukan pula dalam penelitian untuk mendukung prediksi dan klasifikasi teks pada dokumen pengadilan di Inggris oleh Strickson dan La Iglesia (Strickson & De La Iglesia, 2020). Penelitian dilakukan dengan pembuatan data berlabel untuk serangkaian keputusan pengadilan dan selanjutnya diterapkan untuk model machine learning dengan menggunakan beberapa algoritma yaitu SVM, Logistic Regression, Random Forest, k-Nearest Neighbour, Single Layer Perceptron (SLP) dan Multi Layer Perceptron (MLP). Penelitian menghasilkan model prediktif yang berakurasi tinggi dan mudah diinterpretasikan pada kombinasi algoritma Logistic Regression dan TF-IDF dengan F1-score sebesar 69,02, Precision sebesar 69.05% dan recall sebesar 69.02%.

Penelitian lainnya dilakukan oleh Shaikh, dkk (Shaikh et al., 2020). Menurutnya, memprediksi hasil suatu perkara dapat membantu dalam memahami

proses pengambilan keputusan pengadilan. Prediksi dapat dilakukan berdasarkan yang pertama faktor hukum spesifik perkara, seperti jenis bukti, dan faktor kedua adalah hukum tambahan, misalnya arah ideologis pengadilan. Rincian fakta hukum spesifik perkara dapat diambil dari keputusan hukum. Dalam penelitian ini, dijabarkan faktor-faktor penting yang mempengaruhi hasil putusan pada perkara pidana pembunuhan (diambil dari Pengadilan Negeri Delhi) dengan dataset sebanyak 86 perkara. Perbandingan dilakukan pada algoritma Logistic Regression, k-Nearest Neighbour, Classification and Regression Trees (CART), Naïve Bayes, SVM, Bagging, Random Forest dan Boosting. Hasil akhir menunjukkan CART memiliki performa terbaik dalam hal Akurasi dan Skor F1. Namun semua algoritma klasifikasi berkinerja baik terbukti dari akurasi berkisar antara 85% dan 92% dan Skor F1 berkisar antara 86% dan 92%.

Penelitian lainnya dilakukan oleh Mandal, dkk (Mandal et al., 2022) yang mengusulkan model penandaan supervised neural sequence untuk ekstraksi kata kunci terhadap serangkaian dokumen perkara di Mahkamah Agung India. Argumen penelitian ini adalah bahwa kata kunci (catchphrase/keyphrase) adalah frasa khusus dokumen dan oleh karena itu model sequence labelling harus dapat menggunakan informasi spesifik dokumen untuk mengekstraksi kata kunci yang lebih akurat, sehingga mencapai Precision dan Recall yang lebih bahkan ketika mengidentifikasi serangkaian kata kunci yang lebih kecil dibandingkan metode lainnya. Secara khusus, penelitian ini juga menambahkan penyematan Doc2Vec ke lapisan tertentu dari model pelabelan urutan yang menggunakan blok GRU dua arah sebagai intinya, disebut D2V-BIGRU-CRF. Penambahan vektor dokumen ini

memungkinkan model mengekstraksi slogan spesifik untuk dokumen tertentu dengan lebih baik. Penelitian ini juga membandingkan penggunaan pendekatan n-gram dengan pendekatan *noun phrase* (frase kata benda) untuk melakukan ekstraksi pada dokumen putusan pengadilan dan menemukan bahwa pendekatan noun phrase memberikan hasil pengujian (Akurasi, presisi, recall dan F1 score) yang lebih baik dibandingkan pendekatan n-gram.

Penelitian dengan memanfaatkan dokumen hasil mediasi perkara – perkara sebelumnya yang serupa untuk memprediksi hasil mediasi perkara baru dilakukan pada tahun 2022 oleh Hsieh, dkk (Hsieh et al., 2022) pada dokumen pengadilan berbahasa Cina yang bertujuan mengurangi beban pengadilan dalam penyelesaian perkara. Penelitian menggunakan framework LSTM (*Long Short Term Memory)* untuk mengolah data teks mediasi, dengan tools yang disebut LSTMEnsembler untuk memprediksi hasil mediasi dengan menggunakan beberapa *classifier*, diantaranya XGBoost dan LightGBM untuk modelling data numerik, serta TextCNN dan BERT untuk modelling data teks. Hasil eksperimen dalam penelitian tersebut menunjukkan bahwa LSTM Ensembler menghasilkan skor F sebesar 85,6% untuk data – data mediasi. Penelitian tersebut menggunakan karakter Bahasa Cina sebagai token dalam *textual mining* dan hasilnya menunjukkan bahwa teks vector berkontribusi positif terhadap kemampuan model, khususnya model BERT dalam memprediksi keberhasilan mediasi.

Masih sangat jarang ditemukan penelitian yang dilakukan dengan menggunakan dataset berbahasa Indonesia berupa dokumen mediasi maupun putusan pengadilan terdahulu dari pengadilan negeri di Indonesia untuk

memprediksi hasil mediasi. Berdasarkan uraian latar belakang dan berbeda dengan penelitian – penelitian sebelumnya, penelitian ini akan fokus pada pengambilan keputusan dengan melakukan klasifikasi pada data mediasi dan putusan terdahulu untuk memprediksi hasil mediasi dengan tujuan menyingkat waktu pelaksanaan mediasi. Selain itu, untuk memberikan gambaran kepada pihak berperkara tentang penyelesaian perkara sebagai upaya mendorong keberhasilan mediasi sehingga dapat mengurangi beban perkara perdata yang dilanjutkan ke tahapan pemeriksaan dan persidangan. Penelitian ini secara empiris akan melakukan perbandingan terhadap performa berbagai algoritma klasifikasi teks yang diterapkan pada dataset terpilih yang telah melalui proses ekstraksi fitur dengan metode n-gram, sehingga dapat diperoleh gambaran algoritma mana yang paling efektif menghasilkan prediksi hasil mediasi.

1.2. Rumusan Masalah

Berdasarkan latar belakang permasalahan pada bagian sebelumnya maka rumusan permasalahan yang akan menjadi fokus penelitian ini, yaitu:

a. Apakah dokumen hasil mediasi terdahulu dapat digunakan untuk memprediksi hasil mediasi pada perkara baru?
b. Bagaimana pengaruh hasil ekstraksi fitur dengan n-gram terhadap dataset pada model Machine Learning yang digunakan?
c. Berapa tingkat akurasi dari model - model yang diusulkan dan sejauh mana model tersebut berkontribusi pada penghematan waktu mediasi?

1.3. Batasan Masalah

Agar penelitian ini terarah dan memiliki pembahasan yang relevan serta membatasi ruang lingkup yang luas maka ditentukan batasan - batasan masalah sebagai berikut:

a. Sebagai sumber sampel data adalah peradilan umum tingkat pertama atau Pengadilan Negeri dalam wilayah hukum Pengadilan Tinggi Sulawesi Utara sebanyak 8 Pengadilan Negeri yaitu Pengadilan Negeri Manado, Pengadilan Negeri Tondano, Pengadilan Negeri Bitung, Pengadilan Negeri Kotamobagu, Pengadilan Negeri Tahuna, Pengadilan Negeri Airmadidi, Pengadilan Negeri Amurang dan Pengadilan Negeri Melonguane;
b. Dataset yang digunakan adalah dataset berbahasa Indonesia berupa hasil Mediasi pada perkara perdata yang telah dimediasi pada rentang waktu tahun 2020 sampai dengan tahun 2023, yaitu perkara perdata yang mengecualikan kategori sebagaimana Pasal 4 Ayat 2 Peraturan Mahkamah Agung RI Tahun 2016 Tentang Mediasi sebanyak 2591 perkara. Atribut dataset yang akan digunakan adalah yang mempengaruhi kesimpulan hasil mediasi, yaitu: nomor perkara atau nomor penetapan mediasi, waktu mulai mediasi, waktu selesai mediasi, petitum, dan status hasil mediasi;
c. Untuk Ekstraksi Fitur dibatasi pada penggunaan metode n-gram
d. Untuk pembobotan kata dibatasi pada metode Term Frequency-Inverse Document Frequency (TF-IDF)
e. Compiler menggunakan Bahasa Python

f. Penelitian ini tidak sampai pada pembuatan GUI (Graphical User Interface) namun hanya pada pembuatan model matematis dengan menggunakan algoritma Naïve Bayes (NB), Logistic regression (LR), Decision tree (DT) dan Support Vector Machine (SVM).

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk:

a. Menggunakan dokumen – dokumen mediasi terdahulu untuk memprediksi keberhasilan mediasi pada perkara serupa lainnya
b. Menggunakan pendekatan n-gram untuk mengekstraksi fitur pada teks
c. Membuat model prediksi keberhasilan mediasi
d. Menguji tingkat akurasi serta performa algoritma yang digunakan untuk memprediksi keberhasilan mediasi
e. Memberikan kontribusi mengurangi waktu pelaksanaan mediasi dan mengurangi beban perkara pengadilan dengan mengoptimalkan prediksi hasil mediasi

1.5. Manfaat Penelitian

Dengan adanya penelitian ini, diharapkan dapat membawa manfaat:

a. Bagi pembaca, memberikan kontribusi pengetahuan dan wawasan baru tentang prediksi;
b. Bagi penulis, adanya penelitian ini dapat menjadi sarana implementasi ilmu yang diperoleh serta membawa pembaruan pada pengetahuan tentang machine learning terutama pada ranah hukum secara umum dan secara khusus tentang pelaksanaan mediasi;

c. Bagi peneliti lain, penelitian ini menjadi dasar pengembangan pengetahuan ke depan sehubungan bidang – bidang terkait, terutama yang menjadi rekomendasi penelitian;

d. Bagi bidang terkait dan bagi organisasi, penelitian ini bermanfaat mengimplementasikan machine learning pada bidang hukum, terutama pada penanganan perkara perdata di pengadilan negeri (pengadilan tingkat pertama);

e. Mengoptimalisasi pelaksanaan mediasi, mendorong keberhasilan mediasi dan mengurangi beban persidangan di pengadilan;

f. Mempersingkat waktu mediasi dan mendorong terselenggaranya dokumentasi mediasi yang mumpuni sebagai knowledge base untuk pengembangan mediasi elektronik di pengadilan

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian terkait topik yang diangkat dalam penelitian ini adalah sebagai berikut:

A. Penelitian terkait penyelesaian sengketa dengan menggunakan teknologi informasi:

1) Penelitian dari El Jelali, dkk (El Jelali et al., 2015) menyebutkan bahwa dibutuhkan 90 hari untuk menemukan kesepakatan antara pihak yang bersengketa. Inilah menurutnya dimana terjadi pergesaran dari ADR (Alternative Dispute Resolution) menjadi ODR (Online Dispute resolution). ODR adalah tipe penyelesaian sengketa dengan memanfaatkan teknologi dan internet untuk memfasilitasi dan mempercepat resolusi penyelesaian sengketa di luar pengadilan. Di antara skema ODR, eMediasi menjadi sebuah metodologi untuk mendorong penyelesaian positif sehingga tercapai kesepakatan di antara pihak-pihak yang berperkara dengan menyediakan alat dan fungsi untuk mediasi online. Secara khusus, untuk memungkinkan penerapan sistem eMediasi, dua persyaratan utama harus dipenuhi: (1) argumentasi bahasa alami dari klaim dan persyaratan yang terkait dengan perkara dan (2) pemahaman atas perkara serupa melalui proses konsultasi dan penyamaan persepsi didasarkan pada keputusan pengadilan sebelumnya. Alasan utama yang mendasari persyaratan ini adalah perlunya meningkatkan kesadaran pihak yang berperkara dan mediator mengenai sengketa tersebut. Meninjau perkara - perkara serupa untuk memahami hak dan kewajiban, norma-norma yang relevan, waktu dan biaya dari proses persidangan di pengadilan dan kemungkinan hasil dari perselisihan tersebut

dapat meningkatkan pengetahuan para pihak yang terlibat dalam proses emediasi.

2) Penelitian oleh Zeleznikow (Zeleznikow, 2021) telah menyelidiki dua jenis sistem komputer yang memberikan saran cerdas untuk mendukung proses negosiasi yaitu Sistem Pendukung Negosiasi dan Sistem Penyelesaian Sengketa Online (ODR). Penelitian ini kemudian mengusulkan pembangunan sistem cerdas yang berpusat pada pengguna. Sistem terdiri dari 6 fitur yaitu (1) Case management, (2) Triaging, (3) The provision of Advisory tools, (4) Communication tools, (5) Decision Support Tools dan (6) Drafting software and Agreement Technologies. Menurutnya, sangat sedikit sengketa yang memerlukan penggunaan keenam proses tersebut sekaligus. Semua sistem ODR mencakup langkah 4 (komunikasi) dan sebagian besar sistem sekarang mencakup langkah 1 (manajemen kasus). Membangun sistem ODR hybrid yang menggunakan keenamnya membutuhkan anggaran dan waktu yang tidak sedikit, tidak saja karena membutuhkan semua fitur tersebut tapi juga memastikan bahwa semuanya saling terhubung satu dengan yang lain. Penelitian ini kemudian mengusulkan penelitian lebih lanjut untuk pembangunan sistem cerdas menggunakan keenam fitur tersebut.
3) Pada tahun 2021, terdapat suatu usulan dari Lumbantoruan, dkk (Lumbantoruan et al., 2021) untuk melengkapi sistem informasi e-litigasi dengan e-mediasi yaitu dengan menyediakan ruang virtual bagi pelaksanaan mediasi perkara di Pengadilan Negeri di Indonesia terutama karena terjadinya pandemik covid-19 yang tidak memungkinkan adanya pertemuan tatap muka. Penelitian ini mengusulkan penggunaan ruang virtual bagi para pihak berperkara yang terintegrasi dengan sistem yang dimiliki oleh pengadilan yaitu Sistem Informasi Penelusuran Perkara (SIPP). Dengan adanya e-mediasi dimaksud, menurut penelitian ini, para pihak dapat

memperoleh akses keadilan yang lebih besar dalam menemukan penyelesaian sengketa, lebih memuaskan dan memenuhi rasa keadilan.

4) Penelitian dari Cohen, dkk (Cohen et al., 2022) membahas bagaimana data science dan AI dapat digunakan untuk membantu pengacara dan pihak yang berperkara dalam memprediksi hasil hukum, termasuk pekerjaan terkini mengenai perselisihan ketenagakerjaan, pelanggan, dan asuransi. Penelitian ini juga mengeksplorasi keterbatasan model machine learning di bidang hukum saat ini, yaitu terbatasnya prediktabilitas hasil hukum. Terakhir, ditinjau cara-cara potensial untuk mengatasi keterbatasan penelitian pembelajaran mesin tradisional yang pada akhirnya disimpulkan dapat meningkatkan kekuatan prediktif dari suatu model berbasis data. Namun di sisi lain, Cohen, dkk juga berpendapat bahwa membuat prediksi hanya berdasarkan data hukum dapat menimbulkan masalah dan menghasilkan prediksi yang tidak akurat, karena menurutnya data hukum tidak dapat mewakili cara penyelesaian sebagian besar perselisihan sehingga tetap dibutuhkan jasa seorang ahli. Dalam penelitian ini disebutkan pula secara singkat beberapa kemajuan terkini yang menemui persimpangan antara deep learning dan hukum. Bahwa selain dampak transformatif pada sistem peradilan, kemajuan dalam penerapan AI pada hukum dan negosiasi dapat secara signifikan mengubah cara kita memahami penyelesaian sengketa. Misalnya, membuat model perjanjian penyelesaian bersama dengan keputusan pengadilan dapat lebih meningkatkan prediksi. Hal ini juga akan menyoroti poin-poin perbedaan antara perjanjian penyelesaian dan keputusan pengadilan, serta menjelaskan sejauh mana bias tertentu dapat mempengaruhi putusan pengadilan. Pada akhirnya, hal ini dapat membantu mengungkap apakah beberapa kelompok individu menerima kesepakatan penyelesaian yang lebih baik dibandingkan kelompok lainnya, sehingga meningkatkan keadilan dan kesetaraan.

Penelitian ini juga merekomendasikan penelitian AI yang dapat membuka batas baru dalam mengidentifikasi hubungan sebab akibat dan penalaran kontrafaktual, yang merupakan masalah inti dalam data science dan ekonomi. Hal ini akan sangat berdampak apabila model pembelajaran dapat dikembangkan secara memadai untuk memahami teks dan penalaran peradilan. Rekomendasi selanjutnya dari penelitian ini adalah pengembangan representasi semantik dari reasoning, terutama karena penalaran kausalitas dan kontrafaktual merupakan komponen penting dari data hukum dan negosiasi. Penelitian ini juga mengusulkan inovasi dalam penelitian AI yang berfokus pada model deep learning yang dapat digunakan untuk mengembangkan sistem penyelesaian sengketa secara komprehensif dengan mempertimbangkan konsep – konsep hukum.

B. Penelitian tentang Prediksi menggunakan algoritma Machine Learning untuk melakukan text mining terhadap dokumen – dokumen pengadilan,

1) Penelitian dari Zadgaokar, dkk pada tahun 2021(Zadgaonkar & Agrawal, 2021) tentang perbaikan atas kelemahan pemrosesan manual adalah pemrosesan dokumen hukum secara otomatis, yang akan sangat bermanfaat bagi pemahaman masyarakat umum terhadap sistem hukum. Makalah ini mengkaji perkembangan terkini di bidang pengolahan teks hukum dan menawarkan studi perbandingan berbagai metode yang digunakan di dalamnya. Kami telah memisahkan metodologi yang digunakan dalam pekerjaan ini menjadi tiga kelas: berbasis KBP, berbasis pembelajaran mendalam, dan berbasis NLP. Kami memberikan perhatian khusus pada pendekatan KBP karena kami yakin pendekatan ini mampu menangani seluk-beluk bidang hukum. Terakhir, kami membahas beberapa jalur penelitian potensial di masa depan untuk pemrosesan dan analisis dokumen hukum.

2) Penelitian oleh (Sullivan & Beel, 2019) menentukan seberapa akurat penilaian dapat dibuat ECHR (European Court of Human Rights) menjadi dapat diprediksi. Hal ini dilakukan dengan menggunakan dokumen Putusan akhir yang dihasilkan oleh ECHR, sebagai masukan. Menggunakan teknik Natural Language Processing (NLP), fitur tekstual diperoleh dari dokumen-dokumen ini. Model machine learning kemudian dilatih, menggunakan fitur-fitur ini, untuk memprediksi apakah suatu aplikasi mengandung frase mengakibatkan "pelanggaran" atau "tidak adanya pelanggaran" terhadap hak asasi manusia. Pada akhirnya, sebuah model prediktif dapat digunakan untuk membantu mengatasi backlog aplikasi. ECHR dapat menggunakan model prediksi yang akurat untuk membuat atau membantu membuat penilaian. Model seperti ini juga dapat digunakan untuk menentukan prioritas perkara.
3) Penelitian dari (Medvedeva et al., 2020) melakukan beberapa eksperimen yang melibatkan analisis bahasa dalam putusan EHCR untuk memprediksi apakah hal tersebut akan terjadi suatu perkara dinilai melanggar hak seseorang atau tidak. Hasil menunjukkan hal itu menggunakan informasi yang relatif sederhana dan dapat diperoleh secara otomatis. Model yang dihasilkan mampu memprediksi keputusan dengan benar pada sekitar 75% kasus, dan ini jauh lebih tinggi dari peluang kinerja 50%. Telah dibahas kemungkinan analisis bobot yang ditetapkan ke frasa berbeda oleh algoritme machine learning, dan bagaimana hal ini dapat digunakan untuk mengidentifikasi pola dalam teks persidangan.
4) Penelitian dari (Strickson & De La Iglesia, 2020) meneliti dua masalah saat mencoba Legal Judgement Prediction (LJP) pada dokumen pengadilan di Inggris. Yang pertama adalah terbatasnya kemampuan teknik Natural Language Processing (NLP) untuk mengenali struktur semantik yang kompleks seperti argumen. Masalah kedua khusus terjadi di Inggris; saat ini

tidak ada kumpulan data publik terstruktur mengenai kasus-kasus pengadilan di Inggris. Tujuan penelitian ini adalah untuk membangun model prediktif yang dapat ditafsirkan untuk kasus-kasus pengadilan Inggris hanya dengan menggunakan dokumen pengadilan. Tujuan penelitian ini adalah: (1) Untuk membuat kumpulan data keputusan pengadilan Inggris yang diberi label dengan variabel hasil yang dapat digunakan dalam tugas prediksi. (2) Untuk membangun model prediksi menggunakan teknik machine learning yang sebelumnya diterapkan oleh penelitian pembanding. (3) Untuk menguji teknik text mining alternatif seperti fitur penyematan kata dengan model jaringan saraf tiruan. Dalam penelitian ini dijelaskan bagaimana kumpulan data berlabel dibangun untuk keputusan pengadilan Inggris dan kemudian digunakan untuk menguji teknik text yang ada dan yang lebih baru untuk mendapatkan akurasi yang baik dan fitur yang diharapkan.

7) Penelitian oleh (Alghazzawi et al., 2022) yang berangkat dari peningkatan signifikan pada jumlah dokumen peradilan, sehingga dirasa penting untuk mengumpulkan dan menganalisis data tersebut untuk memperkirakan keputusan pengadilan. Untuk melakukan hal ini, model deep learning yang efisien dirancang dan dikembangkan. Model yang disarankan memiliki tiga peran yaitu: (i) memperoleh tolok ukur untuk pengumpulan data peradilan, (ii) pemilihan fitur, dan (iii) memprediksi putusan perkara pengadilan neural network yaitu LSTM+CNN. Beberapa pengujian juga dilakukan dengan menggunakan kumpulan data. Dalam pengumpulan data yang disediakan, pemilihan fitur digunakan untuk memilih fitur-fitur utama saja dengan memprioritaskan dan memilih fitur-fitur dengan peringkat teratas. Terakhir, model LSTM+LSTM digunakan untuk meramalkan putusan perkara pengadilan. Hasil eksperimen yang ada cukup baik. Namun demikian, terdapat kekurangan yang signifikan dari model yang disarankan berikut ini:

(i) sejumlah kecil data dari domain tertentu yang digunakan (kumpulan data peradilan standar), (ii) hanya satu teknik statistik, yaitu pengukuran RFE, yang diterapkan pada kumpulan masukan. data untuk memilih fitur-fitur penting (variabel prediktor), (iii) penyematan digunakan daripada model CNN yang telah dilatih sebelumnya, (iv) tidak ada teknik pengurangan noise yang efisien, dan (v) akan lebih baik jika model tersebut bekerja dengan cara yang sama dari waktu ke waktu, dalam kasus yang berbeda, dan bahkan dengan hakim yang berbeda. Kemanjuran model-model yang ada saat ini juga sangat bervariasi dari waktu ke waktu dan antar hakim. Hal ini terbukti menjadi tantangan bagi para ahli hukum dan praktisi hukum untuk memanfaatkan model prediksi secara efektif.

C. Penelitian lain tentang text mining menggunakan pendekatan n-gram:

1) Penelitian oleh (Vernanda et al., 2020) dengan mendeteksi email spam berbahasa Indonesia dengan menggunakan metode text mining yaitu Naïve Bayes classifier yang disempurnakan dengan metode N-gram. Pada penelitian ini diusulkan dan diimplementasikan metode Naïve Bayes dan N-gram sebagai layanan web menggunakan desain REST API. Penelitian ini menyimpulkan bahwa nilai akurasi berkisar antara 0,615 hingga 0,94, nilai presisi berkisar antara 0,566 hingga 0,924, nilai recall berkisar antara 0,96 hingga 1, dan nilai f-measure berkisar antara 0,721 hingga 0,942. Metode 6 gram dan yang lebih baru tidak mengalami perubahan yang berarti. Sedangkan metode N-gram terbaik yang memberikan nilai akurasi, presisi, dan f-measure tertinggi dalam mendeteksi spam berbahasa Indonesia adalah metode 5 gram jika dipadukan dengan algoritma Naïve Bayes.
2) Penelitian oleh (Avasthi et al., 2021) yang melakukan penelitian terkait dokumen terstruktur dan penggunaan teknik untuk meningkatkan pengalaman pengguna adalah kemampuan untuk memprediksi kata, huruf, atau kalimat berikutnya saat pengguna mengetik ke aplikasi yang

meminimalkan upaya mengetik dan kesalahan ejaan. Menurutnya, karena platform media sosial lebih sering digunakan, maka ukuran data teks pun bertambah. Tujuan utama dari artikel penelitian ini adalah untuk memproses korpus teks besar dan menerapkan model probabilistik yang mirip dengan N-gram untuk memprediksi kata berikutnya ketika pengguna memberikan masukan. Dalam penelitian eksplorasi ini, model n-gram dibahas dan dievaluasi menggunakan Good Turing Estimation, perplexity-measure, dan rasio tipe-to-token.

3) Penelitian oleh Khoirunnisa, dkk dilakukan untuk mengkaji pengaruh n-gram terhadap klasifikasi dokumen pada algoritma Naïve Bayes Classifier. Tujuan pemrosesan teks adalah untuk mengetahui dan mengekstrak informasi yang berguna dari sumber data dengan cara mengidentifikasi dan mengeksplorasi pola menarik dalam *text mining*. Sumber data yang digunakan merupakan kumpulan dokumen yang tidak terstruktur dan diperoleh secara manual dengan mengunduh dokumen teks berbahasa Indonesia dalam format teks (*.txt) di situs http://news.kompas.com dan http://www.republika. Dokumen tersebut terdiri dari lima kategori berita online, yaitu berita kesehatan, berita politik, berita ekonomi, berita teknologi, dan berita olahraga. Setiap kategori terdiri dari 60 buah dokumen teks berjenis teks (*.txt). Perbandingan yang dilakukan menunjukkan bahwa hasil klasifikasi dokumen pada algoritma Naïve Bayes Classifier menggunakan pemodelan N-Gram menghasilkan akurasi yang lebih rendah, jika dibandingkan dengan klasifikasi dokumen dengan algoritma Naïve Bayes Classifier tanpa menggunakan pemodelan N-Gram. Hal ini menunjukkan bahwa penggunaan model N-Gram dapat mempengaruhi hasil klasifikasi pada dokumen berita online. Efeknya karena sistem dapat lebih akurat dalam mengklasifikasikan menggunakan kata dibandingkan karakter yang dihasilkan dari proses pemodelan N-Gram. Akurasinya

adalah sebesar 32,68% dan hasil akurasi tanpa menggunakan pemodelan N-Gram yaitu 84,97% dan pada penelitian ini pengaruh proses pemodelan N-Gram terhadap hasil klasifikasi dokumen berita online terletak pada dokumen kategori kesehatan (Khoirunnisa et al., 2020).

4) Penelitian oleh (Kruczek et al., 2020) menunjukkan tiga domain: atribusi kepenulisan, profil penulis, dan analisis sentimen bahwa pilihan jenis n-gram hanya menghasilkan sedikit peningkatan akurasi klasifikasi dibandingkan n-gram tradisional. Informasi tentang profil penulis didistribusikan ke seluruh kategori n-gram. Tidak ada satu kategori pun yang dapat disarankan untuk klasifikasi. Penelitian ini menyarankan untuk melakukan upaya lebih besar dalam pengoptimalan hyperparameter dan pemilihan model yang efektif dibandingkan beralih dari n-gram ke n-gram yang diketik atau kategori tertentu dari n-gram yang diketik.
5) Penelitian oleh (Georgieva-Trifonova & Duraku, 2021) dikhususkan untuk penerapan metode pemilihan fitur untuk klasifikasi teks. Untuk tujuan ini, penyajian dokumen teks didasarkan pada N-gram kata yang diekstraksi. Fitur metode pemilihan dijalankan dan perubahan kinerja klasifikasi dalam hal akurasi dan pengukuran F dilacak dengan jumlah atribut yang dipilih berbeda untuk classifier yang berbeda dan kumpulan data tertentu. Hasil yang diperoleh diilustrasikan dan kegunaannya untuk penggunaan di masa depan guna meningkatkan kinerja klasifikasi teks.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian
Komparasi Algoritma Klasifikasi Teks dengan Metode Ekstraksi N-gram pada Hasil Mediasi Perkara Perdata di Pengadilan Negeri

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Predicting the Success of Mediation Requests Using Case Properties and Textual Information for Reducing the Burden on the Court	Hsieh, Hsun-Ping, et al. *Digital Government: Research and Practice* 2.4 (2022): 1-18. © 2022 Association for Computing Machinery. 2639-0175/2022/01-ART30 $15.00 https://doi.org/10.1145/3469233	untuk membangun kerangka kerja yang efektif, disebut LSTMensembler, dengan enganalisis data dan memprediksi apakah ada kasus perselisihan di komite mediasi akan diselesaikan dengan sukses, yang berarti bahwa kedua belah pihak encapai kesepakatan secara damai berdasarkan konsiliasi dari mediator	Hasil percobaan menunjukkan bahwa teks vektor memiliki kontribusi positif terhadap keterampilan model, khususnya model BERT. Selain itu, di antara semua kombinasi yang digunakan, ditemukan bahwa dengan tiga classifier yang berbeda, vektor BERT mencapai kinerja terbaik.	Di masa mendatang, penulis ingin meningkatkan kemampuan menjelaskan framework dan meningkatkan kinerjanya. Salah satunya adalah dengan mengembangkan mekanisme perhatian yang mempelajari pentingnya setiap aspek representasi fitur. Cara menghasilkan peringkat mediator yang direko-mendasikan dan andal untuk setiap kasus juga merupakan topik yang menarik dan praktis.	Penelitian Hsieh,dkk menggunakan dataset pengadilan berbahasa Cina dengan metode LSTM untuk melakukan prediksi sementara penelitian yang dirancang akan melakukan komparasi penggunaan algoritma Naïve Bayes, Logistic regression, Decision Tree dan SVM dan metode n-gram untuk mengolah data teks

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
2	Using Artificial Intelligence to provide intelligent Dispute Resolution Support	John Zeleznikow1 Accepted: 29 March 2021 / Published online: 13 April 2021 © The Author(s), under exclusive licence to Springer Nature B.V. 2021. Group Decision and Negotiation (2021) 30:789–812 https://doi.org/10.1007/s10726-021-09734-1	- Melakukan eksaminasi terhadap proses – proses Kecerdasan buatan untuk memperoleh manfaat bagi mediator. - Melakukan investigasi terhadap sistem komputer yang menyediakan intelligent advice untuk mendukung proses negosiasi baik Sistem Pendukung Negosiasii maupun Sistem Penyelesaian Sengketa Online	Penelitian berhasil mengembangkan model untuk membangun Penyelesaian Sengketa Online yang berpusat pada pengguna sistem. Model ini mengintegrasikan fitur-fitur berikut: (1) Manajemen kasus, (2) Triase, (3) Penyediaan Sarana Sarana, (4) Alat Komunikasi, (5) Penetapan Alat Pendukung dan (6) perangkat lunak dan Teknologi Penyusunan Perjanjian.	- Model ini belum dapat digunakan untuk perselisihan individu - Model ini dapat dikembangkan untuk menbangun Sistem Pendukung Negosiasi Sistem Penyelesaian Sengketa Online yang berbasis kecerdasan buatan	Penelitian ini tidak secara spesifik menggunakan algoritma Artificial Intelligence untuk melakukan perbandingan dan eksaminiasi namun lebih kepada riset tentang komponen – komponen penting yang harus disediakan pada saat membangun model untuk menbangun Sistem Pendukung Negosiasi Sistem Penyelesaian Sengketa Online yang berbasis kecerdasan buatan, sementara penelitian yang dirancang akan menggunakan algoritma Naïve bayes, Logistic Regression, Decision Tree dan SVM dengan pendekatan n-gram untuk mengolah data teks dan membandingkan tingkat akurasinya.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Legal Judgement Prediction for UK Courts	Benjamin Strickson, Beatriz De La Iglesia. ACM International Conference Proceeding Series, 2020, 204-209	- Untuk membangun dataset berlabel putusan pengadilan Inggris dengan variabel hasil yang dapat digunakan dalam tugas prediksi. - Untuk membangun model prediksi menggunakan teknik machine learning yang sebelumnya diterapkan oleh studi pembanding. - Untuk menguji teknik text mining alternatif seperti fitur penyisipan kata dengan model jaringan saraf tiruan.	- Tercipta putusan berlabel. - Fitur TFIDF yang dipasangkan dengan algoritme LR mencapai skor F1 tertinggi sebesar 69,02. Mengekstraksi fitur paling penting dari ruang vektor dan model cluster topik adalah tugas yang relatif mudah dan menunjukkan kegunaan model potensial yang baik. - Tujuan ketiga dicapai dengan penerapan penyisipan kata dan JST pada tugas – tugas LJP	Hasil penelitian ini tidak dapat menunjukkan bahwa penyematan kata yang dikombinasikan dengan jaringan saraf pilihan kami dapat meningkatkan kinerja model. Peneliti merekomendasikan future work untuk penggunaan algoritma JST yang advance untuk melakukan klasifikasi teks untuk meningkatkan LJP. Peneliti juga merekomendasikan pengujian lebih jauh pada penggunaan n-gram dan clustering topik dengan melakukan eksaminasi independen dan pengujian langsung terhadap fitur – fitur teerkstrak oleh pengacara	Penelitian ini menggunakan dataset putusan pengadilan Inggris dan melakukan pengujian dengan membandingkan akurasi hasil kombinasi beberapa algoritma dengan TF-IDF. Penelitian yang dirancang menggunakan dataset mediasi perkara perdata yang berbahasa Indonesia, akan menggunakan beberapa algoritma klasifikasi teks yang dikombinasikan dengan n-gram kemudian diuji tingkat akurasinya masing – masing.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
4	Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique.	Sueno,H. T., Gerardo, B. D., & Medina, R. P. (2020). *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3), 3937–3944. https://doi.org/10.30534/ijatcse/2020/216932020	- mengurangi dimensi data -untuk memvektorisasi dokumen menurut distribusi probabilitas	Berdasarkan penelitian ini, Naive Bayes-SVM classifier yang ditingkatkan mengungguli TF-IDF mengklasifikasikan dokumen dengan sangat akurat. Penggunaan Laplace smoothing hingga penyempurnaan Naive Bayes-SVM telah mencapai klasifikasi dengan akurasi lebih tinggi dibandingkan dengan TF-IDF. Hasil ini menunjukkan bahwa teknik vektorisasi Naive Bayes memberikan kontribusi proses transformasi data tekstual yang lebih efektif ke pengklasifikasi SVM, dibandingkan dengan penggunaan teknik vektorisasi TFIDF untuk tujuan yang sama.	Melakukan eksplorasi fitur dan bobot lain untuk menghasilkan vektor kata dan menyelidiki pengaruh metode smoothing lainnya terhadap vektorisasi Naive Bayes	Penelitian yang dirancang akan menggunakan dataset berupa dokumen hasil mediasi dan menggunakan metode ekstaksi fitur n-gram yang disandingkan dengan TF-IDF. Perbedaan juga terdapat dalam algoritma yang di gunakan yai'tyu NB,DT,LR dan SVM

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	Effect of N-Gram on Document Classification on the Naïve Bayes Classifier Algorithm	Fitria Khoirunnisa, Novi Yusliani, M.T, Desty Rodiah, M.T. Sriwijaya Journal of Informatic and Applications 1.1, 2020, 26-33 DOI: https://doi.org/10.36706/sjia.v1i1.13	Untuk meneliti pengaruh N-Gram terhadap akurasi hasil klasifikasi dokumen menggunakan algoritma Naïve Bayes Classifier	proses pemodelan menggunakan N-Gram dapat memberikan pengaruh terhadap hasil klasifikasi dokumen berita online dengan menggunakan algoritma Naïve Bayes Classifier, dimana akurasi sebesar 32,68% dan hasil akurasi tanpa menggunakan pemodelan N-Gram sebesar 84,97% dan pada penelitian ini pengaruh proses pemodelan N-Gram pada hasil klasifikasi dokumen berita online terletak pada dokumen kategori kesehatan. Jumlah karakter yang dihasilkan dari proses pemodelan N-Gram dominan terhadap kategori kesehatan.	Penelitian ini diharapkan dapat menerapkan metode text mining lainnya dalam kategorisasi dokumen selain menerapkan algoritma Naïve Bayes Classifier yang dikombinasikan dengan pemodelan N-Gram dan menerapkan metode yang dapat menentukan posisi karakter dalam suatu term dalam proses klasifikasi dalam pemodelan N-Gram	Penelitian ini mengkombinasikan algoritma Naïve Bayes Classifier dengan n-gram justru menghasilkan akurasi yang rendah pada dataset yang berasal dari berita. Penelitian yang direncanakan akan menggunakan dataset yang berbeda yaitu dokumen

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6	Processing large text corpus using N-gram language modeling and smoothing.	Zudha Avasthi, Sandhya Chauhan, Ritu Acharjya., and Debi Prasanna Proceedings of the Second International Conference on Information Management and Machine Intelligence. Lecture Notes in Networks and Systems, vol 166. Springer, Singapore. https://doi.org/10.1007/978-981-15-9689-6_3	memproses korpus teks besar dan menerapkan model probabilistik seperti N-gram untuk memprediksi kata berikutnya saat pengguna memberikan masukan.	Kosakata, jenis kata, dan keragaman kumpulan data eksperimental dihitung dan hubungan antara langkah-langkah ini diamati. Analisisnya menjelaskan hubungannya dengan jumlah kata dalam korpus berbanding terbalik. Peningkatan algoritma menyeimbangkan kedalaman informasi dengan kecepatan Pemrosesan Analisis menjelaskan model prediksi berguna untuk memprediksi asal kata umum dan nilai confusion matriksnya berkurang. Ini membuktikan bahwa stopword harus selalu disertakan dalam model prediktif	Pekerjaan di masa depan mencakup peningkatan aplikasi online untuk memungkinkan pemahaman yang lebih baik tentang masukan pengguna, memanfaatkan informasi kontekstual untuk meningkatkan akurasi prediksi.	Perbedaan terletak pada pemilihan algoritma. Penelitian yang dirujuk menggunakan Good Turing Estimation, sedangkan penelitian yang akan dilakukan menggunakan algoritma klasifikasi teks berupa NB, LR, DT dan SVM

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
7	Efficient Prediction of Court Judgments Using an LSTM + CNN Neural Network Model with an Optimal Feature Set	Alghazzawi, Daniyal Bamasag, Omaimah Albeshri, Aiiad Sana, Iqra Ullah, Hayat, MDPI.com https://doi.org/10.2290/math10050683 2022	RO1: Untuk memperkirakan keputusan kasus pengadilan menggunakan LSTM+CNN berdasarkan data historis peradilan. RO2: Untuk membandingkan efisiensi yang disarankan Pendekatan LSTM+CNN dengan pembelajaran mesin dan teknik pembelajaran mendalam RO3: Untuk membandingkan keefektifan teknik yang diusulkan dalam memperkirakan putusan perkara pengadilan dari catatan peradilan masa lalu	mengusulkan peramalan putusan pengadilan menggunakan model jaringan saraf hybrid, yaitu LSTM dengan CNN, untuk meramalkan putusan pengadilan secara efektif menggunakan kumpulan data historis peradilan. Dengan memprioritaskan dan memilih fitur-fitur yang memiliki skor tertinggi dalam kumpulan data hukum yang disediakan, hanya fitur-fitur yang paling relevan yang akan dipilih. Setelah itu, model LSTM+CNN digunakan untuk memperkirakan putusan gugatan	Penelitian ini merekomendasikan future work berupa penyelidikan pada penggunaan kumpulan data peradilan dari domain yang berbeda (data peradilan dari berbagai pengadilan), penggunaan metode pemilihan fitur tambahan selain RFE, penggunaan model terlatih seperti word2Vec, Glove, atau fastText, dan menyelidiki teknik pengurangan kebisingan yang canggih.	Penelitian ini menggunakan LSTM dan CNN untuk memprediksi putusan pengadilan berdasarkan dataset putusan pengadilan sebelumnya. dataset putusan pengadilan sebelumnya. Penelitian yang direncanakan menggunakan NB, LR, DT dan SVM untuk memprediksi hasil mediasi

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
8	E-Mediation in E-Litigation Stages in Court	Lumbantoruan, Parulian Mawuntu, Ronald Waha, Caecilia J J Tangkere, Cornelius. Journal of Law, Policy and Organization HeinOnline DOI: 10.7176/JLPG/108-02021	Mengoptimalisasi perdamaian melalui mediasi elektronik, sehingga para pihak atau advokat tidak perlu hadir di gedung Pengadilan namun mengajukan penawaran dan kesepakatan melalui sarana elektronik.	Untuk melengkapi e-litigasi di pengadilan, e-mediasi dapat menjadi solusi untuk menghadirkan ruang virtual bagi para pihak untuk menempuh mediasi. E-mediasi dapat mngurangi jumlah atau beban perkara di pengadilan karena dilakukan pada tahap awal. Dengan e-mediasi, para pihak dapat memperoleh akses keadilan yang lebih nyata dan lebih fair.	Penelitian ini hanya mengubah bentuk mediasi dari manual menjadi virtual tapi tata cara yang digunakan tetaplah sama. Mediator tetap perlu melakukan identifikasi masalah secara manual dan mencari aturan – aturan terkait perkara yang dimohonkan para pihak.	Penelitian ini hanya mengusulkan penggunaan ruang virtual untuk pelaksanaan mediasi. Penelitian yang direncanakan akan lebih jauh memanfaatkan dokumen – dokumen yang sudah ada untuk menjadi knowledge base bagi mediator untuk melaksanakan mediasi.
9.	Research on N-grams feature selection methods for text classification	Georgieva-Trifonova, Tsvetanka Duraku, Mahmut IOP Conference Series: Materials	Meningkatkan performa klasifikasi teks dengan seleksi fitur. Metode pemilihan fitur dilakukan dalam hal akurasi dan ukuran F klasifikasi teks dengan jumlah atribut yang dipilih	K-NN: Nilai Macro F-measure setelah menerapkan pemilihan fitur Information gain untuk jumlah fitur = 100 Decision tree: Pengukuran Makro F mendapatkan hasil yang lebih baik dengan	Hasil yang diperoleh dapat digunakan untuk menerapkan langkah pra-pemrosesan lebih lanjut, termasuk memodifikasi model ruang vektor untuk mencapai perbaikan dalam hal klasifikasi teks.	Penelitian ini menggunakan seleksi fitur N-gram untuk klasifikasi teks. Penelitian yang dirancang akan menerapkan algoritma NB, LR, DT dan SVM dengan pendekatan n-gram untuk melakukan prediksi pada dokumen

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
9.		Science and Engineering IOP Publishing Ltd DOI: 10.1088/1757-899X/1031/1/012048 ISSN: 1757899X 2021	(N-gram kata) yang berbeda untuk pengklasifikasi yang berbeda dan kumpulan data yang berbeda.	algoritma Relief ketika Jumlah fitur adalah 50, 100, 200. Untuk metode pemilihan fitur lainnya, kedua ukuran menunjukkan peningkatan untuk semua nilai Jumlah fitur. Deep Learning H2 menghasilkan Nilai yang lebih baik dari kedua ukuran yang diamati untuk semua metode pemilihan fitur untuk Jumlah fitur minimal 50. Rip: Terdapat perbaikan pada hasil ketika menerapkan semua fitur metode seleksi. Ridor: Pengukuran Makro F menunjukkan sedikit peningkatan		

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
9.				untuk Jumlah fitur = 500 saat diterapkan Pemilihan fitur perolehan informasi. Part: Kedua ukuran tersebut memberikan nilai yang lebih baik ketika menerapkan pemilihan fitur Information gain dan indeks Giniuntuk jumlah fitur minimal = 100		
10.	Predicting the Outcome of Judicial Decisions made by the European Court of Human Rights	Sullivan, Conor O. & Beel, Joeran Arxiv.org https://doi.org/10.48550/arXiv.1912.10819 27th AIAI Irish Conference on Artificial Intelligence	Untuk menentukan seberapa akurat penilaian yang dibuat oleh ECHR dapat diprediksi. Hal ini dilakukan dengan menggunakan dokumen Putusan akhir, yang dihasilkan oleh ECHR, sebagai input 1. Dengan menggunakan teknik Natural	Dengan menggunakan kumpulan data yang realistis, model mencapai rata-rata tertimbang sebesar 68,83% di seluruh pasal. Dengan menggunakan kumpulan data yang realistis, model tersebut mencapai rata-rata tertimbang sebesar 68,83% di seluruh Artikel. Dimana bobotnya diberikan berdasarkan	Keterbatasan penelitian ini adalah bahwa model yang dibangun hanya memberikan prediksi akhir untuk setiap Keputusan. Dimana tidak memberikan indikasi apa pun tentang bagaimana prediksi dibuat. Pada kenyataannya, Hakim harus membenarkan keputusannya sehingga mereka tidak dapat mengandalkan model	Topik penelitian ini adalah bagaimana memprediksi putusan pengadilan HAM di Eropa. Penelitian yang dirancang akan melakukan prediksi pada hasil mediasi.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
10.		and Cognitive Science 2019	Language Processing (NLP), fitur tekstual diperoleh dari dokumen-dokumen tersebut. Model pembelajaran mesin kemudian dilatih, menggunakan fitur-fitur ini, untuk memprediksi apakah suatu aplikasi telah menghasilkan "violation" atau "non-violation" terhadap hak asasi manusia.	jumlah Judgment dalam tes yang ditetapkan untuk setiap Pasal. Oleh karena itu, diperkirakan jika model tersebut digunakan oleh ECHR, lebih dari 30% keputusan mengenai pelanggaran hak asasi manusia akan salah. Konsekuensi dari hal ini bisa sangat parah mengingat Pengadilan ini dibentuk untuk melindungi hak asasi manusia. Seperti yang telah dibahas, model masih dapat menjadi alat yang berguna. Model tersebut dapat memberikan indikasi aplikasi mana dalam backlog yang harus diprioritaskan.	yang hanya memberikan prediksi akhir.	

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
11.	Are n-gram categories helpful in text classificat ion?	Kruczek, Jakub Kruczek, Paulina Kuta, Marcin Computational Science–ICCS 2020: 20th International Conference DOI: 10.1007/978-3-030-50417-5_39 Springer Science and Business Media Deutschland GmbH 2020	untuk memperluas penelitian dan menjawab pertanyaan tentang apakah n-gram yang diketik dapat menjadi fitur yang efektif dalam pembuatan profil penulis dan analisis sentimen seperti halnya dalam atribusi authorship	Makalah ini menunjukkan tiga domain: atribusi kepenulisan, profil penulis, dan analisis sentimen bahwa pilihan jenis n-gram hanya menghasilkan sedikit peningkatan akurasi klasifikasi dibandingkan n-gram tradisional. Informasi tentang profil penulis didistribusikan ke seluruh kategori n-gram. Tidak ada satu kategori pun yang disarankan untuk klasifikasi. Apache Spark memungkinkan klasifikasi yang efisien dengan jumlah fitur yang sangat banyak pada corpora teks besar. Jejak memori adalah aspek yang paling menghalangi klasifikasi tersebut, yang menghalangi eksperimen dengan n-gram yang lebih panjang dari 5	Merekomendasikan untuk mengoptimalkan hyperparameter dan pemilihan model secara efektif dibandingkan beralih dari n-gram ke n-gram yang diketik atau kategori tertentu dari n-gram yang diketik.	Dalam penelitian ini, n-gram digunakan untuk profiling dan atribusi penulis. Penelitian yang dirancang akan menggunakan n-gram untuk dikombinasikan dengan naïve bayes memprediksi hasil mediasi dengan menggunakan dataset berupa dokumen hasil mediasi terdahulu.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
12.	Using machine learning to predict decisions of the European Court of Human Rights	Medvedeva, Masha Vols, Michel Wieling, Martijn *Artificial Intelligence and Law* 28 (2020): 237-266. Springer	untuk menghadapi calon korban pelanggaran hak asasi manusia, memperkirakan keputusan apa yang akan diambil dalam kasus mereka. Menggunakan putusan pengadilan sebelumnya untuk memprediksi putusan bagi perkara baru.	Penelitian ini menggunakan informasi yang relatif sederhana dan dapat diperoleh secara otomatis, model yang dibuat mampu memprediksi keputusan dengan benar pada sekitar 75% kasus, yang jauh lebih tinggi daripada peluang performa sebesar 50%.	Penelitian ini merekomendasikan untuk melakukan evaluasi apakah mungkin untuk menggunakan pendekatan jaringan saraf untuk beberapa bagian pemrosesan data, sambil tetap mempertahankan kemampuan untuk menganalisis hasil sistem.	Penelitian ini menggunakan ngram pada dokumen putusan pengadilan HAM Eropa. Penelitian yang dirancang akan menggunakan ngram pada dokumen hasil mediasi pengadilan negeri di Indonesia.
13.	Legal retrieval as support to eMediation: matching disputant's case and court decisions	El Jelali, Soufiane Fersini, Elisabetta Messina, Enza Artificial Intelligence and Law Springer.	mengatasi masalah pencocokan perkara yang diajukan dengan putusan pengadilan (terutama diungkapkan dalam bahasa alami yang ringkas dan tanpa organisasi formal) yang biasanya terstruktur, panjang,	Sistem ini terdiri dari empat komponen utama: pengindeksan keputusan pengadilan, penambangan inti, pemeringkatan, dan pemrosesan kueri. Pertama-tama, putusan pengadilan diproses terlebih dahulu untuk disimpan dalam database. Dalam modul	untuk meningkatkan kinerja eJRM-IRS. Ide utamanya adalah mengekstraksi konsep semantik (disebut faktor) baik dari keputusan pengadilan maupun deskripsi kasus yang disengketakan untuk kemudian dicocokkan. Faktor-faktor ini, yang mewakili konsep-konsep	Peneltian ini memberikan rekomendasi kepada sistem eMediasi berdasarkan putusan pengadilan untuk perkara serupa. Penelitian yang dirancang akan melakukan prediksi dengan mengolah data teks hasil mediasi untuk memberikan fakta – fakta hukum kepada mediator dalam perkara serupa.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
13.		DOI: 10.1007/s10506-015-9162-1 2015	ditulis dengan baik dan ditandai dengan pola bahasa peradilan dengan mengusulkan pendekatan kecerdasan komputasi untuk pengambilan informasi hukum	mining, keputusan pengadilan direpresentasikan sebagai matriks TF-IDF untuk mengaktifkan teknik pembelajaran mesin. Untuk mengatasi dimensi yang disebabkan oleh sifat data, metode berbasis PCA telah dimanfaatkan sebagai strategi reduksi. Selanjutnya, mesin klasifikasi—yang mampu memprediksi bidang hukum terkait dengan deskripsi perkara telah dikembangkan. Untuk mengurangi noise dalam teks perkara, dan agar lebih sesuai dengan dokumen putusan pengadilan, algoritma Deteksi Istilah Khusus telah diusulkan. Terakhir, metrik kesamaan baru yang disebut Kesamaan Koherensi telah ditentukan	hukum yang berkaitan dengan bidang hukum tertentu, akan diekstraksi dengan mengikuti paradigma tanpa pengawasan (berdasarkan algoritma Deteksi Istilah Khusus) untuk menjaga independensi eJRM-IRS pada domain tertentu.	

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
14	Conflict Analytics: When Data Science Meets Dispute Resolution	Cohen, Maxime C Dahan, Samuel Rule, Colin Branting, L Karl *Manag Business Rev* 2.2 (2022): 86-93	Penelitian ini bertujuan untuk mendeskripsikan bagaimana data science dapat digunakan untuk memberikan rekomendasi, prediksi dan pertimbangan – pertimbangan hukum dalam resolusi penanganan perselisihan.	Penelitian ini telah membahas bagaimana data science dan AI dapat membantu pengacara dan pihak yang berperkara memprediksi hasil hukum, termasuk penelitian terbaru mengenai perselisihan ketenagakerjaan, pelanggan, dan asuransi, mengeksplorasi keterbatasan model pembelajaran mesin di bidang hukum saat ini, terutama terbatasnya prediktabilitas hasil hukum.	mengharapkan penelitian AI juga membuka batas baru dalam mengidentifikasi hubungan sebab akibat dan penalaran kontrafaktual, yang merupakan masalah inti dalam ilmu data dan ekonomi. Hal ini akan sangat berdampak apabila model pembelajaran dapat dikembangkan secara memadai untuk memahami teks dan penalaran hukum	Penelitian yang dirujuk terutama membahas cara – cara penggunaan Data Science secara umum dalam penanganan maupun resolusi perselisihan/sengketa sedangkan penelitian yang direncanakan secara spesifik akan meneliti penggunaan algoritma klasifikasi teks dalam teks mining (sebagai bagian dari data science) pada dokumen hasil mediasi perkara

2.3. Landasan Teori

2.3.1 Mediasi

Mediasi merupakan salah satu metode dalam rangka penyelesaian sengketa di pengadilan. Dalam Pasal 1 angka 1 Peraturan Mahkamah Agung Nomor 1 Tahun 2016 tentang Prosedur Mediasi di Pengadilan dinyatakan bahwa mediasi adalah cara penyelesaian sengketa melalui proses perundingan untuk memperoleh kesepakatan para pihak dengan dibantu oleh mediator. Ketua Pengadilan Negeri wajib mengusahakan perdamaian bagi kedua belah pihak. Upaya perdamaian di persidangan merupakan hal yang wajib dilakukan oleh Hakim sebagaimana diatur dalam Pasal 131 HIR dan jika Hakim tidak berhasil mendamaikan, maka harus disebutkan dalam Berita Acara Persidangan. Untuk saat ini, pengaturan teknis dari Pasal 130 HIR/154 RBG diatur dalam Perma No 1 Tahun 2016 tentang Prosedur Mediasi di Pengadilan. Perdamaian itu sendiri pada dasarnya harus mengakhiri perkara, harus dinyatakan dalam bentuk tertulis, perdamaian harus dilakukan oleh seluruh pihak yang terlibat dalam perkara dan oleh orang yang mempunyai kuasa untuk itu, dan ditetapkan dengan akta perdamaian. Hasil Mediasi dapat berupa *Mediasi Berhasil dengan Kesepakatan Perdamaian, Mediasi Berhasil dengan Pencabutan Gugatan, Mediasi Berhasil Sebagian, Mediasi Tidak Berhasil dan Mediasi Tidak Dapat dilaksanakan.*

Suatu mediasi dapat dikatakan berhasil apabila antara dua pihak mencapai kesepakatan untuk berdamai atau sepakat mencabut gugatan. Mediasi dikatakan gagal apabila salah satu atau kedua belah pihak tidak beritikad baik (misalnya dengan tidak hadir pada waktu mediasi dilaksanakan) ataupun karena para pihak

tidak mau berdamai sehingga perkara harus dilanjutkan ke tahap pemeriksaan atau persidangan.

2.3.2 Text Mining

Text mining adalah salah satu area dalam ilmu komputer yang mengkombinasikan data mining, machine learning, NLP (Natural Language Processing), information retrieval dan manajemen knowledge. Preprocessing pada text mining mengikuti tahapan yang kurang lebih sama dengan data preprocessing yaitu data preparation dan data reduction. Data preparation mencakup data integration, data cleaning, data normalization, data transformation, missing data imputation dan noise identification. Data reduction mencakup feature selection, Instance selection, Discretization dan Feature Extraction (Garcia et al., 2015).

Dalam text mining, text preprocessing berarti mengubah dokumen ke dalam format yang mudah dimengerti, diprediksi dan dapat dianalisis melalui berbagai algoritma machine learning. Terdapat beberapa teknik pre-processing yang sering digunakan diantaranya segmentasi kalimat, *lower/upper case conversion, Tokenizing, POS (Part-Of-Speech) Tagging, Stopwords Removal, Punctuation Removal, Stemming dan Lemmatization* (Tabassum & Patil, 2020) dimana setiap tipe memiliki fungsi dan tujuannya masing - masing.

Segmentasi Kalimat disebut juga Deteksi Batas Kalimat, mengacu pada proses pemecahan Teks dokumen atau korpus menjadi kalimat tersendiri yang membantu dalam mengidentifikasi batasan kata sehingga dapat diproses lebih lanjut dilakukan pada setiap kalimat. Segmentasi dilakukan di terjadinya tanda titik atau tanda baca dalam kalimat tokenizer.

Lower/Upper Case Conversion adalah tipe preprocessing dimana setiap huruf dalam kalimat diubah menjadi huruf kecil atau huruf besar. Tahapan ini sering pula disebut Case Folding.

Tokenization adalah proses memecah suatu materi besar menjadi potongan-potongan kecil, seperti frasa dan kata. Token adalah komponen terkecil. Token dalam sebuah kalimat, misalnya adalah sebuah kata, sedangkan kalimat adalah sebuah token dalam sebuah paragraf. Kriteria pemisahan terutama pada kemunculan spasi atau tanda baca. Langkah ini membantu menyaring kata-kata yang tidak diinginkan dalam langkah pemrosesan lebih lanjut. Contoh tokenisasi pada kalimat adalah sebagai berikut:

Contoh kalimat:

> "Mediasi dilakukan atas dasar kesepakatan oleh para pihak, yang mengajukan permohonan pemeriksaan atas sengketa yang dialami."

Kalimat ini dapat dipecah menjadi:

> "Mediasi", "dilakukan", "atas", "dasar", "kesepakatan", "oleh", "para", "pihak", ",", "yang", "mengajukan", "permohonan", "pemeriksaan", "atas", "sengketa", "yang", "dialami".

Selanjutnya adalah *Part-of-Speech* (POS) *Tagging* yaitu cara untuk menetapkan jenis kata pada setiap kata (seperti kata benda, kata kerja, kata sifat, dan lain-lain) dalam teks tertentu berdasarkan makna dan konteksnya. Semakin baik segmentasi kalimatnya yang dilakukan, semakin baik bagi POS tagger untuk mengidentifikasi bagian-bagian dalam speech.

Stopwords Removal adalah bagian dari tahapan preprocessing teks yang bertujuan untuk menghapus kata yang tidak relevan didalam suatu kalimat berdasarkan daftar stopword. *Stopword* adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Dalam teks preprocessing, stopwords removal sering disebut juga filtering dimana kata – kata penting diambil dari hasil tokenizing dengan membuang kata yang tidak atau kurang penting dan menyimpan kata yang penting. Untuk stopwords removal pada teks berbahasa Indonesia dapat digunakan library python bernama sastrawi (Rosid et al., 2020). Contoh *stopword* dalam bahasa Indonesia yang berada dalam kamus kata sastrawi adalah "yang", "dan", "di", "dari", dll.

Punctuation Removal adalah penghapusan tanda baca seperti titik, koma, tanda seru dan sebagainya. Tanda baca ini seringkali dianggap sebagai noise oleh mesin khususnya dalam sebuah dokumen yang tidak terstruktur. Dalam removal, dikenal adanya *specialized removal* yaitu penghapusan specialized character, misalnya tag HTML.

Stemming adalah proses dimana dilakukan pemotongan pada akhir kata atau imbuhan yang bertujuan untuk memperoleh akar kata. Contohnya adalah "memperlihatkan", "terlihat", "dilihat" akar katanya adalah "lihat". Namun demikian, setiap Bahasa memiliki perbedaan dalam proses stemming yang disebabkan oleh struktur kata. Contohnya pada teks berbahasa Inggris, yang perlu dihilangkan hanyalah sufiks. Lain halnya dengan teks berbahasa Indonesia, semua kata imbuhan baik sufiks maupun prefix dihilangkan/dipotong dalam proses stemming.

Lemmatization adalah proses menghilangkan atau mengganti sufiks dari kata untuk membawanya ke basisnya disebut lemma. Lemma adalah kata yang bermakna, berbeda dengan kata dasar. Terdapat perbedaan antara stemming dengan lemmatization. Stemming melibatkan penghilangan sufiks dari kata-kata untuk mendapatkan bentuk dasarnya, sedangkan lemmatization melibatkan pengubahan kata menjadi bentuk dasar morfologisnya.

Dalam pemrosesan teks juga dikenal adanya proses ekstraksi teks yaitu proses untuk mengidentifikasi maupun mengambil informasi spesifik atau struktur dari teks (Khurana et al., 2023). Ekstraksi teks melibatkan penarikan entitas, frase, atau data penting dalam dokumen teks dan bertujuan untuk mengubah teks yang tidak terstruktur menjadi lebih terstruktur agar dapat diperoleh informasi yang relevan dari korpus teks yang besar dan kompleks.

Terdapat beberapa metode yang sering digunakan dalam ekstraksi teks, diantaranya metode VSM (Vector Space Model) dengan "bag of words" (BoW), N-gram (Georgieva-Trifonova & Duraku, 2021), Named Entity Recognition (NER), TF-IDF (Term Frequency-Inverse Document Frequency), (Tabassum & Patil, 2020), Keyword Extraction (Locke & Zuccon, 2022) dan lain sebagainya.

2.3.3 N-gram

Referensi pertama N-gram berasal dari makalah Claude Shannon "A Mathematical Theory of Communications" yang diterbitkan pada tahun 1948. N-gram merupakan salah satu metode dalam *feature extraction atau feature engineering*. Sebuah fitur yang diekstrak dari teks dapat menentukan karakteristik tugas NLP dan menentukan bagaimana algoritma machine learning menafsirkan

data teks tersebut. N-gram dapat digunakan sebagai model probabilistik untuk memprediksi item berikutnya dalam urutan item. Item bisa berupa huruf / karakter, kata, atau yang lain. Salah satunya, model n-gram yang berbasis kata digunakan untuk memprediksi kata berikutnya dalam urutan kata tertentu.

Dalam hal item atau term yang digunakan adalah kata, maka N-gram dapat dihitung dengan menggunakan persamaan 1 di bawah ini:

$$Ngram_k = X - (N - 1) \quad (1)$$

Dimana $Ngram_k$ adalah banyaknya n-gram dalam k kalimat , X adalah jumlah kata dalam kalimat k, N adalah tipe n-gram yang digunakan, bisa berupa unigram, bigram, trigram dan seterusnya. N-gram adalah barisan token berurutan dengan panjang (terbatas) n; satu token adalah 1 gram, dua berturut-turut tokennya disebut 2-gram dan seterusnya (Ferrario & Naegelin, 2020).

Dengan kelebihan – kelebihan yang dimilikinya dalam menangkap konteks dan struktur bahasa tersebut, penggunaan n-gram dapat dipertimbangkan apabila tujuan analisis adalah untuk melakukan ekstraksi informasi dari teks.

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

TF-IDF adalah salah satu metode pembobotan kata untuk menentukan frekuensi relatif kata dalam sebuah dokumen tertentu (Term Frequency) dan frekuensi kemunculan kata tersebut di seluruh korpus dokumen (Inverse Document Frequency). Hasil TF-IDF berupa matriks yang diperoleh dari perkalian IDF dengan TF (Nurhidayat & Dewi, 2023). TF-IDF dihitung dengan persamaan 2 berikut ini:

$$TF * IDF\ (d,t) = TF(d,t) * log\frac{N}{df(t)} \quad (2)$$

Dimana $TF * IDF$ *(d,t)* adalah Bobot TF-IDF, *TF(d,t)* adalah Jumlah munculnya term t pada dokumen d, *N* adalah Total dokumen (korpus) dan *df(t)* adalah Jumlah dokumen yang di dalamnya mengandung term t.

2.3.5 Algoritma Klasifikasi - Prediksi

Prediksi adalah salah satu kelompok Data mining yang hampir sama dengan klasifikasi dan estimasi. Perbedaaanya adalah bahwa prediksi nilai dari hasil akan berada di masa mendatang. Klasifikasi digunakan apabila terdapat target variabel kategori (Larose, 2005) (Kusrini & Luthfi, 2009).

Dalam penelitian ini, akan digunakan Algoritma Prediksi Klasifikasi berupa:

i. Naïve Bayes, yaitu salah satu algoritma yang digunakan untuk melakukan prediksi pada kemungkinan keanggotaan suatu class (Kusrini & Luthfi, 2009). Bentuk Umum dari Teorema Bayes adalah persamaan 3 di bawah ini:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

Dimana, X adalah data dengan class yang belum diketahui, H adalah Hipotesis data X, merupakan suatu class spesifik, P(H|X) adalah probabilitas hipotesis H berdasar kondisi X (posteriori probability), P(H) adalah probabilitas hipotesis H (prior probability), P(X|H) adalah probabilitas X berdasar kondisi pada hipotesis H, dan P(X) adalah probabilitas dari X.

ii. Logistic Regression, merupakan algoritma Klasifikasi yang menggunakan fungsi logistic untuk memprediksi probabilitas kelas target berdasarkan kombinasi linear dari variabel – variabel independen (x) (Sengupta & Dave,

2022). Fungsi sigmoid digunakan untuk memodelkan probabilitas sebagaimana persamaan 4 berikut ini:

$$P(y = 1 \mid X) = 1 / (1 + \exp(-z)) \quad (4)$$

Dimana P (y=1 | X) adalah probabilitas dari variabel target pada adalah probabilitas variabel target menjadi 1 dengan prediktor X yang diberikan, z adalah kombinasi linear dari predictor dan koefisiennya z = w0 + w1*X1 + w2*X2 + ... + wn*Xn dan exp () adalah fungsi eksponensial.

iii. Decision tree. Penggunaan Decision tree adalah dengan classifier yang menunjukkan keberadaan kata dalam kategorisasi teks dan percabangan keputusan dari suatu node didasarkan pada respon fitur. Decision tree memadukan eksplorasi data dan pemodelan dimana kumpulan data teks yang heterogen dibagi – bagi menjadi kalimat kemudian menjadi kata atau token yang lebih homogen dengan memperhatikan variable targetnya. Dengan kata lain, terdapat proses untuk mengubah data dari tabel menjadi berbentuk pohon, kemudian diubah menjadi rule dan langkah terakhir adalah menyederhanakan rule. Dalam klasifikasi teks, terdapat langkah – langkah yang harus dilakukan yaitu mengumpulkan nilai frekuensi token muncul dalam sebuah kalimat atau dokumen dan menghitung entropi (ketidakpastian dan keacakan dataset) awal dengan rumus:

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i * \log_2 p_i \quad (5)$$

Dimana S adalah himpunan kasus, n adalah jumlah partisi S dan p_i adalah proporsi S_i terhadap S. Nilai Entropi selalu berada dalam rentang 0 dan 1(Charbuty & Abdulazeez, 2021). Langkah selanjutnya adalah menghitung gain, menggunakan rumus seperti yang tertera dalam persamaan 6 berikut.

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (6)$$

Dimana S adalah himpunan kasus, A adalah atribut, n adalah jumlah partisi atribut A, $|S_i|$ adalah jumlah kasus pada partisi ke-i dan $|S|$ adalah jumlah kasus dalam S.

iv. Support Vector Machine (SVM) bertujuan untuk menemukan hyperplane yang paling baik memisahkan kelas-kelas dalam ruang fitur. Fungsi keputusan untuk memprediksi label kelas dari titik data baru. SVM dapat dihitung menggunakan persamaan 5 sebagai berikut:

$$f(x) = sign(w \cdot x + b) \quad (7)$$

Dimana x adalah titik data input, w adalah vektor bobot, b adalah bias term, · menunjukkan perkalian titik, dan sign () mengembalikan tanda dari nilai fungsi.

2.3.6 Evaluasi Model

Penelitian ini adalah tentang menentukan mediasi berhasil atau gagal, yang berarti klasifikasi biner. Untuk itu akan digunakan metric performa untuk mengevaluasi model yang dibuat dengan menghitung nilai akurasi (Persamaan 8), presisi (Persamaan 9), recall (Persamaan 10) dan F1-score (Persamaan 11) menggunakan rumus sebagai berikut:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1-score} = \frac{2 * presisi * recall}{presisi + recall} \quad (11)$$

Hasil perhitungan performa algoritma kemudian dibandingkan satu sama lain dalam suatu komparasi untuk menunjukkan algoritma mana memiliki performa paling baik.

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini adalah penelitian eksperimental dan menggunakan data primer berupa dokumen hasil mediasi berbahasa Indonesia. Dataset yang digunakan diperoleh dari Pengadilan Negeri di Indonesia.

Penelitian ini bersifat deskriptif dikarenakan menggambarkan objek yang akan diteliti dan menjabarkan hasil pengujian dari model yang dihasilkan dari dataset yang ada sehingga diketahui tingkat akurasi dan kesalahan (error) yang mungkin terjadi.

Penelitian ini menggunakan pendekatan penelitian kuantittatif karena penelitian ini dilakukan secara sistematis, terencana dan terstruktur. Pada bagian akhir dari penelitian akan dilakukan pengujian performa model dengan menggunakan formula akurasi, presisi, recall dan F1-score.

3.2. Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah hasil Mediasi perkara perdata di Pengadilan Negeri dalam wilayah hukum Pengadilan Tinggi Sulawesi Utara sebanyak 8 (delapan) Pengadilan Negeri yaitu Pengadilan Negeri Manado, Pengadilan Negeri Tondano, Pengadilan Negeri Bitung, Pengadilan Negeri Kotamobagu, Pengadilan Negeri Tahuna, Pengadilan Negeri Airmadidi, Pengadilan Negeri Amurang dan Pengadilan Negeri Melonguane.

Teknik pengumpulan data untuk penelitian ini adalah Observasi langsung di pengadilan negeri dan studi dokumen (sampel dokumen elektronik). Observasi dilakukan dengan mengamati dan mencatat semua hal terkait proses pelaksanaan mediasi dan dokumen – dokumen yang digunakan dan dihasilkan dalam mediasi di pengadilan negeri. Studi dokumen dilakukan terhadap dokumen hasil mediasi terdahulu, baik mediasi gagal maupun mediasi berhasil dalam rentang waktu tahun 2020 sampai dengan tahun 2023 yang diekspor dari database sistem informasi perkara dalam bentuk file .csv. Pengumpulan data dari 8 (delapan) pengadilan negeri menghasilkan sebanyak 2591 data awal sebagaimana Tabel 1.1. pada Bab Pendahuluan.

3.3. Metode Analisis Data

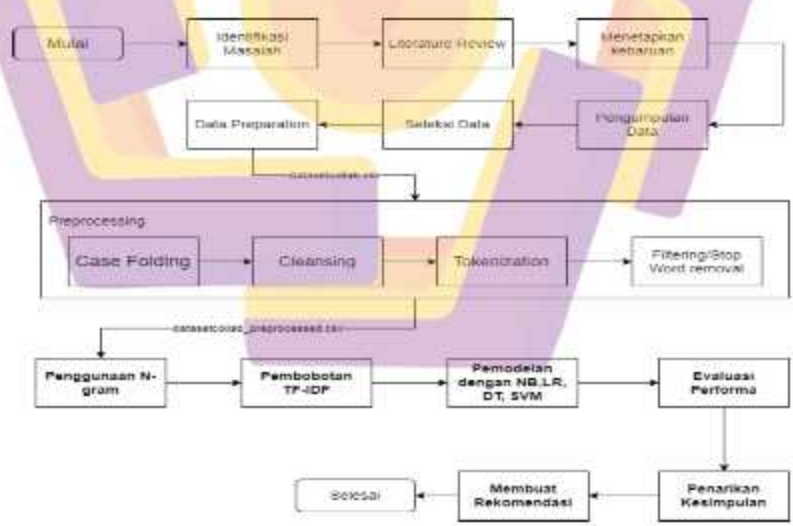
Proses analisis data dilakukan melalui beberapa tahapan, yaitu Seleksi Data, Data Preparation, Preprocessing, Implementasi N-gram, Pembobotan TF-IDF, Pemodelan, dan Evaluasi Model. Seleksi data dilakukan pada 2591 data yang telah dikumpulkan dan mengubah dataset dengan menyisakan kolom nomor perkara atau nomor penetapan mediasi, kolom petitum dan kolom status hasil mediasi sebagai variabel yang akan dianalisis. Variabel sumber adalah kolom petitum dan variabel target adalah kolom status hasil mediasi sedangkan kolom nomor perkara atau nomor penetapan mediasi berfungsi sebagai identitas per baris dataset.

Data Preparation dilakukan dengan menentukan label status hasil mediasi yang akan menjadi target prediksi yaitu mediasi berhasil atau mediasi gagal. Preprocessing adalah tahapan untuk membersihkan dataset dengan proses Case Folding, Cleansing, Tokenization dan Stop Word Removal atau Filtering.

Implementasi N-gram adalah proses ekstraksi fitur untuk menghasilkan term dari dataset yang telah mengalami preprocessing. Pembobotan TF-IDF adalah proses penghitungan bobot dari setiap term yang dihasilkan sebagai output dari implementasi N-gram. Pemodelan menggunakan nilai TF-IDF yang dimodelkan dengan algoritma klasifikasi teks yaitu Naïve Bayes, Logistic Regression, Decision tree dan Support Vector Machine. Model – model yang telah dibuat kemudian dievaluasi kinerjanya dengan menggunakan perhitungan akurasi, presisi, recall dan F1-score.

3.4. Alur Penelitian

Alur penelitian yang diaplikasikan dalam penelitian ditunjukkan pada gambar 3.1 di bawah ini.



Gambar 3.1. Alur penelitian

Alur penelitian dalam Gambar 3.1 di atas dapat dijelaskan sebagai berikut:

Identifikasi masalah dilakukan dengan menentukan rumusan – rumusan masalah berdasarkan latar belakang masalah yang diperoleh pada proses mediasi. Permasalahan yang diangkat kemudian dibuat batasan – batasannya sehingga memiliki pembahasan yang relevan serta membatasi ruang lingkup penelitian yang luas. Identifikasi masalah didukung pula dengan uraian singkat dari beberapa penelitian yang terkait dengan implementasi algoritma klasifikasi teks dan n-gram yang mendasari perlunya diadakan penelitian ini.

Literature review adalah tahapan dimana dilakukan tinjauan terhadap referensi – referensi penelitian baik berupa jurnal atau prosiding, buku, maupun Peraturan perundang – undangan yang mendukung penelitian ini. Literature Review dilakukan untuk mendapatkan gambaran terhadap penelitian – penelitian terdahulu maupun teori – teori terkait penelitian. Dari literature review diperoleh pula perbandingan apa yang telah dilakukan oleh peneliti terdahulu dan apa yang akan dilakukan dalam peneltian ini.

Dari proses identifikasi masalah dan literature review, dapat ditetapkan kebaruan/novelty yang akan diangkat dan apa kontribusi yang dapat disumbangkan melalui penelitian ini.

Pengumpulan Data dilakukan saat sudah ditetapkan topik penelitian yaitu pemanfaatan dokumen mediasi terdahulu. Data dikumpulkan melalui observasi dan studi dokumen dari beberapa pengadilan negeri, sebagaimana telah dijelaskan pada bagian 3.2 bab ini. Data yang dikumpulkan adalah sebanyak 2591 data mentah.

Seleksi Data kemudian dilakukan pada data yang telah dikumpulkan tersebut. Dengan menganggap kolom 'nomor perkara atau nomor penetapan

mediasi' adalah identitas bagi setiap baris data yang membedakan dengan data yang lain, maka langkah pertama adalah memfilter seluruh data yang memiliki isi (tidak NULL) pada kolom 'nomor perkara atau nomor penetapan mediasi'. Selanjutnya, dihitung berapa banyak data yang memiliki label pada kolom 'status hasil mediasi' yaitu keterangan status hasil mediasi tidak berhasil (T), tidak dilaksanakan (D), Berhasil dengan Kesepakatan Damai (Y1), Berhasil dengan Pencabutan (Y2) dan Berhasil Sebagian (S). Dihitung pula berapa banyak data yang tidak memiliki label (NULL).

Tahapan selanjutnya adalah Data Preparation, dimana pada tahapan ini, data dengan label Y1 dan label Y2 diubah labelnya menjadi Y karena keduanya dianggap sebagai kondisi mediasi berhasil. Selanjutnya dilakukan penghapusan data dengan label hasil mediasi D, karena mediasi yang tidak dilaksanakan adalah sepenuhnya kesadaran dari pihak berperkara namun tidak serta merta menjadikan mediasi tersebut dianggap tidak berhasil. Kemudian, dilanjutkan dengan penghapusan terhadap data dengan label hasil mediasi S atau mediasi berhasil sebagian. Data ini dihapus karena berdasarkan pertimbangan mediator merekomendasikan agar perkara diperiksa dan disidangkan untuk poin – poin tertentu dalam posita maupun petirum.

Data hasil preparation menjadi dataset yang akan digunakan dalam penelitian variabel adalah:

- Nomor perkara atau nomor penetapan, adalah identitas unik dari setiap mediasi, yang berbeda satu sama lain
- petitum adalah isi gugatan penggugat kepada tergugat, sebagai variabel sumber

- status hasil mediasi adalah label yang diberikan terhadap hasil mediasi, yaitu mediasi berhasil (Y) atau mediasi gagal (T), sebagai variabel target

Dataset kemudian diproses dalam tahapan Data Pre-processing agar lebih mudah digunakan untuk proses klasifikasi, lebih seragam dan untuk menghilangkan noise. Teknik preprocessing yang akan digunakan dalam penelitian ini adalah: Proses *Case Folding* yaitu tahapan mengubah huruf capital atau huruf besar menjadi huruf kecil. Cleansing yaitu tahapan untuk membersihkan data seperti menghilangkan angka, tanda baca maupun symbol pada dataset. Dalam penelitian ini, akan dilakukan HTML removal yaitu menghilangkan tag HTML yang ada pada dataset, pembersihan tanda baca, symbol – symbol serta penghapusan baris data yang tidak. Proses selanjutnya adalah *Tokenization* terhadap teks dari dokumen yang telah dibersihkan dari tag HTML. Teks dalam dokumen akan dipecah menjadi bagian – bagian yang lebih kecil. Dilanjutkan dengan proses *Filtering* atau *Stop Word Removal* dilakukan terhadap token yang dihasilkan pada proses sebelumnya. Proses ini akan menghilangkan *stop word* seperti "yang", "di", "dari", "oleh". Library yang akan digunakan adalah *Sastrawi* karena library ini secara spesifik menyediakan daftar stop word dalam Bahasa Indonesia.

Tahapan berikutnya adalah menggunakan n-gram dalam proses *feature extraction*. Dalam tahapan ini, kata – kata yang telah melalui tahapan pre-processing akan dihitung frekuensinya muncul dalam dokumen (n). Apabila muncul > 1 kali maka akan dicatat penambahannya. Perhitungan n-gram akan menggunakan Persamaan 1. Penggunaan n-gram adalah dengan menggunakan python dengan n = 1 (unigram) sampai dengan n = 5 (poligram). Keluaran n-gram

berupa term. Selanjutnya, dilakukan proses Pembobotan dengan TF – IDF (Term Frequency – Inverse Document Frequency). Pada tahapan ini, term keluaran n-gram menjadi nilai t pada persamaan 2.

Tahapan selanjutnya adalah pembuatan model dengan menggunakan algoritma Naïve Bayes, Logistic Regression, Decision tree dan SVM (Support Vector machine). Model kemuadian akan diimplementasikan untuk mengecek berapa banyak akar kata yang muncul dokumen yang telah di-grouping pada tahapan preprocessing, selanjutnya membuat rekomendasi apakah mediasi berhasil atau gagal.

Untuk mengevaluasi akurasi dan performa model, dilakukan perhitungan nilai akurasi, presisi, recall dan F1-score menggunakan persamaan 6, 7, 8 dan 9.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Penelitian ini melalui tahapan Data Selection, Data preparation, Data Preprocessing, Pemisahan Data Latih dan Data Uji, Implementasi N-gram, Implementasi TF-IDF, Pembuatan Model dengan Algoritma Klasifikasi dan tahapan Evaluasi performa model yang dihasilkan. Secara rinci, tahapan – tahapan tersebut dijelaskan dalam sub bab – sub bab di bawah ini:

4.1. Data Selection

Data awal yang dikumpulkan dari 8 (delapan) pengadilan negeri adalah sebanyak 2591 data dalam file hasil ekspor dari sistem informasi bernama datasetcollab_awal.csv sebagaimana contoh pada Tabel 4.1 di bawah ini. Dataset berbentuk tabel yang terdiri dari 6 (enam) kolom atau field yaitu [nomor perkara atau nomor penetapan mediasi] yang berformat varchar (50), [klasifikasi perkara] yang berformat varchar (100), [mulai mediasi] yang berformat date, [selesai mediasi] yang berformat date, [petitum] yang berformat text dan [status hasil mediasi] yang berformat char (2). Dataset tersebut pada kolom petitum masih mengandung tag – tag html karena untuk beberapa field dalam sistem informasi perkara memang dibuat untuk mendukung rich text dari hasil upload dokumen agar format teks tetap terjaga.

Tabel 4.1. Contoh Dataset awal

nomor perkara atau nomor penetapan mediasi	klasifikasi perkara	mulai mediasi	selesai mediasi	petitum	status hasil mediasi
9/Pdt.G/2020/PN Mgn	Perceraian	1/30/2020	1/30/2020	- Menerima dan mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan menurut hukum Perkawinan antara Penggugat dan Tergugat yang menikah secara agama Kristen Protestan pada tanggal 15 Desember 2016 di Melonguane sebagaimana tercatat menurut Akta Perkawinan Nomor 7104/CPK/15122016.00057, putus karena Perceraian; - Menetapkan anak Penggugat dan Tergugat yaitu : - ADRIVAN MANGALO yang lahir di Kalongan tanggal 10 Oktober 2014 jenis kelamin Laki-laki sebagaimana tercatat dalam Akta Kelahiran Nomor 7104-LT-27012017-0001; ……dst	T
16/Pdt.G/2021/PN Mgn	Perceraian	3/12/2021	3/29/2021	- Mengabulkan gugatan Penggugat untuk seluruhnya ; - Menyatakan dalam hukum bahwa perkawinan Penggugat dan Tergugat yang telah dilangsungkan pada tanggal 25 Juni Tahun 2012 di GEREJA MASEHI INJILI TALAUD (GERMITA) Jemaat Gereja Paradise Melonguane, dan telah pula didaftarkan dan dicatatkan pada Kantor Dinas Kependudukan dan Pencatatan Sipil Kabupaten Kepulauan Talaud, dengan Kutipan Akta Perkawinan Nomor: 7104/CPK/25062012.0294 tanggal 25 Juni 2012 adalah putus karena perceraian dengan segala akibat hukumnya; - Menetapkan Penggugat dan juga Tergugat sebagai wali asuh anak-anak masih di bawah umur dari hasil perkawinan Penggugat dan juga Tergugat, yaitu bernama …dst	Y2

Tabel 4.1. Lanjutan

nomor perkara atau nomor penetapan mediasi	klasifikasi perkara	mulai mediasi	selesai mediasi	Petitum	status hasil mediasi
50/Pdt.G/2021/PN Mgn	Perbuatan Melawan Hukum	10/6/2021	10/18/2021	- Mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan Tergugat telah melakukan Perbuatan Melawan Hukum (PMH); - Menghukum Tergugat untuk membyara kerugian materil yakni akibat telah dilakukannya Penangkapan dan Penahanan kepada adik Penggugat sejumlah Rp. 50.000.000,-(lima puluh juta rupiah) atau suatu jumlah yang dipandang layak dan adil oleh Pengadilan cq Majelis Hakim, jumlah kerugian mana harus dibayarkan secara tunai, sekaligus dan seketika oleh Tergugat; - Menghukum Tergugat untuk membayar ganti kerugian immateril sebesar Rp. 100.000.000,-(seratus juta rupiah) atau suatu jumlah yang dipandang layak dan adil oleh Pengadilan cq majelis hakim, jumlah kerugian mana harus dibayarkan secara tunai, sekaligus dan seketika Tergugat; - Membebankan biaya perkara kepada para Tergugat; <u>Subsidair :</u> Jika Majelis Hakim berpendapat lain mohon putusan yang seadil-adilnya (ex aequo et bono);	Y1
62/Pdt.G/2021/PN Mgn	Perceraian	11/22/2021	12/20/2021	- Menerima dan mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan perkawinan Penggugat dan Tergugat yang yang di langsungkan di Niampak pada tanggal 24 Februari 2017, Perkawinan tersebut telah dicatatkan/didaftarkan di Kantor Dinas Kependudukan dan Catatan Sipil Kabupaten Talaud, sesuai dengan Kutipan Akta Perkawinan Nomor : 71.04/ CPK/ 24022017 yang dikeluarkan di Tarohan pada tanggal 28 Februari 2017 putus karena perceraian; - Menetapkan Penggugat dan tergugat sebagai Hak Asuh bersama terhadap seorang anak .....dst	D

Tabel 4.1. Lanjutan

nomor perkara atau nomor penetapan mediasi	klasifikasi perkara	mulai mediasi	selesai mediasi	Petitum	status hasil mediasi
55/Pdt.G/2023/PN Mgn	Perbuatan Melawan Hukum	8/16/2023	10/9/2023	Mengabulkan gugatan Penggugat untuk seluruhnya; Menyatakan sah menurut hukum Akta Jual Beli dengan No.30/JB/1988 tertanggal 14 April 1988 yang dibuat dihadapan Pejabat Pembuat Akta Tanah Camat Lirung Drs. Alfrits Areros dengan batas-batas sebagai berikut : - Utara : Sungai Sarosoga - Timur : Abdon Bulawan - Selatan : Silas garasut - Barat : Tepi Laut Menyatakan menurut hukum bahwa Tanah objek sengketa seluas ± 270 m2 (kurang lebih dua ratus tujuh puluh meter persegi) dengan batas batas sebagai berikut : - Utara : Kalpein Sarean - Timur : Djepri Garasut dan Alpres Doner Garasut atau Salah satunya - Selatan : Djepri Garasut dan Alpres Doner Garasut atau Salah satunya - Barat : Djepri Garasut dan Alpres Doner Garasut atau Salah satunya - adalah sah tanah milik Penggugat, karna Tanah Objek sengketa tersebut adalah termasuk kedalam keseluruhan bidang tanah seluas ± 13.498,5 m2 (kurang lebih tiga belas ribu empat ratus Sembilan puluh delapan koma lima meter persegi) berdasarkan Akta Jual Beli No.30/JB/1988 tertanggal 14 April 1988 yang dibuat dihadapan Pejabat Pembuat Akta Tanah Camat Lirung Drs. Alfrits Areros, serta lampiran Akta tentang Gambar Situasi Tanah Jual beli tanggal 26 Desember 1987 atas nama Kalpein Sarean; Menyatakan menurut hukum bahwa dalam proses Penerbitan Sertifikat Hak Milik (SHM) atas nama Tergugat I atau Tergugat II atau atas nama keduanya yang sebagian tanahnya ada di atas objek tanah sengketa tersebut adalah mengandung cacat hukum dan harus dinyatakan tidak sah dan batal demi hukum; Menyatakan menurut hukum, bahwa karena Tergugat I dan Tergugat II telah dengan cara-cara paksa masuk dengan pengancaman dan menguasai tanah objek sengketa seluas ± 270 m2 (kurang lebih dua ratus tujuh puluh meter persegi) padahal tanah objek ….dst	S

Seleksi data dilakukan terhadap 2591 data awal untuk memilah data mana yang dapat digunakan dan mana yang tidak. Data awal disimpan dalam file datasetcollab_awal.csv. Sesuai dengan tujuan penelitian yaitu melakukan prediksi, maka kolom petitum dan kolom status hasil mediasi adalah variabel yang akan dianalisis. Variabel sumber adalah kolom petitum dan variabel target adalah kolom status hasil mediasi. Terhadap dataset awal dilakukan penghapusan kolom – kolom yang tidak akan digunakan yaitu kolom klasifikasi perkara, kolom mulai mediasi dan kolom selesai mediasi sehingga tabel menjadi dua kolom saja sebagaimana Tabel 4.2 di bawah ini. Kolom nomor perkara atau nomor penetapan tidak dihapus karena digunakan untuk identitas perkara. Pada pemrosesan di tahapan – tahapan berikutnya, kolom nomor perkara digunakan untuk mengidentifikasi baris – baris hasil pemrosesan.

Tabel 4.2. Contoh Dataset setelah penghapusan kolom

nomor perkara atau nomor penetapan mediasi	petitum	status hasil mediasi
9/Pdt.G/2020/ PN Mgn	- Menerima dan mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan menurut hukum Perkawinan antara Penggugat dan Tergugat yang menikah secara agama Kristen Protestan pada tanggal 15 Desember 2016 di Melonguane sebagaimana tercatat menurut Akta Perkawinan Nomor 7104/CPK/15122016.00057, putus karena Perceraian; - Menetapkan anak Penggugat dan Tergugat yaitu : - ADRIVAN MANGALO yang lahir di Kalongan tanggal 10 Oktober 2014 jenis kelamin Laki-laki sebagaimana tercatat dalam Akta Kelahiran Nomor 7104-LT-27012017-0001; …….dst	T

Tabel 4.2. Lanjutan

nomor perkara atau nomor penetapan mediasi	petitum	status hasil mediasi
16/Pdt.G/2021 /PN Mgn	- Mengabulkan gugatan Penggugat untuk seluruhnya ; - Menyatakan dalam hukum bahwa perkawinan Penggugat dan Tergugat yang telah dilangsungkan pada tanggal 25 Juni Tahun 2012 di GEREJA MASEHI INJILI TALAUD (GERMITA) Jemaat Gereja Paradise Melonguane, dan telah pula didaftarkan dan dicatatkan pada Kantor Dinas Kependudukan dan Pencatatan Sipil Kabupaten Kepulauan Talaud, dengan Kutipan Akta Perkawinan Nomor: 7104/CPK/25062012.0294 tanggal 25 Juni 2012 adalah putus karena perceraian dengan segala akibat hukumnya; - Menetapkan Penggugat dan juga Tergugat sebagai wali asuh anak-anak masih di bawah umur dari hasil perkawinan Penggugat dan juga Tergugat, yaitu bernama ...dst	Y2
50/Pdt.G/2021 /PN Mgn	- Mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan Tergugat telah melakukan Perbuatan Melawan Hukum (PMH); - Menghukum Tergugat untuk membyura kerugian materil yakni akibat telah dilakukannya Penangkapan dan Penahanan kepada adik Penggugat sejumlah Rp. 50.000.000,- (lima puluh juta rupiah) atau suatu jumlah yang dipandang layak dan adil oleh Pengadilan cq Majelis Hakim, jumlah kerugian mana harus dibayarkan secara tunai, sekaligus dan seketika oleh Tergugat; - Menghukum Tergugat untuk membayar ganti kerugian immateril sebesar Rp. 100.000.000,-(seratus juta rupiah) atau suatu jumlah yang dipandang layak dan adil oleh Pengadilan cq majelis hakim, jumlah kerugian mana harus dibayarkan secara tunai, sekaligus dan seketika Tergugat; - Membebankan biaya perkara kepada para Tergugat; <u>Subsidair :</u> Jika Majelis Hakim berpendapat lain mohon putusan yang seadil-adilnya (ex aequo et bono);	Y1
62/Pdt.G/2021 /PN Mgn	- Menerima dan mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan perkawinan Penggugat dan Tergugat yang yang di langsungkan di Niampak pada tanggal 24 Februari 2017, Perkawinan tersebut telah dicatatkan/didaftarkan di Kantor Dinas Kependudukan dan Catatan Sipil Kabupaten Talaud, sesuai dengan Kutipan Akta Perkawinan Nomor : 71.04/ CPK/ 24022017 yang dikeluarkan di Tarohan pada tanggal 28 Februari 2017 putus karena perceraian; - Menetapkan Penggugat dan tergugat sebagai Hak Asuh bersama terhadap seorang anak .....dst	D

Dataset hasil penghapusan kolom disimpan dalam file 1_filtered_dataset.csv. Setiap baris data memiliki kolom status hasil mediasi sebagai labelnya yaitu label Y1 artinya mediasi "Berhasil Dengan Kesepakatan",

label Y2 artinya mediasi "Berhasil Dengan Pencabutan", label S artinya mediasi "Berhasil Sebagian", label T artinya mediasi "Tidak Berhasil" dan label D artinya mediasi "Tidak dapat dilaksanakan". Selain label di atas, sebagian dataset tidak memiliki isi pada kolom status hasil mediasi atau NULL. Sebagian lagi berisi data namun tertulis "(NULL)" sehingga diasumsikan data tersebut juga adalah NULL.

Dengan menggunakan python dihitung jumlah data untuk masing – masing label dengan hasil berikut ini:

```
Jumlah data dengan status hasil mediasi 'Tidak Berhasil (T)': 2397
Jumlah data dengan status hasil mediasi 'Tidak Dilaksanakan (D)':
57
Jumlah data dengan status hasil mediasi 'Berhasil dengan Kesepakatan
Damai (Y1)': 80
Jumlah data dengan status hasil mediasi 'Berhasil dengan Pencabutan
(Y2)': 28
Jumlah data dengan status hasil mediasi 'Berhasil Sebagian (S)': 4
Jumlah data yang tidak memiliki label (NULL): 17
Jumlah data yang memiliki label '(NULL)': 8
```

Hasil di atas menunjukkan pada dataset awal terdapat baris data status hasil mediasi dengan label T atau mediasi "Tidak Berhasil" sebanyak 2397 perkara, status hasil mediasi dengan label D atau mediasi "Tidak Dilaksanakan" sebanyak 57 perkara, status hasil mediasi dengan label Y1 atau mediasi "Berhasil dengan kesepakatan Damai" sebanyak 80 perkara, status hasil mediasi dengan label Y2 atau mediasi "Berhasil dengan Pencabutan" sebanyak 28 perkara, dan status hasil mediasi dengan label S atau mediasi "Berhasil sebagian" sebanyak 4 perkara. Selain itu, terdapat perkara yang tidak memiliki isi pada kolom status hasil mediasi atau NULL sebanyak 17 perkara dan sebagian lagi berisi data namun tertulis "(NULL)" sehingga diasumsikan data tersebut juga masuk sebagai kategori NULL sebanyak 8 perkara. Sehingga, rekapitulasi jumlah baris per kategori label status hasil mediasi dapat disajikan dalam tabel 4.3 berikut ini:

Tabel 4.3. Jumlah perkara per kategori label

Label	Jumlah Baris
T	2397
Y1	80
Y2	28
D	57
S	4
NULL	17
(NULL)	8
Jumlah dataset	2591

4.2. Data Preparation

Kolom status hasil mediasi pada dataset adalah variabel target yang secara keseluruhan terdiri dari label Y1 yang berarti "Berhasil Dengan Kesepakatan", label Y2 yang berarti "Berhasil Dengan Pencabutan", label S yang berarti "Berhasil Sebagian", label T yang berarti "Tidak Berhasil" dan label D yang berarti "Tidak dapat dilaksanakan".

Dalam tahapan ini, data dipersiapkan menjadi dataset yang hanya memiliki 2 (dua) status hasil mediasi yaitu Y = mediasi berhasil dan T = mediasi gagal. Untuk maksud tersebut, semua baris berlabel "Y1" dan "Y2" diubah menjadi semuanya berlabel "Y" atau sama – sama berisi data mediasi berhasil. Baris dengan label "D" dihapus karena mediasi yang tidak dilaksanakan bukanlah tujuan pemrosesan data. Baris dengan label "S" juga dihapus, karena walaupun mediasi berhasil sebagian, namun beberapa hal dalam pokok perkara masih harus diperiksa dalam persidangan. Baris dengan label "(NULL)" dan yang tidak memiliki labelpun (kosong atau NULL) akan dihapus. Penghapusan ini dilakukan karena label – label tersebut tidak menjadi tujuan pemrosesan data.

Untuk contoh dataset pada Tabel 4.2 sebelumnya, dengan perubahan label akan menjadi dua label saja yaitu "T" atau mediasi gagal dan "Y" atau mediasi berhasil, sebagaimana disajikan dalam tabel 4.4 berikut ini.

Tabel 4.4. Contoh penggabungan dan penghapusan label

petitum	status hasil mediasi
- Menerima dan mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan menurut hukum Perkawinan antara Penggugat dan Tergugat yang menikah secara agama Kristen Protestan pada tanggal 15 Desember 2016 di Melonguane sebagaimana tercatat menurut Akta Perkawinan Nomor 7104/CPK/15122016.00057, putus karena Perceraian; - Menetapkan anak Penggugat dan Tergugat yaitu : …….. tanggal 10 Oktober 2014 jenis kelamin Laki-laki sebagaimana tercatat dalam Akta Kelahiran Nomor 7104-LT-27012017-0001; …….dst</p>	T
- Mengabulkan gugatan Penggugat untuk seluruhnya ; - Menyatakan dalam hukum bahwa perkawinan Penggugat dan Tergugat yang telah dilangsungkan pada tanggal 25 Juni Tahun 2012 di GEREJA MASEHI INJILI TALAUD (GERMITA) Jemaat Gereja Paradise Melonguane, dan telah pula didaftarkan dan dicatatkan pada Kantor Dinas Kependudukan dan Pencatatan Sipil Kabupaten Kepulauan Talaud, dengan Kutipan Akta Perkawinan Nomor: 7104/CPK/25062012.0294 tanggal 25 Juni 2012 adalah putus karena perceraian dengan segala akibat hukumnya; - Menetapkan Penggugat dan juga Tergugat sebagai wali asuh anak-anak masih di bawah umur dari hasil perkawinan Penggugat dan juga Tergugat, yaitu bernama …dst	Y
- Mengabulkan gugatan Penggugat untuk seluruhnya; - Menyatakan Tergugat telah melakukan Perbuatan Melawan Hukum (PMH); - Menghukum Tergugat untuk membyara kerugian materil yakni akibat telah dilakukannya Penangkapan dan Penahanan kepada adik Penggugat sejumlah Rp. 50.000.000,-(lima puluh juta rupiah) atau suatu jumlah yang dipandang layak dan adil oleh Pengadilan cq Majelis Hakim, jumlah kerugian mana harus dibayarkan secara tunai, sekaligus dan seketika oleh Tergugat; - Menghukum Tergugat untuk membayar ganti kerugian immateril sebesar dst </p>	Y

Proses penghapusan baris data untuk label D, S dan NULL dilakukan dengan memastikan kode program tidak menyebabkan anomaly pada baris lainnya dalam dataset misalnya terhapusnya spasi dalam kolom petitum sehingga menjadi kata bersambung ataupun perubahan lain pada kolom petitum misalnya munculnya karakter – karakter atau simbol - simbol. Selain penghapusan baris data untuk label yang tidak diperlukan juga dilakukan penghapusan baris data yang tidak memiliki

isi (NULL) pada kolom petitum yang dapat menjadi noise apabila diproses lebih lanjut. Potongan kode program berikut ini menggambarkan proses dimaksud:

```
# Ubah semua baris berlabel "Y1" dan "Y2" menjadi "Y" hanya pada
kolom 'status hasil mediasi'
df['status hasil mediasi'] = df['status hasil
mediasi'].replace({'Y1': 'Y', 'Y2': 'Y'})

# Hapus baris dengan label "D" dan "S" hanya pada kolom 'status hasil
mediasi'
df = df[df['status hasil mediasi'].isin(['Y']) | ~df['status hasil
mediasi'].isin(['D', 'S'])]

# Hapus baris yang mengandung "(NULL)" hanya pada kolom 'status hasil
mediasi'
df = df[~df['status hasil mediasi'].str.contains(r'\(NULL\)',
na=False)]

# Hapus baris yang tidak memiliki data di dalam kolom 'status hasil
mediasi'
df = df.dropna(subset=['status hasil mediasi'])

# Terapkan pembersihan karakter aneh dan spasi berlebih pada kolom
'petitum'
df['petitum'] = df['petitum'].apply(clean_text)

# Hapus baris di mana kolom 'petitum' hanya berisi tag HTML atau
karakter titik
df = df[~df['petitum'].apply(is_html_only)]
```

Proses pengubahan dan penghapusan label serta pembersihan karakter dan simbol dimaksud turut melibatkan proses encoding untuk mengubah data teks menjadi numerik, lebih tepatnya menjadi byte, agar lebih mudah diproses oleh komputer. Karena penelitian tidak secara spesifik membahas penggunaan metode encoding maka fokus diarahkan pada hasil pembacaan yang berhasil. Hasil proses disajikan berikut ini:

```
# Print the total number of rows and the counts for each label
print("Total number of rows:", total_rows)
print("Number of rows with label 'Y':", count_Y)
print("Number of rows with label 'T':", count_T)
```

Hasil proses pengubahan dan penghapusan label serta pembersihan karakter dan simbol adalah dataset baru dengan kolom 'status hasil mediasi' berisi label T sebanyak 2390 baris dan label Y sebanyak 108 baris sehingga total dataset yang akan dianalisa lebih lanjut dalam tahapan preprocessing menjadi 2498 data.

4.3. Data Preprocessing

Tahapan ini dilakukan agar dataset lebih mudah digunakan untuk proses – proses selanjutnya, menjadikan dataset lebih seragam dan menghilangkan noise pada dataset. Terdapat beberapa teknik pre-processing yang sering digunakan diantaranya segmentasi kalimat, *lower/upper case conversion, Tokenizing, POS (Part-Of-Speech) Tagging, Stopwords Removal, Punctuation Removal, Stemming dan Lemmatization* (Tabassum & Patil, 2020) dimana setiap teknik memiliki fungsi dan tujuannya masing – masing.

Teknik preprocessing yang digunakan dalam penelitian ini adalah Case Folding, Cleansing, Tokenization, dan Filtering atau Stop Word Removal sebagaimana penjelasan berikut ini:

4.3.1 Case-Folding

Pada tahapan ini dataset yang diperoleh dari proses seleksi dan preparation dibersihkan dan di-preprocess dengan melakukan Case-Folding hanya pada kolom 'petitum' menggunakan python sebagai berikut:

```
import pandas as pd
import re

# Nama file input dan output
input_file = "2_dataset_labelubah.csv"
output_file = "3_preprocessed_casefolding.csv"

# Membaca dataset dari file CSV dengan encoding latin1
df = pd.read_csv(input_file, encoding='latin')
```

```
# Fungsi untuk membersihkan karakter aneh
def clean_text(text):
    if pd.ina(text):  # Menghindari kesalahan jika ada nilai NaN
        return ''
    # Menghapus karakter non-standar
    clean_text = re.sub(r'[^\x0-\x7F]+', '', text)
    return clean_text

# Pastikan kolom 'petitum' ada di dataframe
if 'petitum' not in df.column:
    print("Kolom 'petitum' tidak ditemukan di file CSV.")
    exit()

# Melakukan pembersihan karakter aneh dan case-folding pada kolom
'petitum'
df['petitum'] = df['petitum'].aply(clean_text).str.low()

# Menyimpan hasil preprocessing ke dalam file CSV dengan encoding
latin1
df.to_csv(output_file, index=False, encoding='latl')

print("Hasil preprocessing dengan case-folding dan pembersihan
karakter aneh telah disimpan dalam file
'3_preprocessed_casefolding.csv'.")
```

Terhadap kolom petitum dalam contoh dataset di Tabel 4.4 sebelumnya, dilakukan casefolding yang hasilnya disajikan dalam tabel 4.5 di bawah ini:

Tabel 4.5. Hasil Case Folding Kolom Petitum

nomor perkara atau nomor penetapan mediasi	Petitum	Status hasil mediasi
16/Pdt.G/2021/PN Mgn	- mengabulkan gugatan penggugat untuk seluruhnya : - menyatakan dalam hukum bahwa perkawinan penggugat dan tergugat yang telah dilangsungkan pada tanggal25 juni tahun 2012 di gereja masehi injili talaud (germita) jemaat gereja paradise melonguane, dan telah pula didaftarkan dan dicatatkan pada kantor dinas kependudukan dan pencatatan sipil kabupaten kepulauan talaud, dengan kutipan akta perkawinan nomor: 7104/cpk/25062012.0294tanggal25 juni 2012 adalah putus karena perceraian dengan segala akibat hukumnya; - menetapkan penggugat dan juga tergugat sebagai wali asuh anak-anak masih di bawah umur dari hasil perkawinan penggugat dan juga tergugat, yaitu bernama jovanca marselen bertji riung dan gabriela kenya evelin riung berada dalam pengasuhan dan pemeliharaan nafkah penggugat dan juga tergugat sampai dewasa. - memerintahkan kepada panitra pengadilan negeri melonguane untuk mengirim salinan putusan yang telah memperoleh kekuatan hukum tetap kepada kantor dinas kependudukan …dst	Y

Tabel 4.5. Lanjutan

nomor perkara atau nomor penetapan mediasi	Petitum	Status hasil mediasi
18/Pdt.G/2021/PN Mgn	berdasarkan atas alasan-alasan tersebut diatas, maka penggugat mohon dengan hormat kiranya pengadilan cq majelis hakim berkenan memeriksa dan mengadili perkara ini dengan menjatuhkan putusan yang amarnya pada pokoknya berbunyi sebagai berikut : - mengabulkan gugatan penggugat untuk seluruhnya; - menyatakan bahwa para tergugat dalam hal ini tergugat i, tergugat ii, tergugat iii, tergugat iv, tergugat v, tergugat vi dan tergugat vii telah melakukan perbuatan melawan hukum; ...dst	T
	- mengabulkan gugatan penggugat untuk seluruhnya; - menyatakan menurut hukum bahwa perkawinan antara penggugat dan tergugat yang dilangsungkan di melonguane pada tanggal 20 juni 2011, sesuai dengan kutipan akta perkawinan nomor : 710407/cpk/2006110029, yang dikeluarkan oleh pegawai pencatatan sipil kecamatan melonguane kabupaten kepulauan talaud pada tanggal 20 juni 2011, putus karena perceraian - memerintahkan kepada panitera pengadilan negeri melonguane untuk mengirimkan salinan putusan yang telah memperoleh kekuatan hukum yang tetap kepada kepala kantor dinas kependudukan dan catatan sipil kabupaten kepulauan talaud untuk mencatat perceraian ini dalam register yang telah disediakan untuk itu; - menghukum tergugat untuk membayar biaya perkara; - mohon keadilan;	T

Hasil pemrosesan dalam kolom petitum di atas tampak bahwa case folding berlaku hanya untuk huruf, sehingga yang dilakukan adalah mengubah huruf dalam setiap kata semuanya menjadi huruf kecil. Case Folding tidak akan mempengaruhi angka dan karakter. Tampak dalam kolom petitum masih berisi tag HTML, tanda baca, simbol dan juga angka. Hasil case folding disimpan dalam file 3_preprocessed_casefolding.csv. Juga tampak pada kolom nomor perkara atau nomor penetapan mediasi terdapat baris yang tidak memiliki identitas. Untuk tahapan berikut akan dilakukan pembersihan yang diperlukan.

4.3.2 Cleansing

Pada tahapan ini, cleansing dilakukan untuk membersihkan data dengan menghilangkan tag HTML, tanda baca, simbol dan juga angka. Proses cleansing pada kolom 'petitum' hasil case folding dilakukan dengan cermat tanpa mempengaruhi bagian – bagian teks maupun kolom yang lain. Selain itu dilakukan juga penghapusan baris – baris yang tidak memiliki identitas nomor perkara (ditetapkan jumlah karakter pada kolom ini >=8). Proses cleansing dengan python ditunjukkan berikut ini:

```
# Fungsi untuk membersihkan HTML dan karakter aneh
def clean_text(text):
    if pd.isn(text):  # Menghindari kesalahan jika ada nilai NaN
        return ''
    # Menghapus tag HTML
    soup = BeautifulSoup(text, "html.parse")
    clean_text = soup.gettext()
    # Menghapus karakter non-standar dan angka
    clean_text = re.sub(r'[^\x0-\x7F]+', '', clean_text)
    clean_text = re.sub(r'\d+', '', clean_text)
    # Menghapus tanda baca dan simbol
    clean_text = re.sub(r[^\w\s]', '', clean_text)
    # Menghapus spasi yang lebih dari satu
    clean_text = re.sub(r\s+', ' ', clean_text).strip()
    return clean_text

# Pastikan kolom 'petitum' ada di dataframe
if 'petitum' not in df.column:
    print("Kolom 'petitum' tidak ditemukan di file CSV.")
    exit()

# Melakukan pembersihan HTML dan karakter lainnya pada kolom
'petitum'
df['petitum'] = df['petitum'].aply(clean_text)

# Menyimpan hasil pembersihan ke dalam file CSV baru dengan encoding
latin1
df.to_csv(output_file, index=False, encoding='latin1')

print(f"Hasil pembersihan HTML, karakter aneh, simbol, angka, dan

tanda baca telah disimpan dalam file '{output_file}'.")
```

Proses cleansing tersebut menggunakan beberapa Library yaitu Pandas untuk manipulasi dan analisis data, Beautiful Soup untuk HTML Removal dan Regex atau Regular Expression untuk menghapus karakter – karakter non standar, angka, tanda baca/punctutation, simbol dan spasi berlebih.

Dengan menggunakan contoh dari tabel 4.5, hasil cleansing disajikan ke dalam tabel 4.6 berikut ini:

Tabel 4.6. Hasil Cleansing Kolom Petitum

nomor perkara	petitum	status hasil mediasi
16/Pdt.G/2021/PN Mgn	mengabulkan gugatan penggugat untuk seluruhnya menyatakan dalam hukum bahwa perkuwinan penggugat dan tergugat yang telah dilangsungkan pada tanggal juni tahun di gereja masehi injili talaud germita jemaat gereja paradise melonguane dan telah pula didaftarkan dan dicatatkan pada kantor dinas kependudukan dan pencatatan sipil kabupaten kepulauan talaud dengan kutipan akta perkuwinan nomor cpktanggal juni adalah putus karena perceraian dengan segala akibat hukumnya menetapkan penggugat dan juga tergugat sebagai wali asuh anakanak masih di bawah umur dari hasil perkawinan penggugat dan juga tergugat yaitu bernama jovanca marselea bertji riung dan gabriela kenya evelin riung berada dalam pengasuhan dan pemeliharaan nafkah penggugat dan juga tergugat sampai dewasa memerintahkan kepada panitra pengadilan negeri melonguane untuk mengirim salinan putusan yang telah memperoleh kekuatan hukum tetap kepada kantor dinas kependudukan dan catatan sipil kabupaten kepulauan talaud untuk didaftarkan dalam buku register perceraian …dst	Y
18/Pdt.G/2021/PN Mgn	berdasarkan atas alasanalasan tersebut diatas maka penggugat mohon dengan hormat kiranya pengadilan cq majelis hakim berkenan memeriksa dan mengadili perkara ini dengan menjatuhkan putusan yang amarnya pada pokoknya berbunyi sebagai berikut mengabulkan gugatan penggugat untuk seluruhnya menyatakan bahwa para tergugat dalam hal ini tergugat i tergugat ii tergugat iii tergugat iv tergugat v tergugat vi dan tergugat vii telah melakukan perbuatan melawan hukum menyatakan sah serta memiliki kekuatan hukum surat keterangan penyerahan tanah tanggal juni dan surat wasiat tanggal november menyatakan sebagai hukum bahwa tanah obyek sengketa disebut sebagai lapangan bola yang terletak di desa niampak kecamatan beo selatan kabupaten kepulauan talaud ukuran kurang lebih m enam ribu enam ratus lima puluh meter persegi terdiri dari satu hamparan yang sebagian kecilnya di tengahnya ada jalan kampung dengan batasbatas tanah..dst	T

Proses cleansing bertujuan menghilangkan atau memperbaiki elemen – elemen yang tidak diinginkan atau tidak relevan dari data sehingga meningkatkan kualitas dan kebersihan data dalam kolom petitum. Selain itu, kolom yang tidak memiliki nomor perkara juga telah dihapus. Akan tetapi, dalam pemeriksaan manual ditemukan kata – kata yang saling menempel atau bersambung dalam kolom petitum. Sebagian besar kata – kata bersambung ini adalah bawaan dari dataset awal yang berasal dari data yang dieksport dari database Sistem Informasi Perkara dan memerlukan penanganan khusus dalam tahapan – tahapan selanjutnya.

Hasil cleansing adalah dataset baru yang terdiri dari 1018 baris dengan jumlah baris berlabel status hasil mediasi 'Y' sebanyak 40 baris dan jumlah baris berlabel status hasil mediasi 'T' sebanyak 978 baris. Keseluruhan dataset hasil cleansing disimpan dalam file 4_preprocessed_cleaning.csv.

4.3.3 Tokenization

Dalam penelitian ini *Tokenization* dilakukan terhadap kolom 'petitum' pada dataset yang telah melalui tahapan cleansing. Token atau kata merujuk pada unit terkecil dari teks yang diidentifikasi dan diproses secara individu dari sebuah teks. Hasil dari proses ini disimpan dalam kolom baru yang dinamakan 'petitum_tokenized'.

Proses Tokenization menggunakan tokenisasi kata (menggunakan perintah `'word_tokenize'` dari library NLTK) pada setiap entri dalam kolom petitum. Dengan menggunakan contoh hasil cleansing pada Tabel 4.6 di bagian sebelumnya, proses tokenisasi menghasilkan kolom baru petitum_tokenized dengan isinya dalam Tabel 4.7 sebagai berikut:

Tabel 4.7. Contoh hasil tokenisasi

nomor perkara atau nomor penetapan mediasi	petitum	status hasil mediasi	petitum_tokenized
16/Pdt.G/2021/PN Mgn	mengabulkan gugatan penggugat untuk seluruhnya menyatakan dalam hukum bahwa perkawinan penggugat dan tergugat yang telah dilangsungkan pada tanggal juni tahun di gereja masehi injili talaud germita jemaat gereja paradise melonguane dan telah pula didaftarkan dan dicatatkan pada kantor dinas kependudukan dan pencatatan sipil kabupaten kepulauan talaud dengan kutipan akta perkawinan nomor cpktanggal juni adalah putus karena perceraian dengan segala akibat hukumnya menetapkan penggugat dan juga tergugat sebagai wali asuh anakanak masih di bawah umur dari hasil perkawinan penggugat dan juga tergugat yaitu bernama jovanca marselen bertji riung dan gabriela kenya evelin riung berada dalam pengasuhan dan pemeliharaan nafkah penggugat dan juga tergugat sampai dewasa memerintahkan kepada panitra pengadilan negeri melonguane untuk mengirim salinan putusan yang telah memperoleh kekuatan hukum tetap kepada kantor dinas kependudukan dan catatan sipil kabupaten kepulauan talaud untuk didaftarkan dalam buku register perceraian yang sedang berjalan menetapkan biayaperkara menurut hukum apabila pengadilan berpendapat lain mohon putusan yang seadiladilnya ex aequo et bono	Y	['mengabulkan', 'gugatan', 'penggugat', 'untuk', 'seluruhnya', 'menyatakan', 'dalam', 'hukum', 'bahwa', 'perkawinan', 'penggugat', 'dan', 'tergugat', 'yang', 'telah', 'dilangsungkan', 'pada', 'tanggal', 'juni', 'tahun', 'di', 'gereja', 'masehi', 'injili', 'talaud', 'germita', 'jemaat', 'gereja', 'paradise', 'melonguane', 'dan', 'telah', 'pula', 'didaftarkan', 'dan', 'dicatatkan', 'pada', 'kantor', 'dinas', 'kependudukan', 'dan', 'pencatatan', 'sipil', 'kabupaten', 'kepulauan', 'talaud', 'dengan', 'kutipan', 'akta', 'perkawinan', 'nomor', 'cpktanggal', 'juni', 'adalah', 'putus', 'karena', 'perceraian', 'dengan', 'segala', 'akibat', 'hukumnya', 'menetapkan', 'penggugat', 'dan', 'juga', 'tergugat', 'sebagai', 'wali', 'asuh', 'anakanak', 'masih', 'di', 'bawah', 'umur', 'dari', 'hasil', 'perkawinan', 'penggugat', 'dan', 'juga', 'tergugat', 'yaitu', 'bernama', 'jovanca', 'marselen', 'bertji', 'riung', 'dan', 'gabriela', 'kenya', 'evelin', 'riung', 'berada', 'dalam', 'pengasuhan', 'dan', 'pemeliharaan', 'nafkah', 'penggugat', 'dan', 'juga', 'tergugat', 'sampai', 'dewasa', 'memerintahkan', 'kepada', 'panitra', 'pengadilan', 'negeri', 'melonguane', 'untuk', 'mengirim', 'salinan', 'putusan', 'yang', 'telah', 'memperoleh', 'kekuatan', 'hukum', 'tetap', 'kepada', 'kantor', 'dinas', 'kependudukan', 'dan', 'catatan', 'sipil', 'kabupaten', 'kepulauan', 'talaud', 'untuk', 'didaftarkan', 'dalam', 'buku', 'register', 'perceraian', 'yang', 'sedang', 'berjalan', 'menetapkan', 'biayaperkara', 'menurut', 'hukum', 'apabila', 'pengadilan', 'berpendapat', 'lain', 'mohon', 'putusan', 'yang', 'seadiladilnya', 'ex', 'aequo', 'et', 'bono']

Tabel 4.7. Lanjutan

nomor perkara atau nomor penetapan mediasi	petitum	status hasil mediasi	petitum_tokenized
18/Pdt.G/2021/PN Mgn	berdasarkan atas alasanalasan tersebut diatas maka penggugat mohon dengan hormat kiranya pengadilan cq majelis hakim berkenan memeriksa dan mengadili perkara ini dengan menjatuhkan putusan yang amarnya pada pokoknya berbunyi sebagai berikut mengabulkan gugatan penggugat untuk seluruhnya menyatakan bahwa para tergugat dalam hal ini tergugat i tergugat ii tergugat iii tergugat iv tergugat v tergugat vi dan tergugat vii telah melakukan perbuatan melawan hukum menyatakan sah serta memiliki kekuatan hukum surat keterangan penyerahan tanah tanggal juni dan surat wasiat tanggal november menyatakan sebagai hukum bahwa tanah obyek sengketa disebut sebagai lapangan bola yang terletak di desa niampak kecamatan beo selatan kabupaten kepulauan talaud ukuran kurang lebih m enam ribu enam ratus lima puluh meter persegi terdiri dari satu hamparan yang sebagian kecilnya di tengahnya ada jalan kampung dengan batasbatas tanah sebagai berikut utara berbatasan dengan gothar andasia timur berbatasan dengan aris sarendeng selatan berbatasan dengan jalan setapak sebelah dari jalan setaapak olden garasut dan obet ncjo riung barat berbatasan dengan hardi barao a maliatja mes babinsa selanjutnya tanah tersebut disebut sebagai obyek sengketa adalah milik secara sah dari penggugat berdasarkan surat keterangan penyerahan tanah tanggal juni dan surat wasiat tanggal november dari ayah penggugat almarhum bartholomeus masone kepada penggugat menyatakan batal dan tidak mempunyai kekuatan hukum yang mengikat bagi penggugat….dst	T	['berdasarkan', 'atas', 'alasanalasan', 'tersebut', 'diatas', 'maka', 'penggugat', 'mohon', 'dengan', 'hormat', 'kiranya', 'pengadilan', 'cq', 'majelis', 'hakim', 'berkenan', 'memeriksa', 'dan', 'mengadili', 'perkara', 'ini', 'dengan', 'menjatuhkan', 'putusan', 'yang', 'amarnya', 'pada', 'pokoknya', 'berbunyi', 'sebagai', 'berikut', 'mengabulkan', 'gugatan', 'penggugat', 'untuk', 'seluruhnya', 'menyatakan', 'bahwa', 'para', 'tergugat', 'dalam', 'hal', 'ini', 'tergugat', 'i', 'tergugat', 'ii', 'tergugat', 'iii', 'tergugat', 'iv', 'tergugat', 'v', 'tergugat', 'vi', 'dan', 'tergugat', 'vii', 'telah', 'melakukan', 'perbuatan', 'melawan', 'hukum', 'menyatakan', 'sah', 'serta', 'memiliki', 'kekuatan', 'hukum', 'surat', 'keterangan', 'penyerahan', 'tanah', 'tanggal', 'juni', 'dan', 'surat', 'wasiat', 'tanggal', 'november', 'menyatakan', 'sebagai', 'hukum', 'bahwa', 'tanah', 'obyek', 'sengketa', 'disebut', 'sebagai', 'lapangan', 'bola', 'yang', 'terletak', 'di', 'desa', 'niampak', 'kecamatan', 'beo', 'selatan', 'kabupaten', 'kepulauan', 'talaud', 'ukuran', 'kurang', 'lebih', 'm', 'enam', 'ribu', 'enam', 'ratus', 'lima', 'puluh', 'meter', 'persegi', 'terdiri', 'dari', 'satu', 'hamparan', 'yang', 'sebagian', 'kecilnya', 'di', 'tengahnya', 'ada', 'jalan', 'kampung', 'dengan', 'batasbatas', 'tanah', 'sebagai', 'berikut', 'utara', 'berbatasan', 'dengan', 'gothar', 'andasia', 'timur', 'berbatasan', 'dengan', 'aris', 'sarendeng', 'selatan', 'berbatasan', 'dengan', 'jalan', 'setapak', 'sebelah', 'dari', 'jalan', 'setaapak', 'olden', 'garasut', 'dan', 'obet', 'ncjo', 'riung', 'barat', 'berbatasan', 'dengan', 'hardi', 'barao', 'a', 'maliatja', 'mes', 'babinsa', 'selanjutnya', 'tanah', 'tersebut', 'disebut', 'sebagai', 'obyek', 'sengketa', 'adalah', 'milik', 'secara', 'sah', 'dari', 'penggugat', 'berdasarkan', 'surat', 'keterangan', 'penyerahan', 'tanah', 'tanggal', 'juni', 'dan', 'surat', ….dst

Hasil tokenisasi tersebut menunjukkan kalimat – kalimat di dalam kolom petitum diubah menjadi token atau kata dengan frekuensi kemunculan yang berbeda – beda. Selengkapnya hasil tokenisasi disimpan dalam file 5_preprocessed_tokenized_nltk.csv.

Pada beberapa baris terlihat proses tokenisasi tidak berhasil memecah kalimat – kalimat bersambung (ditandai dengan huruf tebal). Misalnya pada baris ke-3 dengan kolom petitum sebagai berikut:

[**'menerimadanmengabulkangugatanpenggugat'**, **'menyatakanmenurut-hukumbahwaperkawinanpenggugatdantergugat'**, **'yangdilaksanakandi'**, 'manado', **'putuskarenaperceraian'**, **'menyatakanbahwaanakdalam-perkawinan'**, 'penggugat', **'dantergugatyang'**, **'bernamaraissa'**, 'gloria', **'manganguwidibawahasuhandanpengawasandari'**, **'penggugatdan'**, **ter-gugatsampaianaktumbuhdewasamenurutundangundang'**, **'memohonkepada-pengadilannegerimelonguaneuntukmengirimturunan'**, **'salinanputusan-perceraianini'**, 'yang', **'sudahmemperolehkekuatanhukumtetapkepadakepala-dinaskependudukandancatatansipilkota'**, 'manado', **'dankepaladinaskepen-dudukandanpencatatansipilkabkep'**, 'talaud', 'di', 'melonguane', **'mohonkeadilan'**]

Terdapat beberapa kata bersambung dalam kolom petitum tersebut yaitu:

'menerimadanmengabulkangugatanpenggugat',

'menyatakanmenuruthukumbahwaperkawinanpenggugatdantergugat',

'yangdilaksanakandi',

'putuskarenaperceraian',

'menyatakanbahwaanakdalamperkawinan',

'dantergugatyang',

'bernamaraissa',

'manganguwidibawahasuhandanpengawasandari',

'penggugatdan',

'tergugatsampaianaktumbuhdewasamenurutundangundang',

'memohonkepadapengadilannegerimelonguaneuntukmengirimturunan',

'salinanputusanperceraianini',

'sudahmemperolehkekuatanhukumtetapkepadakepaladinaskependudukandancatatansipilkota',

''dankepaladinaskependudukandanpencatatansipilkabkep',

'mohonkeadilan'

Tokenisasi tidak dapat mendeteksi kata – kata bersambung tersebut dan tetap menganggapnya sebagai satu kata. Dengan python diperoleh gambaran bahwa dalam keseluruhan dataset terdapat sebanyak 657 kata bersambung (ditetapkan oleh penulis, kata bersambung adalah kata dengan panjang katanya > 16 karakter) dan berada pada 201 baris data. Kata – kata bersambung ini disimpan dalam file long_words.txt dan pada tahapan selanjutnya dilakukan proses stemming untuk menangani kata – kata bersambung tersebut sehingga untuk proses – proses selanjutnya kata – kata tersebut tidak akang mempengaruhi hasil pemodelan maupun akurasi pemodelan.

4.3.4 Stemming

Proses stemming pada tahapan ini dilakukan dengan menggunakan library Sastrawi secara khusus terhadap kata – kata bersambung yang ditemukan setelah proses tokenisasi di sub bab sebelumnya (file long_words.txt). Berikut ini adalah contoh hasil stemming yang dilakukan dengan python pada kata – kata panjang yang ditentukan sebagai kata yang memiliki >16 karakter:

```
# Nama file input
input_file = "long_words.txt"

# Membaca daftar kata panjang dari file teks
with open(input_file, 'r') as f:
    long_words = [line.strip() for line in f.readlines()]

# Melakukan stemming pada kata-kata panjang
stemmed_words = [(word, stemmer.stem(word)) for word in long_words]

# Menampilkan hasil
for original_word, stemmed_word in stemmed_words:
    print(f"Kata asli: {original_word}, Kata dasar:
{stemmed_word}")
```

Kode program tersebut menghasilkan:

```
Kata asli: onrechtmatigedaad, Kata dasar: onrechtmatigedaad
Kata asli: olehtergugatsekaligus, Kata dasar: olehtergugatsekaligus
Kata asli: menghukumtergugat, Kata dasar: menghukumtergugat
Kata asli: memberikanputusan, Kata dasar: memberikanputusan
Kata asli: menerimadanmengabulkangugatanpenggugat, Kata dasar:
menerimadanmengabulkangugatanpenggugat
Kata asli:
menyatakanmenuruthukumbahwaperkawinanpenggugatdantergugat, Kata
dasar: menyatakanmenuruthukumbahwaperkawinanpenggugatdantergugat
Kata asli: yangdilaksanakandi, Kata dasar: yangdilaksanakandi
Kata asli: putuskarenaperceraian, Kata dasar: putuskarenaperceraian
Kata asli: menyatakanbahwaanakdalamperkawinan, Kata dasar:
menyatakanbahwaanakdalamperkawinan
Kata asli: manganguwidibawahasuhandanpengawasandari, Kata dasar:
manganguwidibawahasuhandanpengawasandari
```

Hasil stemming menunjukkan bahwa Library Sastrawi tidak dapat melakukan pemisahan kata – kata panjang yang ditemukan tersebut. Kemungkinan penyebabnya adalah karena stemmer tidak berhasil mengenali pola dan aturan untuk memecah dan menyederhanakan kata – kata tersebut. Penyebab lainnya

adalah kata – kata tersebut tidak memiliki komponen yang dikenal dalam kamus Sastrawi sehingga stemmer kesulitan mengidentifikasi kata dasar yang relevan. Hal yang dapat dilakukan adalah menambahkan kata – kata baru ke dalam kamus kata Sastrawi dan melakukan proses latih yang lebih intens terhadap stemmer Sastrawi. Akan tetapi, karena hal tersebut di luar lingkup penelitian maka tidak akan dilakukan dalam penelitian ini.

Sehingga, untuk kata – kata panjang yang telah diperoleh sebelumnya, agar tidak mempengaruhi pemodelan dan akurasinya tidak akan diikutsertakan dalam pemrosesan lebih lanjut dan dihapus dari dataset. Dengan python dilakukan penghapusan sebagai berikut:

```
input_file = "6_preprocessed_no_long_words.csv"

# Membaca dataset dari file CSV dengan encoding latin
filtered_df = pd.read_csv(input_file, encoding='latin')

# Menghitung jumlah baris dengan label 'T' dan 'Y' pada kolom
'status hasil mediasi'
label_count = filtered_df['status hasil mediasi'].value_count()

# Menampilkan hasil
total_row = filtered_df.shape[0]
label_t_count = label_counts.get('T', 0)
label_y_count = label_counts.get('Y', 0)

print(f"Total jumlah baris: {total_row}")
print(f"Jumlah baris dengan label 'T': {label_t_count}")
print(f"Jumlah baris dengan label 'Y': {label_y_count}")
```

Penghapusan menghasilkan 817 baris dengan jumlah baris berlabel T sebanyak 789 baris dan berlabel Y sebanyak 28 baris. Data ini akan diproses lebih lanjut dalam proses filtering atau stopword removal.

4.3.5 Filtering or Stop word removal

Pada tahapan ini dilakukan filtering dengan stop word removal pada kolom petitum_tokenized. Stopword adalah kata umum yang biasanya muncul namun tidak memiliki makna. Untuk keperluan tersebut digunakan library Sastrawi. Daftar stopword yang ada dalam kamus kata Sastrawi di antaranya adalah:

Proses stop word removal dilakukan dengan membandingkan hasil tokenisasi dengan kamus stopword Sastrawi. Fungsi StopWordRemover akan menghapus kata – kata umum hasil tokenisasi yang tidak memberikan banyak informasi dalam text processing. Proses stopword removal yang diimplementasikan pada contoh hasil tokenisasi di Sub Bab 4.3.3 adalah sebagai berikut.

Hasil tokenisasi:

['menerima', 'dan', 'mengabulkan', 'gugatan', 'para', 'penggugat', 'untuk', 'seluruhnya', 'menyatakan', 'tergugat', 'i', 'telah', 'melakukan', 'ingkar', 'janji', 'wanprestasi', 'menyatakan', 'putusan', 'dalam', 'perkara', 'nomor', 'pdtgspn', 'mgn', 'tidak', 'memiliki', 'kekuatan', 'hukum', 'yang', 'mengikat', 'bagi', 'para', 'penggugat', 'menyatakan', 'tidak', 'sah', 'dan', 'tidak', 'memiliki', 'kekuatan', 'hukum', 'yang', 'mengikat', 'sita', 'jaminan', 'yang', 'diletakan', 'oleh', 'pengadilan', 'negeri', 'melonguane', 'dalam', 'perkara', 'perdata', 'nomor', 'pdtgspn', 'mgn', 'atas', 'tanah', 'dan', 'bangunan', 'rumah', 'milik', 'para', 'penggugat', 'berdasarkan', 'sertifikat', 'hak', 'milik', 'nomor', 'melonguane', 'barat', 'kecamatan', 'melonguane', 'kabupaten', 'kepulauan', 'talaud', 'dengan', 'seluas', 'm', 'empat', 'ratus', 'sembilan', 'puluh', 'dua', 'meter', 'atas', 'nama', 'martji', 'sasauw', 'taliwuna', 'menghukum', 'tergugat', 'i', 'untuk', 'mengembalikan', 'uang', 'pinjaman', 'kepada', 'tergugat', 'ii', 'sejumlah', 'rp', 'dua', 'ratus', 'lima', 'puluh', 'juta', 'rupiah', 'berserta', 'bunga', 'sebagaimana', 'dalam', 'isi', 'surat', 'perjanjian', 'tertanggal', 'januari', 'membebankan', 'biaya', 'perkara', 'kepada', 'tergugat', 'i', 'dan', 'tergugat', 'ii'].

Menjadi:

['menerima', 'mengabulkan', 'gugatan', 'penggugat', 'seluruhnya', 'menyatakan', 'tergugat', 'i', 'melakukan', 'ingkar', 'janji', 'wanprestasi', 'menyatakan', 'putusan', 'perkara', 'nomor', 'pdtgspn', 'mgn', 'memiliki', 'kekuatan', 'hukum', 'mengikat', 'penggugat', 'menyatakan', 'sah', 'memiliki', 'kekuatan', 'hukum', 'mengikat', 'sita', 'jaminan', 'diletakan', 'pengadilan', 'negeri', 'melonguane', 'perkara', 'perdata', 'nomor', 'pdtgspn', 'mgn', 'atas', 'tanah', 'bangunan', 'rumah', 'milik', 'penggugat', 'berdasarkan', 'sertifikat', 'hak', 'milik', 'nomor', 'melonguane', 'barat', 'kecamatan', 'melonguane', 'kabupaten', 'kepulauan', 'talaud', 'seluas', 'm', 'empat', 'ratus', 'sembilan', 'puluh', 'meter', 'atas', 'nama', 'martji', 'sasauw', 'taliwuna', 'menghukum', 'tergugat', 'i', 'mengembalikan', 'uang', 'pinjaman', 'tergugat', 'ii', 'sejumlah', 'rp', 'ratus', 'lima', 'puluh', 'juta', 'rupiah', 'berserta', 'bunga', 'sebagaimana', 'isi', 'surat', 'perjanjian', 'tertanggal', 'januari', 'membebankan', 'biaya', 'perkara', 'tergugat', 'i', 'tergugat', 'ii']

Hasil Stop word removal di atas menunjukkan hal – hal sebagai berikut: Jumlah kata sebelum penghapusan stopword sebanyak 126 kata. Jumlah kata setelah penghapusan stopword sebanyak 100 kata. Jumlah stopword yang dihapus sebanyak 26 kata, yaitu: ['dan', 'para', 'untuk', 'telah', 'dalam', 'tidak', 'yang', 'bagi', 'para', 'tidak', 'dan', 'tidak', 'yang', 'yang', 'oleh', 'dalam', 'dan', 'para', 'dengan', 'dua', 'untuk', 'kepada', 'dua', 'dalam', 'kepada', 'dan']. Hasil stop word removal menunjukkan kata – kata makin sedikit dan makin spesifik.

Hasil stop word removal untuk keseluruhan dataset disimpan dalam file 7_preprocessed_stopwords_removed.csv.

Secara keseluruhan, hasil dari Data Preprocessing menghasilkan dataset yang siap untuk diproses lebih lanjut sesuai alur penelitian yaitu implementasi n-gram.

4.4. Pemisahan Data Latih dan Data Uji

Pada tahapan ini dilakukan Pemisahan data latih dan data uji dari total dataset sebanyak 817 data. Pemisahan dilakukan dengan komposisi 70:30 sehingga total data untuk data latih adalah sebanyak 571 data dan data uji adalah sebanyak 246 data. Masing – masing disimpan dalam file terpisah yaitu 8_data_latih.csv dan 9_data_uji.csv.

4.5 Implementasi N-gram

N-gram adalah urutan n unit yang biasanya terdiri dari karakter tunggal atau string yang dipisahkan oleh spasi. Dalam tahapan penelitian ini implementasi N-grams dilakukan untuk memahami konteks teks yang ada dalam dataset hasil preprocessing pada tahapan sebelumnya. Proses ini meng-*generate* N-grams dengan menggunakan persamaan 1) untuk mengekstrak fitur dari teks, dengan fitur disini adalah item atau term yang berupa kata.

$$Ngram_k = X - (N - 1) \quad (5)$$

Dimana $Ngram_k$ adalah banyaknya n-gram dalam k kalimat , X adalah jumlah kata dalam kalimat k, N adalah tipe n-gram yang digunakan, bisa berupa unigram, bigram, trigram dan seterusnya.

Pada tahapan ini, n-gram diimplentasikan pada dataset data latih yaitu dataset yang telah mengalami proses stop word removal, secara khusus pada kolom petitum_tokenized.

Implementasi n-gram pada penelitian ini akan menggunakan 5 skenario, yaitu n = 1 sampai n = 5. Adapun alasan pemilihan skenario ini adalah yang pertama, kekayaan konteks. Penggunaan n = 1 atau unigram dapat membantu mengidentifikasi kata – kata kunci dan frekuensi kata sehingga kata secara individu dianalisis tanpa konteks sekitar. Penggunaan n = 2 atau bigram pada kalimat bisa lebih informatif dibandingkan unigram karena menangkap pasangan kata dan memberikan informasi tentang kedua kata yang berdampingan tersebut. Penggunaan n = 3 atau trigram memberikan makna yang lebih kaya secara konteks dan membantu memahami struktur kalimat. Penggunaan n = 4 dan n = 5 untuk menangkap pola yang lebih panjang sehingga diperoleh pemahaman atas frase yang lebih kompleks. Alasan lainnya adalah menggunakan n-gram yang bervariasi dapat membantu pembuatan model yang lebih akurat karena mempertimbangkan lebih dari satu gram pada satu waktu.

Sebagai contoh menggunakan isi kolom petitum_tokenized sebagai berikut:

Hasil tokenisasi:

['menerima', 'dan', 'mengabulkan', 'gugatan', 'para', 'penggugat', 'untuk', 'seluruhnya', 'menyatakan', 'tergugat', 'i', 'telah', 'melakukan', 'ingkar', 'janji', 'wanprestasi', 'menyatakan', 'putusan', 'dalam', 'perkara', 'nomor', 'pdtgspn', 'mgn', 'tidak', 'memiliki', 'kekuatan', 'hukum', 'yang', 'mengikat', 'bagi', 'para', 'penggugat', 'menyatakan', 'tidak', 'sah', 'dan', 'tidak', 'memiliki', 'kekuatan', 'hukum', 'yang', 'mengikat', 'sita', 'jaminan', 'yang', 'diletakan', 'oleh', 'pengadilan', 'negeri', 'melonguane', 'dalam', 'perkara', 'perdata', 'nomor', 'pdtgspn', 'mgn', 'atas', 'tanah', 'dan', 'bangunan', 'rumah', 'milik', 'para', 'penggugat', 'berdasarkan', 'sertifikat', 'hak',

'milik', 'nomor', 'melonguane', 'barat', 'kecamatan', 'melonguane', 'kabupaten', 'kepulauan', 'talaud', 'dengan', 'seluas', 'm', 'empat', 'ratus', 'sembilan', 'puluh', 'dua', 'meter', 'atas', 'nama', 'martji', 'sasauw', 'taliwuna', 'menghukum', 'tergugat', 'i', 'untuk', 'mengembalikan', 'uang', 'pinjaman', 'kepada', 'tergugat', 'ii', 'sejumlah', 'rp', 'dua', 'ratus', 'lima', 'puluh', 'juta', 'rupiah', 'berserta', 'bunga', 'sebagaimana', 'dalam', 'isi', 'surat', 'perjanjian', 'tertanggal', 'januari', 'membebankan', 'biaya', 'perkara', 'kepada', 'tergugat', 'i', 'dan', 'tergugat', 'ii'].

Implementasi n-gram untuk contoh unigram di atas dilakukan dengan langkah – langkah sebagai berikut:

1) Menghitung jumlah kata dalam kalimat. Jumlah kata dalam kalimat adalah (X) sehingga (X) = 126
2) Nilai n dapat berupa unigram atau n = 1, bigram atau n = 2, trigram atau n = 3 dan seterusnya. Menghitung N-gram menggunakan persamaan 1) berikut:

$$Ngram_k = X - (N - 1) \quad (6$$

Maka, untuk n = 1 atau unigram dapat dihitung sebagai berikut:

$Ngram_i$ = 126 – (1 – 1)

= 126

Artinya untuk n = 1, tidak ada perubahan pada susunan kalimat. N-gram yang diperoleh adalah sama dengan banyaknya jumlah kata.

Selanjutnya, untuk n = 2 atau bigram,

$Ngram_i$ = 126 – (2 – 1)

= 126 – 1

= 125

Maka untuk kalimat di atas, bigramnya adalah unigram – 1. Artinya, untuk n = 2 kombinasi n-gram akan menjadi sebanyak 125

Selanjutnya untuk n = 3 atau trigram,

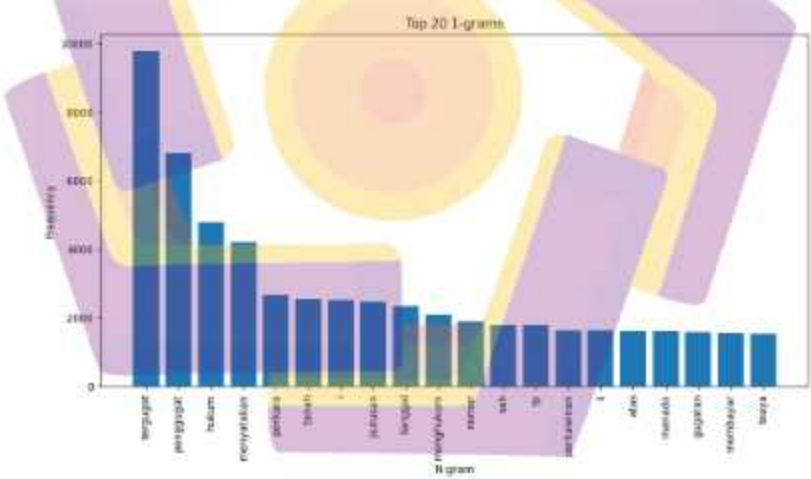
$$Ngram_i = 126 - (3 - 1)$$

$$= 126 - 2$$

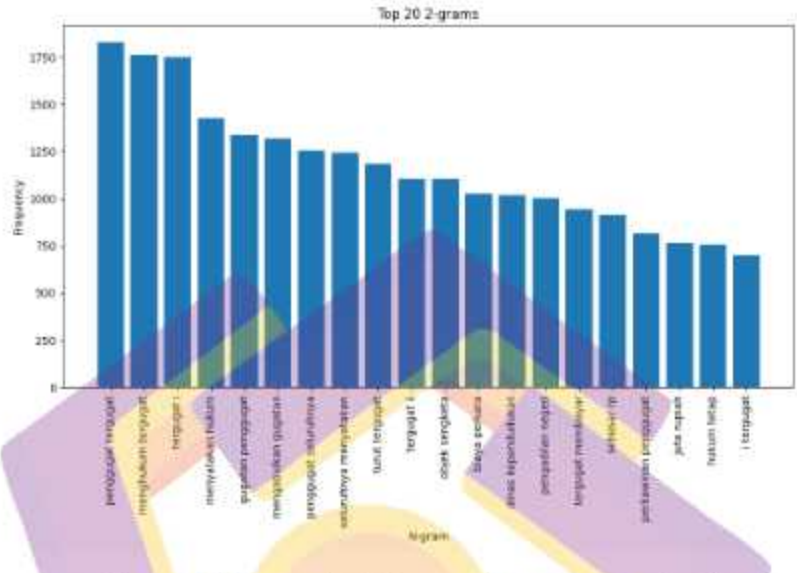
$$= 124$$

Artinya, untuk n = 3 kombinasi n-gram akan menjadi sebanyak 124, dan seterusnya.

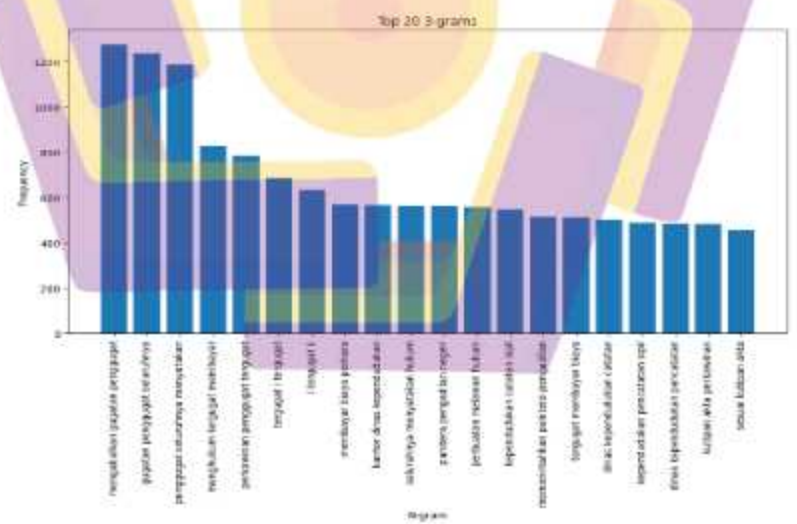
Untuk keseluruhan file hasil preprocessing, hasil plot pada bar graph Gambar 4.1 sampai dengan Gambar 4.5 berikut ini menunjukkan 20 n-grams yang paling banyak muncul untuk n = 1 sampai dengan n = 5 pada data latih.



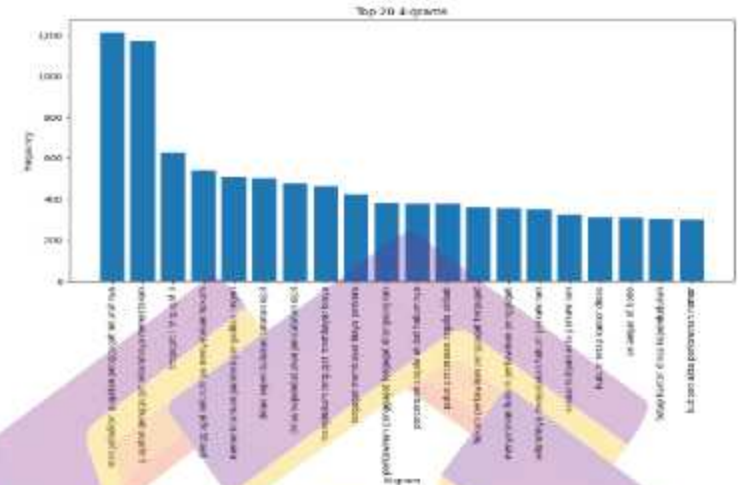
Gambar 4.1. N-grams data latih yang paling sering muncul untuk n=1



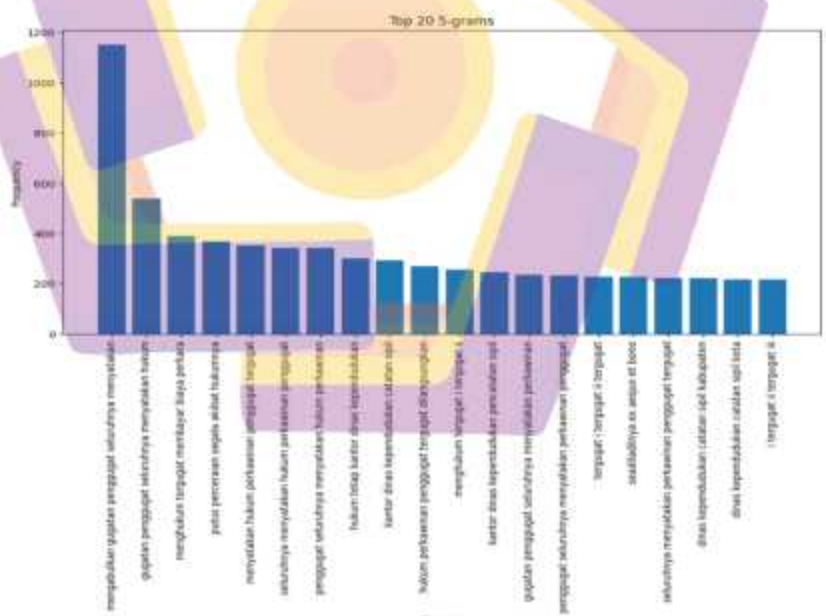
Gambar 4.2. N-grams data latih yang paling sering muncul untuk n=2



Gambar 4.3. N-grams data latih yang paling sering muncul untuk n=3



Gambar 4.4. N-grams data latih yang paling sering muncul untuk n=4



Gambar 4.5. N-grams data latih yang paling sering muncul untuk n=5

Berdasarkan hasil – hasil perhitungan di atas, dapat ditunjukkan bahwa bentuk n-gram untuk setiap baris akan berbeda. Hal ini dapat disebabkan oleh panjang – pendeknya isi petitum. Dalam konteks dokumen pengadilan, petitum dalam surat gugatan oleh pihak berperkara berisi permintaan pihak berperkara atau tuntutan untuk dapat diputuskan Majelis Hakim. Petitum dapat berbentuk tunggal (gugatan primer) maupun memiliki gugatan alternatif atau subsidair.

Di dalam penelitian ini digunakan kolom petitum yang telah dipreprocessing untuk menghitung n-gram. Untuk melakukan ekstraksi n-gram ditentukan n = 1 sampai dengan n = 5. Hasil implementasi n-gram pada data latih menunjukkan jumlah fitur dalam dataset dimana setiap n-gram yang dihasilkan dari teks dianggap sebagai fitur yang terpisah. Jumlah total fitur dapat menjadi sangat besar, apabila korpus teksnya besar karena dimensinya akan bertambah secara eksponensial. Pada implementasi dengan python, peningkatan dimensi dapat memperbesar kebutuhan memori dan waktu komputasi.

Pada n yang kecil, terutama unigram, informasi kontekstual tidak dapat ditangkap secara sempurna karena kata – kata berdiri sendiri dan bersifat umum. Pada n yang besar, misalnya trigram ke atas, walaupun dapat menangkap lebih banyak informasi kontekstual namun frekuensi kemunculannya akan makin jarang. Untuk itu dibutuhkan proses lebih lanjut, misalnya dengan pembobotan agar dapat diidentifikasi fitur – fitur atau term mana yang lebih penting.

4.6 Implementasi TF – IDF (Term Frequency-Inverse Document Frequency)

Dalam penelitian ini TF-IDF digunakan untuk memberikan bobot pada fitur – fitur atau term yang dihasilkan pada implementasi N-gram. Untuk mengkonversikan data teks ke dalam fitur numerik, digunakan TF-IDF (Term Frequency-Inverse Document frequency) yang menghitung frekuensi kemunculan frase dan signifikansinya dalam dokumen. Untuk menghitung TF-IDF dari n-gram yang diperoleh, menggunakan persamaan 2 berikut ini:

$$TF * IDF\ (d,t) = TF(d,t) * log\frac{N}{df(t)} \quad (7$$

Dimana $TF * IDF\ (d,t)$ adalah Bobot TF-IDF, $TF(d,t)$ adalah Jumlah munculnya term t pada dokumen d, N adalah Total dokumen (korpus) dan $df(t)$ adalah Jumlah dokumen yang di dalamnya mengandung term t. Nilai TF dan IDF dihitung terpisah kemudian dikalikan. Dalam penelitian ini, TF-IDF akan dihitung dari 5 skenario n-gram yang dihasilkan dari data latih pada tahapan sebelumnya, yaitu dengan n=1 sampai n=5.

Contoh perhitungan TF-IDF disajikan berikut ini: Diketahui baris perkara 310/Pdt.G/2023/PN Tnn, hasil implementasi unigram atau n = 1 adalah:

['menerima', 'mengabulkan', 'permohonan', 'penggugat', 'seluruhnya', 'menyatakan', 'surat', 'harta', 'nikah', 'tertanggal', 'agustus', 'disahkan', 'turut', 'tergugat', 'i', 'surat', 'keterangan', 'waris', 'disahkan', 'turut', 'tergugat', 'i', 'teregsiter', 'nomor', 'skwx', 'tanggal', 'oktober', 'turut', 'tergugat', 'ii', 'teregister', 'nomor', 'tanggal', 'april', 'sah', 'memiliki', 'kekuatan', 'hukum', 'mengikat', 'menyatakan', 'hukum', 'penggugat', 'pemilik', 'sah', 'atas', 'satu', 'bidang', 'tanah', 'bangunan', 'rumah', 'setengah', 'bagian', 'belakang', 'diatas', 'tanah', 'teletak', 'kelurahan', 'tonsaru', 'lingkungan', 'iv',

'kecamatan', 'tondano', 'selatan', 'kabupaten', 'minahasa', 'sulawesi', 'utara', 'batasbatas', 'sekarang', 'berikut', 'sebelah', 'utara', 'berbatas', 'tanah', 'pratasik', 'kapoyos', 'polii', 'kawilaran', 'sebelah', 'selatan', 'berbatas', 'jalan', 'utama', 'sebelah', 'timur', 'berbatas', 'tanah', 'r', 'kabesi', 'kabesi', 'mawicere', 'sebelah', 'barat', 'berbatas', 'sumanti', 'korengkeng', 'boyoh', 'sumuruk', 'menyatakan', 'tergugat', 'turut', 'tergugat', 'melakukan', 'perbuatan', 'melawan', 'hukum', 'onrecht', 'matigedaad', 'menyatakan', 'hukum', 'segalah', 'alas', 'hak', 'timbul', 'atas', 'objek', 'sengketa', 'bukan', 'atas', 'nama', 'penggugat', 'sah', 'karenanya', 'batal', 'hukum', 'baik', 'tergugat', 'siapa', 'memperoleh', 'hak', 'daripadanya', 'mengosongkan', 'menyerahkan', 'objek', 'sengketa', 'penggugat', 'keadaan', 'baik', 'beban', 'hak', 'apapun', 'memerintahkan', 'turut', 'tergugat', 'i', 'sd', 'turut', 'tergugat', 'iii', 'tunduk', 'patuh', 'isi', 'putusan', 'menghukum', 'tergugat', 'membayar', 'biaya', 'perkara', 'timbul']

Untuk menghitung nilai TF-IDF dari hasil unigram, maka langkah – langkahnya adalah sebagai berikut:

1) Hitung Jumlah munculnya unigram $TF_{(d,t)}$

$TF_{(d,t)}$ = Jumlah munculnya term (t) pada dokumen (d)

= Jumlah munculnya unigram pada kalimat tersebut

Jumlah munculnya unigram $TF_{(d,t)}$ *adalah* sebagaimana Tabel 4.8 berikut ini:

Tabel 4.8. TF unigram dalam kalimat

Unigram	TF
menerima	0.006289
mengabulkan	0.006289
permohonan	0.006289
penggugat	0.025157
seluruhnya	0.006289
menyatakan	0.025157
surat	0.012579
harta	0.006289
nikah	0.006289
tertanggal	0.006289
agustus	0.006289
disahkan	0.012579
turut	0.037736
tergugat	0.056604
i	0.018868
surat	0.012579
keterangan	0.006289
waris	0.006289
disahkan	0.012579
turut	0.037736
tergugat	0.056604
i	0.018868
teregister	0.012579
nomor	0.012579
skwx	0.006289
tanggal	0.012579
oktober	0.006289
turut	0.037736
tergugat	0.056604
ii	0.006289
teregister	0.012579
nomor	0.012579
tanggal	0.012579
april	0.006289
sah	0.018868
memiliki	0.006289

Unigram	TF
kekuatan	0.006289
hukum	0.031447
mengikat	0.006289
menyatakan	0.025157
hukum	0.031447
penggugat	0.025157
pemilik	0.006289
sah	0.018868
atas	0.018868
satu	0.006289
bidang	0.006289
tanah	0.025157
bangunan	0.006289
rumah	0.006289
setengah	0.006289
bagian	0.006289
belakang	0.006289
diatas	0.006289
tanah	0.025157
teletak	0.006289
kelurahan	0.006289
tonsaru	0.006289
lingkungan	0.006289
iv	0.006289
kecamatan	0.006289
tondano	0.006289
selatan	0.012579
kabupaten	0.006289
minahasa	0.006289
sulawesi	0.006289
utara	0.012579
batasbatas	0.006289
sekarang	0.006289
berikut	0.006289
sebelah	0.025157
utara	0.012579

Unigram	TF
berbatas	0.025157
tanah	0.006289
pratasik	0.006289
kapoyos	0.006289
polii	0.006289
kawilaran	0.025157
sebelah	0.012579
selatan	0.025157
berbatas	0.006289
jalan	0.006289
utama	0.025157
sebelah	0.006289
timur	0.025157
berbatas	0.025157
tanah	0.006289
r	0.012579
kabesi	0.012579
kabesi	0.006289
mawicere	0.025157
sebelah	0.006289
barat	0.025157
berbatas	0.006289
sumanti	0.006289
korengkeng	0.006289
boyoh	0.006289
sumuruk	0.025157
menyatakan	0.056604
tergugat	0.037736
turut	0.056604
tergugat	0.006289
melakukan	0.006289
perbuatan	0.006289
melawan	0.031447
hukum	0.006289
onrecht	0.006289
matigedaad	0.025157

Tabel 4.8. Lanjutan

Unigram	TF
Menyatakan	0.025157
hukum	0.031447
segalah	0.006289
alas	0.006289
hak	0.018868
timbul	0.012579
atas	0.018868
objek	0.012579
sengketa	0.012579
bukan	0.006289
atas	0.018868
nama	0.006289
penggugat	0.025157
sah	0.018868
karenanya	0.006289
batal	0.006289
hukum	0.031447

Unigram	TF
baik	0.012579
tergugat	0.056604
siapa	0.006289
memperoleh	0.006289
hak	0.018868
daripadanya	0.006289
mengosongkan	0.006289
menyerahkan	0.006289
objek	0.012579
sengketa	0.012579
penggugat	0.025157
keadaan	0.006289
baik	0.012579
beban	0.006289
hak	0.018868
apapun	0.006289
memerintahkan	0.006289

Unigram	TF
turut	0.037736
tergugat	0.056604
i	0.018868
sd	0.006289
turut	0.037736
tergugat	0.056604
iii	0.006289
tunduk	0.006289
patuh	0.006289
isi	0.006289
putusan	0.006289
menghukum	0.006289
tergugat	0.056604
membayar	0.006289
biaya	0.006289
perkara	0.006289
timbul	0.012579

Berdasarkan tabel di atas, dapat ditunjukkan bahwa jumlah keseluruhan unigram dalam kalimat adalah sebanyak 159.

2) Menghitung IDF (Inverse Document Frequency)

IDF(t) dihitung dengan rumus $log\frac{N}{df(t)}$ dimana N adalah total jumlah dokumen dalam korpus dan *df(t)* adalah jumlah dokumen yang mengandung token. Untuk menghitung IDF, dihitung jumlah dokumen dalam korpus. Dalam hal menggunakan contoh kalimat di atas sebagai satu dokumen dan dokumen tersebut berada dalam dataset data latih sebanyak 571 data, maka banyaknya dokumen dalam korpus adalah 571. Selanjutnya, dihitung jumlah dokumen yang mengandung token - token

tersebut dalam korpus. Setelah diperoleh jumlah dokumen yang mengandung token dalam korpus, selanjutnya adalah menghitung nilai IDF untuk setiap token.

3) Menghitung nilai TF-IDF

Nilai TF-IDF diperoleh dengan mengalikan hasil TF dan hasil IDF, sehingga untuk contoh di atas maka nilai TF-IDFnya adalah sebagaimana disajikan dalam tabel 4.9 di bawah ini:

Tabel 4.9. Hasil Perhitungan TF-IDF

n-gram	TF	IDF	TFIDF
agustus	0.052838	3.115032	0.164591
alas	0.083984	4.951244	0.415823
apapun	0.059715	3.520498	0.210227
april	0.052353	3.086459	0.161585
atas	0.100679	1.978501	0.199193
bagian	0.063408	3.738221	0.237034
baik	0.105675	3.115032	0.329181
bangunan	0.049102	2.894792	0.14214
barat	0.045878	2.704748	0.124089
batal	0.055208	3.254794	0.179691
batasbatas	0.049501	2.918322	0.14446
beban	0.0856	5.046554	0.431986
belakang	0.106022	6.250527	0.662695
berbatas	0.40457	5.962845	2.412388
berikut	0.050542	2.979691	0.150599
biaya	0.018128	1.068743	0.019374
bidang	0.067541	3.981843	0.268936
boyoh	0.1129	6.655992	0.75146
bukan	0.09165	5.403229	0.495207
daripadanya	0.073015	4.304617	0.314303
diatas	0.058301	3.437116	0.200387
disahkan	0.2258	6.655992	1.50292
hak	0.103601	2.035933	0.210925
harta	0.059715	3.520498	0.210227
hukum	0.090166	1.063141	0.095859
ii	0.039001	2.299283	0.089674
Iii	0.045555	2.6857	0.122348
isi	0.052353	3.086459	0.161585

Tabel 4.9. Lanjutan

n-gram	TF	IDF	TFIDF
iv	0.048906	2.883231	0.141006
jalan	0.050542	2.979691	0.150599
kabesi	0.2258	6.655992	1.50292
kabupaten	0.033959	2.002031	0.067986
kapoyos	0.1129	6.655992	0.75146
karenanya	0.094265	5.55738	0.523866
kawilaran	0.1129	6.655992	0.75146
keadaan	0.066409	3.915152	0.260002
kecamatan	0.041204	2.429158	0.10009
kekuatan	0.037345	2.201645	0.082219
kelurahan	0.050118	2.95469	0.148083
keterangan	0.066965	3.947942	0.264376
korengkeng	0.1129	6.655992	0.75146
lingkungan	0.062956	3.711553	0.233664
matigedaad	0.1129	6.655992	0.75146
mawicere	0.1129	6.655992	0.75146
melakukan	0.044026	2.595549	0.114272
melawan	0.036017	2.123392	0.076479
membayar	0.024536	1.446506	0.035491
memerintahkan	0.025836	1.523139	0.039351
memiliki	0.05935	3.498991	0.207666
memperoleh	0.05935	3.498991	0.207666
menerima	0.043593	2.570015	0.112034
mengabulkan	0.017474	1.030171	0.018001
menghukum	0.02426	1.430245	0.034698
mengikat	0.048331	2.849329	0.13771
mengosongkan	0.056378	3.323787	0.18739
menyatakan	0.069408	1.02299	0.071004
menyerahkan	0.047414	2.795262	0.132533
minahasa	0.038571	2.273965	0.08771
nama	0.043593	2.570015	0.112034
nikah	0.089385	5.269697	0.471033
nomor	0.052894	1.559179	0.082471
objek	0.083669	2.466337	0.206355
oktober	0.050758	2.99243	0.15189
onrecht	0.1129	6.655992	0.75146
Patuh	0.062086	3.66026	0.22725
pemilik	0.056378	3.323787	0.18739

Tabel 4.9. Lanjutan

n-gram	TF	IDF	TFIDF
penggugat	0.069166	1.019418	0.070509
perbuatan	0.035391	2.086449	0.073841
perkara	0.017907	1.05572	0.018905
permohonan	0.07075	4.171085	0.295106
polii	0.106022	6.250527	0.662695
pratasik	0.1129	6.655992	0.75146
putusan	0.018546	1.093389	0.020278
rumah	0.054652	3.222005	0.176089
sah	0.091204	1.792311	0.163466
satu	0.041328	2.436484	0.100695
sd	0.087387	5.151914	0.450212
sebelah	0.27501	4.053302	1.1147
segalah	0.097358	5.739701	0.558803
sekarang	0.061667	3.635567	0.224194
selatan	0.088052	2.595549	0.228543
seluruhnya	0.017782	1.048353	0.018642
sengketa	0.081673	2.407497	0.196626
setengah	0.097358	5.739701	0.558803
siapa	0.052116	3.072473	0.160124
Skwx	0.1129	6.655992	0.75146
sulawesi	0.057965	3.417313	0.198084
sumanti	0.106022	6.250527	0.662695
sumuruk	0.1129	6.655992	0.75146
surat	0.087472	2.578454	0.225543
tanah	0.145544	2.145132	0.312211
tanggal	0.048611	1.432937	0.069657
teletak	0.1129	6.655992	0.75146
teregister	0.1129	6.655992	0.75146
teregsiter	0.1129	6.655992	0.75146
tergugat	0.156442	1.02478	0.160319
tertanggal	0.042759	2.520825	0.107787
timbul	0.080953	2.386294	0.193178
timur	0.045082	2.657791	0.119818
tondano	0.046544	2.743969	0.127714
tonsaru	0.1129	6.655992	0.75146
tunduk	0.04706	2.774428	0.130565
Turut	0.280284	2.754019	0.771908
utama	0.097358	5.739701	0.558803

Tabel 4.9. Lanjutan

n-gram	TF	IDF	TFIDF
Utara	0.078883	2.325258	0.183423
waris	0.050758	2.99243	0.15189

Perhitungan nilai TF-IDF di atas dibuat dengan menggunakan term implementasi n-gram untuk unigram atau n = 1. Sehingga, untuk n = 2, n = 3 dan seterusnya, TF-IDF dihitung dengan mengalikan TF (frekuensi munculnya n-gram dalam keseluruhan dataset) dengan IDF (seberapa penting n-gram dalam keseluruhan konteks dataset. Untuk memperhalus skala, digunakan logaritma). Dari nilai hasil perhitungan TF-IDF tersebut, nilai yang tinggi menunjukkan bahwa kata tersebut sering muncul dalam dokumen contoh, namun jarang muncul di dokumen yang lain sehingga kata ini dianggap penting untuk dokumen tersebut. Nilai yang tinggi menandakan kata tersebut memiliki dan membawa informasi penting dalam keseluruhan kalimat. Sedangkan nilai TF-IDF yang rendah menunjukkan bahwa kata tersebut jarang muncul dalam dokumen contoh namun sering muncul di dokumen lain di dalam korpus.

Dalam penelitian ini ditunjukkan bahwa Metode TF-IDF (Term Frekuensi-Inverse Document Frekuensi) dapat memberikan keuntungan dan memungkinkan untuk menetapkan bobot yang sesuai pada n-gram dalam dokumen sehingga dapat diterapkan dalam pembuatan model untuk memprediksi status hasil mediasi. Berdasarkan dokumen – dokumen yang sudah ada, bobot TF-IDF akan menentukan apakah dokumen yang sudah ada dapat dijadikan referensi untuk memilih dan memberi peringkat dokumen berdasarkan relevansinya dengan permintaan pengguna.

4.7 Pengaruh N-gram terhadap nilai bobot TF-IDF

Berdasarkan hasil perhitungan TF-IDF pada sub bab sebelumnya, terdapat beberapa hal yang menunjukkan pengaruh n-gram terhadap TF-IDF yaitu:

- Frekuensi Kata (TF)

  Untuk n = 1 atau unigram, kata atau term dihitung terpisah per unit kata/individu sehingga bobot TF mencerminkan seberapa sering kata tersebut muncul dalam dokumen. Sedangkan untuk n = 2 atau bigram, n = 3 atau trigram dan seterusnya n > 1, cenderung muncul lebih jarang sehingga nilai TF akan lebih tinggi.

- Dokumen yang mengandung term atau token (DF) dan inversenya (IDF)

  Untuk n = 1 kata – kata umum akan cenderung sering muncul dalam dokumen sehingga nilai DFnya menjadi tinggi namun bobot IDFnya menjadi rendah. Sebaliknya, untuk untuk n = 2 atau bigram, n = 3 atau trigram dan seterusnya n > 1, karena urutan kata lebih panjang dan spesifik maka kecenderungannya untuk muncul dalam dokumen rendah, menghasilkan nilai DF rendah namun IDFnya tinggi.

- Pembobotan (TF-IDF)

  Pembobotan TF-IDF diperoleh dengan mengalikan TF dengan IDF sehingga TF-IDF akan cenderung lebih rendah untuk kata – kata umum yang sering muncul dalam dokumen (n = 1 atau unigram). Sedangkan untuk n-gram dengan n > 1 bobot TF-IDF akan menjadi lebih tinggi apabila n-gram tersebut muncul dalam konteks yang lebih spesifik.

Nilai TF-IDF ini selanjutnya akan digunakan dalam pemodelan pada tahapan berikut.

4.8 Pemodelan dengan Algoritma Klasifikasi teks

Dalam penelitian ini digunakan beberapa algoritma klasifikasi teks dimana fitur – fitur TF-IDF yang dihasilkan sebelumnya dilatih dan menggunakan algoritma – algoritma tersebut untuk memprediksi variable hasil, yaitu label pada kolom "status hasil mediasi" berdasarkan variable sumber yaitu kolom petitum. Hasil akhir yang diharapkan adalah keluaran berupa label T untuk prediksi mediasi gagal dan label Y untuk prediksi mediasi berhasil. Untuk melatih model, digunakan data latih yang dibandingkan dengan data hasil perhitungan TF-IDF n-gram pada tahapan sebelumnya. Kedua data ini disandingkan satu sama lain dengan menggunakan nomor perkara atau nomor penetapan mediasi sebagai identittas unik. Kolom 'nomor perkara atau nomor penetapan mediasi' tidak menentukan akurasi pemodelan karena kolom tersebut hanya digunakan untuk menggabungkan data asli dengan data TF-IDF. Setelah penggabungan, kolom 'nomor perkara' tidak digunakan dalam proses pemodelan itu sendiri. Kolom ini walaupun dihilangkan, proses pemodelan dan akurasinya tidak akan terpengaruh, karena fitur-fitur yang digunakan untuk memprediksi 'status hasil mediasi' berasal dari nilai TF-IDF dari n-gram yang berbeda (n=1 sampai n=5). Jika dihapuspun, kolom 'nomor perkara' setelah penggabungan data akan tetap menghasilkan model yang sama dengan akurasi yang sama.

Berikut ini dipaparkan hasil implementasi setiap algoritma klasifikasi teks:

a. Naïve Bayes

Hasil perhitungan TF-IDF untuk n-gram (n=1 sampai n=5) yang telah diproses dalam tahapan sebelumnya akan digunakan untuk inisialisasi dan pelatihan model dengan Naive Bayes, yaitu persamaan 3:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (8$$

Dimana,

X adalah data dengan class yang belum diketahui,

H adalah Hipotesis data X, merupakan suatu class spesifik,

P(H|X) adalah probabilitas hipotesis H berdasar kondisi X (posteriori probability),

P(H) adalah probabilitas hipotesis H (prior probability),

P(X|H) adalah probabilitas X berdasar kondisi pada hipotesis H, dan

P(X) adalah probabilitas dari X.

Langkah – langkah pemodelan dapat dijabarkan sebagai berikut:

1) Probabilitas prior atau P(H) untuk setiap kelas dalam 'status hasil mediasi'yaitu membagi jumlah dokumen untuk setiap kelas dengan total jumlah dokumen. Jumlah dokumen dalam data latih adalah sebanyak 571 dokumen. Adapun jumlah dokumen untuk setiap kelas dalam data latih adalah 552 dokumen untuk status hasil mediasi = "T" dan 19 dokumen untuk status hasil mediasi = "Y"

   Sehingga,

   P(H) [status hasil mediasi = "T"] = 552/571 = 0.996672504

   P(H) [status hasil mediasi = "Y"] = 19/571 = 0.03327496

2) Maka, penggunaan TF-IDF dalam setiap bagiannya:

1. P(X|H) (Probabilitas Likelihood):
    - P(X|H) mengacu pada probabilitas bahwa bukti X (yaitu, n-gram yang diamati dalam dokumen) akan muncul jika hipotesis H (yaitu, kelas dokumen) benar.
    - Dalam konteks TF-IDF, setiap elemen $X=\{t_1, t_2,...,t_n\}$ merupakan sekumpulan n-gram dari dokumen. Untuk setiap n-gram t_i, nilai TF-IDF-nya menunjukkan seberapa penting atau diskriminatif n-gram tersebut dalam konteks keseluruhan korpus.
    - Penggunaan TF-IDF dalam P(X|H) dapat dinyatakan sebagai: $P(X|H) = \prod_{i=1}^{N} P(t_i|H)^{TF-IDF(t_i)}$ di mana,

    $P(t_i|H)$ adalah probabilitas kemunculan n-gram t_i dalam kelas H, dan $TF\text{-}IDF(t_i)$ adalah nilai TF-IDF dari n-gram t_i.

Likelihood dihitung menggunakan bobot TF-IDF. Likelihood adalah probabilitas kemunculan kata – kata (atau n-gram) dalam kelas tertentu. TF-IDF telah dihitung untuk setiap n-gram (disimbolkan dengan w_i)

3) Misalkan, untuk contoh di atas pada baris perkara 310/Pdt.G/2023/PN Tnn pada hasil 1-gram atau unigram, dapat dibuat pemodelan dengan naïve bayes sebagai berikut:

Dari perhitungan sebelumnya untuk P(H):

P(H) [status hasil mediasi = "T"] = 552/571 = 0.996672504

P(H) [status hasil mediasi = "Y"] = 19/571 = 0.03327496

$$P(X|H) = \prod_{i=1}^{N} P(w_i|H)$$

$$P(w_i|H) \approx \frac{TF-IDF\ dari\ w_i\ dalam\ kelas\ H}{\sum_j TF-IDF\ dari\ semua\ n-gram\ dalam\ kelas\ H}$$

Dengan nilai $\sum_j TF - IDF$ dari semua n-gram (n=1 sampai n=5) dalam kelas 'status hasil mediasi' = "T" sedangkan 'status hasil mediasi' = "Y".

Contohnya, TF-IDF untuk unigram perkara tersebut adalah sebagaimana dalam tabel 4.9 sehingga $P(w_i|H)$ atau probabilitas dari unigram dapat dihitung. Selengkapnya nilai TF-IDF seluruh baris untuk kedua kelas tersebut disajikan pada lampiran.

4) Untuk menentukan $P(X|H) = \prod_{i=1}^{N} P(w_i|H)$ diperoleh dengan mengalikan semua P(wi|H) atau probabilitas dari kelas. Dari perhitungan yang dilakukan maka diperoleh P(X|'status hasil mediasi = "T") dan P(X|'status hasil mediasi = "Y").

Dengan demikian, penggunaan TF-IDF dalam Naive Bayes tidak hanya membantu dalam memperkirakan probabilitas kemunculan n-gram dalam konteks kelas (P(X|H)), tetapi juga memperbaiki probabilitas prior (P(H)) dengan mempertimbangkan signifikansi n-gram dalam korpus secara keseluruhan. Klasifikasi dengan Data Uji dilakukan setelah melatih model Naive Bayes, dimana digunakan nilai TF-IDF dari n-gram yang ada dalam data uji (yang juga telah diproses sebelumnya) untuk melakukan prediksi kelas.

Pengolahan Data Uji menggunakan model Naïve Bayes menghasilkan klasifikasi yang memiliki kelas "T" untuk seluruh 246 data uji. Hal ini dapat disebabkan adanya ketidakseimbangan data (data imbalanced).

Untuk menghitung performa model, terlebih dahulu dibuat confusion matriksnya sebagai berikut: Aktual (Y), Prediksi (T) sebagai berikut

	T	Y
T	TN	FP
Y	FN	TP

Dengan

- TP = Jumlah data di mana label aktualnya adalah Y dan model memprediksi Y = 0
- TN: Jumlah data di mana label aktualnya adalah T dan model memprediksi T = 300
- FP: Jumlah data di mana label aktualnya adalah T, namun model memprediksi Y = 0
- FN: Jumlah data di mana label aktualnya adalah Y, namun model memprediksi T = 11

sehingga dengan menggunakan Persamaan 8, 9, 10, 11 sebagai berikut dapat dihitung performa algoritma:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1\text{-}score = \frac{2 * presisi * recall}{presisi + recall} \quad (11)$$

Performa model yang dibuat dengan algoritma Naïve Bayes dihitung berdasarkan hasil TF-IDF untuk setiap n-gram dengan n = 1 sampai n = 5 dan menghasilkan hasil evaluasi performa sebagai berikut:

Untuk TFIDF dari n gram n = 1 atau unigram diperoleh Akurasi Model sebesar 0.96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9619181946403385	0.994169096209 3126	0.977777777777 7777	686
Y	0.2	0.035714285714 28571	0.060606060606 06061	28
'accuracy':	0.9565826330532213			
'macro avg'	0.5809590973201693	0.514941690962 0992	0.519191919191 9192	714
'weighted avg'	0.9320390 49752482	0.956582633053 2213	0.941810259457 3184	714

Untuk TFIDF dari n gram n = 2 atau bigram diperoleh Akurasi Model sebesar 0.96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.960729312762 9734	1.0	0.979971387696 7096	685
Y	0.0	0.0	0.0	28
'accuracy':	0.9607293127629734			

'macro avg'	0.480364656381 4867	0.5	0.489985693848 3548	713
'weighted avg'	0.941487237829 2372	0.9607293 127629734	0.941487237829 2372	713

Untuk TFIDF dari n gram n = 3 atau trigram diperoleh Akurasi Model sebesar 0.96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.960784313725 4902	1.0	0.98	588
Y	0.0	0.0	0.0	24
'accuracy':	0.9607843137254902			
'macro avg'	0.480392156862 7451	0.5	0.49	612
'weighted avg'	0.923106497500 9613	0.9607843 137254902	0.941568627450 9804	612

Untuk TFIDF dari n gram n = 4 atau fourgram diperoleh Akurasi Model sebesar 0.96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.959527824620 5734	1.0	0.979345955249 5697	569
Y	0.0	0.0	0.0	24
'accuracy':	0.9595278246205734			
'macro avg'	0.479763912310 2867	0.5	0.489672977624 78484	593

'weighted avg'	0.920693646221 0897	0.9595278 246205734	0.939709693991 5769	593

Untuk TFIDF dari n gram n = 5 atau fivegram diperoleh Akurasi Model sebesar 0.96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.958415841584 1584	1.0	0.978766430738 1193	484
Y	0.0	0.0	0.0	21
'accuracy':	0.9584158415841584			
'macro avg'	0.479207920792 0792	0.5	0.489383215369 05965	505
'weighted avg'	0.918560925399 4707	0.9584158 415841584	0.938065252430 1976	505

Hasil Evaluasi performa algoritma di atas untuk setiap nilai TF-IDF n-gram dapat dijelaskan dalam analisis berikut:

1. Akurasi Model

Semua file TF-IDF menunjukkan akurasi yang tinggi, sekitar 96%. Namun, akurasi yang tinggi tidak selalu menunjukkan performa model yang baik, terutama dalam kasus data tidak seimbang seperti dataset yang digunakan

2. Precision, Recall, dan F1-Score

Untuk Kelas 'T' atau mediasi gagal sebagai label Mayoritas, nilai Precision, recall, dan F1-score sangat tinggi, mendekati atau mencapai 1.0. Ini menunjukkan bahwa model sangat baik dalam mengklasifikasikan data yang termasuk dalam

kelas mayoritas ('T'). Untuk kelas 'Y' atau mediasi berhasil sebagai label Minoritas, nilai Precision, recall, dan F1-score untuk kelas 'Y' sangat rendah, bahkan 0.0 dalam beberapa kasus. Ini menunjukkan bahwa model sangat buruk dalam mengklasifikasikan data yang termasuk dalam kelas minoritas ('Y').

3. Macro Average dan Weighted Average

Macro average untuk precision, recall, dan F1-score sekitar 0.48-0.50. Ini mengindikasikan bahwa performa model tidak merata antara kelas mayoritas dan minoritas. Nilai macro average yang rendah mencerminkan bahwa model tidak dapat menangani kelas minoritas dengan baik. Sementara, Weighted average lebih tinggi daripada macro average, menunjukkan bahwa nilai ini lebih dipengaruhi oleh kelas mayoritas karena jumlah data yang lebih banyak. Weighted average mendekati akurasi keseluruhan model, karena kelas mayoritas mendominasi.

4. Ketidakseimbangan Data

Ketidakseimbangan data sangat jelas terlihat dari jumlah support untuk kelas 'T' (sekitar 685-484) dan kelas 'Y' (sekitar 28-21). Model tidak cukup untuk belajar dengan baik untuk kelas 'Y' karena jumlah sampel yang sangat sedikit dibandingkan dengan kelas 'T'.

b. Logistic Regression

Dalam penelitian ini, LogisticRegression diimpor dari scikit-learn. Classifier Logistic Regression diinisialisi dan dilatih menggunakan menggunakan fitur TF-IDF. Persamaan yang digunakan adalah persamaan (4:

$$P(y = 1 \mid X) = 1 / (1 + \exp(-z)) \qquad (9$$

Dimana P (y=1 | X) adalah probabilitas dari variabel target pada adalah probabilitas variabel target menjadi 1 dengan prediktor X yang diberikan, z adalah kombinasi linear dari predictor dan koefisiennya z = w0 + w1*X1 + w2*X2 + ... + wn*Xn dan exp () adalah fungsi eksponensial.

Langkah – langkahnya adalah:

1) Ambil hasil perhitungan TF-IDF.
2) Gabungkan hasil perhitungan TF-IDF menjadi satu matriks fitur, dimana hasil TF-IDF berasal dari n-gram dengan n=1 sampai n=5
3) Menginisialisasi bobot dan bias.

Bobot (weights) z = TF-IDF Dokumen

Misalnya, diketahui TF IDF sebagai berikut:

0.091578	0.128313	0.26046	0.102061	0.041897	0.083502
0.032554	0.028914	0.045031	0.148421	0.017057	0.031998
0.028038	0.183155	0.035832	0.059038	0.056843	0.055714
0.055714	0.062398	0.102535	0.038808	0.066263	0.054158
0.113687	0.054659	0.032618	0.050829	0.031284	0.054327
0.036873	0.283379	0.049487	0.097987	0.072545	0.059038

Inisialisasi Bobot dan bias ditetapkan sebagai berikut:

Bobot (panjangnya disesuaikan dengan TF-IDF, dengan memilih nilai nol atau nilai random yang kecil):

0	0.2	-0.1	0.1	0	0.2
-0.2	0.1	0.1	0.1	0.2	-0.1
0.2	0.1	0	0.2	0.1	-0.2
0.1	0.2	0.2	-0.1	0.2	0.1
0	-0.2	-0.1	0.2	0.1	0.2
0.1	-0.1	0.2	-0.2	0.2	0.2

Inisialisasi bias: 0.0

Hitung nilai z

= 0.0 + 0*0.091578 + 0.2*0.128313 + (-0.1*0.26046) + (0.1*0.102061) + (0*0.041897) + (0.2*0.083502) + (-0.2*0.032554) + (0.1*0.028914) + (0.1*0.045031) + (0.1*0.148421) + (0.2*0.017057) + (-0.1*0.031998) + (0.2*0.028038) + (0.1*0.183155) + (0*0.035832) + (0.2*0.059038) + (0.1*0.056843) + (-0.2*0.055714) + (0.1*0.055714) + (0.2*0.062398) + (0.2*0.102535) + (-0.1*0.038808) + (0.2*0.066263) + (0.1*0.054158) + (0*0.113687) + (-0.2*0.054659) + (-0.1*0.032618) + (0.2*0.050829) + (0.1*0.031284) + (0.2*0.054327) + (0.1*0.036873) + (-0.1*0.283379) + (0.2*0.049487) + (-0.2*0.097987) + (0.2*0.072545) + (0.2*0.059038)

= 0 + 0 + 0.0025663 + (-0.02605) + 0.010206 + 0 + 0.0167 + (-0.00651) + 0.002891 + 0.004503 + 0.014842 + 0.003411 + (-0.0032) + (0.005608) + (0.018316) + 0 + 0.011808 + 0.005684 + (-0.01114) + 0.05571 + 0.01248 + 0.020507 + (-0.00388) + 0.013253 + 0.005416 + 0 + (-0.01093) + (-0.00326) + 0.010166 + 0.003128 + 0.010865 + 0.003687 + (-0.08234) + (0.009897) + (-0.0196) + 0.014509 + 0.011808

= 0.12801

4) Hitung probabilitas: P(y=1|X) atau fungsi sigmoid

$\sigma(0.12801)=1+e^{-0.128011} = 0.5319$

untuk nilai probabilitas ini, menunjukkan bahwa data berada pada nilai ambang batas sehingga perbandingan antara kedua kelas adalah sama dan berimbang. Nilai ambang batas 0.5 adalah threshold untuk klasifikasi biner.

5) Hitung error.

Error = 0.5319 – 1 = - 0.4681

Nilai ini adalah sama dengan bias

6) Hitung gradient untuk bobot, dengan python diperoleh

$\frac{\partial L}{\partial m}$ untuk baris pertama adalah

gradients = (–0.0215245–0.0609211–0.1212621–0.0476145–0.020548 –0.038961))

Untuk seluruh perhitungan di atas, dibuat pemodelan dari data latih. Selanjutnya, model tersebut digunakan untuk data uji, sehingga dengan menggunakan Persamaan 8, 9, 10, 11 sebagai berikut dapat dihitung performa algoritma:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1\text{-}score = \frac{2 * presisi * recall}{presisi + recall} \quad (11)$$

Hasil evaluasi model menggunakan Algoritma Logistic Regression adalah sebagai berikut: Untuk TFIDF dari n gram n = 1 atau unigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9970930 23255814	1.0	0.998544395924 3085	686
Y	1.0	0.928571428571 4286	0.962962962962 963	28
'accuracy':	0.9971988795518207			
'macro avg'	0.9985465 11627907	0.964285714285 7143	0.980753679443 6358	714
'weighted avg'	0.9972070 223438212	0.997198879551 8207	0.997149045612 0988	714

Untuk TFIDF dari n gram n = 2 atau bigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422740524781	1.0	0.99927060539752	685
Y	1.0	0.9642857142857143	0.9818181818181818	28
'accuracy':	0.9985974754558204			
'macro avg'	0.9992711370262390	0.9821428571428572	0.9905443936078508	713
'weighted avg'	0.9985995199522405	0.9985974754558204	0.9985852367296076	713

Untuk TFIDF dari n-gram n = 3 atau trigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422740524781	1.0	0.99927060539752	685
Y	1.0	0.9642857142857143	0.9818181818181818	28
'accuracy':	0.9985974754558204			
'macro avg'	0.9992711370262390	0.9821428571428572	0.9905443936078508	713
'weighted avg'	0.9985995199522405	0.9985974754558204	0.9985852367296076	713

Untuk TFIDF dari n-gram n = 4 atau fourgram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422740524781	1.0	0.99927060539752	685
Y	1.0	0.9642857142857143	0.9818181818181818	28
'accuracy':	0.9985974754558204			
'macro avg'	0.9992711370262390	0.9821428571428572	0.9905443936078508	713
'weighted avg'	0.9985995199522405	0.9985974754558204	0.9985852367296076	713

Untuk TFIDF dari n-gram n = 5 atau Fivegram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422740524781	1.0	0.99927060539752	685
Y	1.0	0.9642857142857143	0.9818181818181818	28
'accuracy':	0.9985974754558204			
'macro avg'	0.9992711370262390	0.9821428571428572	0.9905443936078508	713
'weighted avg'	0.9985995199522405	0.9985974754558204	0.9985852367296076	713

Evaluasi Performa pemodelan dengan Algoritma Logistic regression memberikan gambaran sebagai berikut:

1. Accuracy:
    - tfidf_results_1gram.csv: 99.72%
    - tfidf_results_2gram.csv: 99.86%
    - tfidf_results_3gram.csv: 99.86%
    - tfidf_results_4gram.csv: 99.86%
    - tfidf_results_5gram.csv: 99.86%

    Accuracy mengukur seberapa sering model membuat prediksi yang benar. Nilai di atas yang mendekati 100% menunjukkan bahwa model hampir selalu membuat prediksi yang benar.
2. Precision, Recall, dan F1-score. Precision mengukur berapa proporsi prediksi positif yang benar-benar positif. Precision untuk kelas T: Sekitar 0.998 - 0.999. Precision untuk kelas Y: 1.0. Recall mengukur berapa proporsi kasus positif yang terprediksi dengan benar oleh model. Recall untuk kelas T: 1.0. Recall untuk kelas Y: Sekitar 0.964. F1-score adalah rata-rata harmonis dari precision dan recall. F1-score untuk kelas T: Sekitar 0.999. F1-score untuk kelas Y: Sekitar 0.982. Ini menunjukkan bahwa model sangat baik dalam mengklasifikasikan kedua kelas, terutama kelas mayoritas (T).
3. Macro Avg menunjukkan Rata-rata precision, recall, dan F1-score untuk kedua kelas tanpa memperhatikan jumlah sampel di setiap kelas dan

Weighted Avg menunjukkan Rata-rata precision, recall, dan F1-score untuk kedua kelas dengan memperhitungkan jumlah sampel di setiap kelas. Macro avg dan weighted avg menunjukkan bahwa model memiliki performa yang konsisten di seluruh kelas.

c. **Decision tree**

DecisionTreeClassifier diimpor dari scikit-learn. Classifier Decision Tree diinisialisasi dan dilatih menggunakan karakteristik TF-IDF. Selanjutnya diprediksi label yang akan muncul pada kolom 'status hasil mediasi'.

Langkah – langkah yang dilakukan adalah:

1. Mengumpulkan TFIDF dalam satu tabel. TFIDF dihitung untuk n-gram dengan n=1 sampai n=5.

   Misalkan, diketahui baris – baris perkara sebagaimana tabel 4.10 berikut:

   Tabel 4.10. Contoh baris perkara dengan perhitungan TFIDF

nomor perkara	status hasil mediasi	TFIDF 1-gram	TFIDF 2-gram	TFIDF 3-gram	TFIDF 4-gram	TFIDF 5-gram
193/Pdt.G/2021/PN Amr	T	0.224921	0.467174	0.529631	0.508059	0.491006
135/Pdt.G/2022/PN Ktg	Y	0.050897	0.566896	0.833574	0.071077	0.393303
441/Pdt.G/2023/PN Mnd	T	0.433896	0.017374	0.473828	0.071077	0.078775
368/Pdt.G/2021/PN Tnn	T	0.2101	0.172386	0.219242	0.669679	0.501414

2. Hitung Entropi awal menggunakan persamaan 5

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \qquad (5$$

Dari tabel baris perkara di atas diperoleh

Status Hasil Mediasi	Jumlah Dokumen
T	3
Y	1

$H(S) = -(\frac{1}{4} log_2 \frac{1}{4} + \frac{3}{4} log_2 \frac{3}{4})$

$log_2 \frac{1}{4} = log_2 (1) - log_2 (4) = 0 - 2 = -2$

$log_2 \frac{3}{4} = log_2 (3) - log_2 (4) = log_2 (3) - 2$

$H(S) = -(\frac{1}{4} * (-2) + \frac{3}{4} * log_2 (3) - 2)$

$= -(-\frac{2}{4} + \frac{3}{4} * log_2 (3) - 2)$

$= -(-\frac{1}{2} + \frac{3}{4} * log_2 (3) - \frac{3}{2})$

$= -(-\frac{1}{2} - \frac{3}{2} + \frac{3}{4} * log_2 (3))$

$= -(-2 + \frac{3}{4} * log_2 (3))$

$= 2 - \frac{3}{4} * log_2 3$

$= 2 - \frac{3}{4} * 1.58496$

$= 2 - 1.18872 = 0.81128$

3. Tentukan threshold untuk split setiap nilai TFIDF dari n = 1 sampai n = 5.

   Gunakan threshold 0.2 untuk split:

   - Subset 1: TF-IDF 1-gram < 0.2
   - Subset 2: TF-IDF 1-gram ≥ 0.2

   Split data:

   - Subset 1: 135/Pdt.G/2022/PN Ktg

- Subset 2: 193/Pdt.G/2021/PN Amr, 441/Pdt.G/2023/PN Mnd, 368/Pdt.G/2021/PN Tnn

4. Hitung entropi masing-masing subset:
   - *H*(Subset 1): hanya memiliki satu kelas (Y), jadi entropi = 0
   - *H*(Subset 2): hanya memiliki satu kelas (T), jadi entropi = 0
5. Hitung Information Gain dihitung dengan menggunakan persamaan 6 berikut:

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (6)$$

   Information Gain untuk n=1 = 0.81125

6. Menyusun Pohon Keputusan (Decision Tree):

```
        [TF-IDF 1-gram < 0.2]
        /                \
  status Y        [TF-IDF 1-gram ≥ 0.2]
                  /                \
            status T    [Cek fitur berikutnya]
```

   Tentukan fitur (nilai TF-IDF) dan label untuk setiap dokumen.

Misalkan kita membuat aturan sederhana seperti berikut:

1. Jika TFIDF 1-gram > 0.3, maka status hasil mediasi adalah 'T'.
2. Jika TFIDF 1-gram <= 0.3 dan TFIDF 2-gram > 0.5, maka status hasil mediasi adalah 'Y'.

3. Jika TFIDF 1-gram <= 0.3 dan TFIDF 2-gram <= 0.5, maka status hasil mediasi adalah 'T'.

Klasifikasi status hasil mediasi berdasarkan aturan di atas:

1. 193/Pdt.G/2021/PN Amr: TFIDF 1-gram = 0.224921 -> TFIDF 2-gram = 0.467174 -> Prediksi: 'T'
2. 135/Pdt.G/2022/PN Ktg: TFIDF 1-gram = 0.050897 -> TFIDF 2-gram = 0.566896 -> Prediksi: 'Y'
3. 441/Pdt.G/2023/PN Mnd: TFIDF 1-gram = 0.433896 -> Prediksi: 'T'
4. 368/Pdt.G/2021/PN Tnn: TFIDF 1-gram = 0.210100 -> TFIDF 2-gram = 0.172386 -> Prediksi: 'T'

7. Klasifikasi Dokumen Baru

Misalkan kita memiliki dokumen baru dengan nilai TF-IDF 1-gram = 0.15, maka cek nilai TF-IDF pada aturan di atas, sehingga prediksi kelas adalah Y. Model Decision Tree kemudian diimplementasikan pada data uji sehingga dengan menggunakan Persamaan 8, 9, 10, 11 dapat dihitung performa algoritma sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Akurasi = 0.946073194702882

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

Presisi = 0.9098361542820413

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

Recall = 0.946073194702882

$$F1\text{-}score = \frac{2 * presisi * recall}{presisi + recall} \quad (11)$$

F1-score = 0.927120709674359

Hasil Pemodelan dari TF-IDF dengan menggunakan algoritma Decision Tree menghasilkan performa sebagai berikut:

Untuk TFIDF dari n-gram n = 1 atau unigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985443 959243085	1.0	0.999271667880 5535	686
Y	1.0	0.964285714285 7143	0.981818181818 1818	28
'accuracy': 0.9985994397759104				
'macro avg'	0.9992721 979621543	0.982142857142 8572	0.990544924849 3676	714
'weighted avg'	0.9986014 784370808	0.998599439775 9104	0.998587217446 7351	714

Untuk TFIDF dari n-gram n = 2 atau bigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422 740524781	1.0	0.999270605397 52	685
Y	1.0	0.964285714285 7143	0.981818181818 1818	28
'accuracy': 0.9985974754558204				

'macro avg'	0.9992711 37026239	0.982142857142 8572	0.990544393607 8508	713
'weighted avg'	0.9985995 199522405	0.998597475455 8204	0.998585236729 6076	713

Untuk TFIDF dari n-gram n = 3 atau trigram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422 740524781	1.0	0.999270605397 52	685
Y	1.0	0.964285714285 7143	0.981818181818 1818	28
'accuracy': 0.9985974754558204				
'macro avg'	0.9992711 37026239	0.982142857142 8572	0.990544393607 8508	713
'weighted avg'	0.9985995 199522405	0.998597475455 8204	0.998585236729 6076	713

Untuk TFIDF dari n-gram n = 4 atau fourgram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422 740524781	1.0	0.999270605397 52	685
Y	1.0	0.964285714285 7143	0.981818181818 1818	28
'accuracy': 0.9985974754558204				

'macro avg'	0.9992711 37026239	0.982142857142 8572	0.990544393607 8508	713
'weighted avg'	0.9985995 199522405	0.998597475455 8204	0.998585236729 6076	713

Untuk TFIDF dari n-gram n = 5 atau fivegram diperoleh Akurasi Model sebesar 100%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9985422 740524781	1.0	0.999270605397 52	685
Y	1.0	0.964285714285 7143	0.981818181818 1818	28
'accuracy':	0.9985974754558204			
'macro avg'	0.9992711 37026239	0.982142857142 8572	0.990544393607 8508	713
'weighted avg'	0.9985995 199522405	0.998597475455 8204	0.998585236729 6076	713

Dari hasil evaluasi model Decision Tree yang dilatih pada lima skenario TF-IDF, berikut adalah analisis performa model: Model dievaluasi berdasarkan empat metrik utama yaitu Akurasi terkait proporsi prediksi yang benar dari keseluruhan prediksi, Precision yaitu proporsi prediksi positif yang benar, Recall yaitu proporsi kasus positif yang terdeteksi oleh model dan F1-Score yaitu Rata-rata harmonis dari precision dan recall. Model Decision Tree menunjukkan performa yang sangat baik dengan akurasi mendekati 100% untuk semua skenario TF-IDF. Precision dan

recall untuk kategori 'T' hampir sempurna, sementara untuk kategori 'Y', recall sedikit lebih rendah tetapi tetap dalam kisaran tinggi (~0.964), menunjukkan bahwa model ini sangat efisien dalam mendeteksi kelas minoritas dengan akurasi tinggi. Metrik performa menunjukkan ketidakseimbangan kelas dengan jumlah 'T' yang jauh lebih banyak dibandingkan 'Y' sehingga recall untuk 'Y' sedikit lebih rendah. Performansi yang hampir identik di semua file menunjukkan bahwa model secara konsisten dapat menangani berbagai jenis n-gram dengan efektif.

d. Support Vector Machine (SVM)

Dengan memaksimalkan karakteristik TF-IDF, classifier SVM diinisialisasi dan dilatih untuk memprediksi label yang dihasilkan pada kolom 'status hasil mediasi'. Misalkan diketahui data sebagaimanan Tabel 4.10 yang telah digunakan pada contoh implementasi algoritma sebelumnya.

Langkah – langkah yang dilakukan adalah sebagai berikut:

1) **Persiapkan Dataset dalam Format yang Sesuai**:

Dataset menggunakan data latih dengan hasil TF-IDF telah dihitung sebelumnya dengan mempertimbangkan n-gram pada kalimat, mulai dari n = 1 sampai n = 5.

2) **Konversi Label Kategori menjadi Nilai Numerik**:

Label kategori (T dan Y) harus diubah menjadi nilai numerik (misalnya, T = 0 dan Y = 1). Untuk contoh data pada tabel 4.10 di atas, hasil perhitungan TF-IDFnya adalah sebagai berikut:

x = [[0.224921, 0.467174, 0.529631, 0.508059, 0.491006],

[0.050897, 0.566896, 0.833574, 0.071077, 0.393303],

[0.433896, 0.017374, 0.473828, 0.071077, 0.078775],

[0.2101, 0.172386, 0.219242, 0.669679, 0.501414]]

y = [0, 1, 0, 0]

3) **Latih Model SVM**:

Misalkan w=[0.5,–0.1,0.3,0.2,0.1] dan b=–0.05 maka untuk dokumen dengan TFIDF:

X = [0.224921,0.467174,0.529631,0.508059,0.491006]

Perhitungan prediksi dapat menggunakan persamaan 7 berikut untuk melatih model dengan dataset yang diberikan.

$$f(x) = sign(w \cdot x + b) \quad (7)$$

f(x) = (0.5∗0.224921) + (–0.1∗0.467174) + (0.3∗0.529631) + (0.2∗0.508059) + (0.1∗0.491006) – 0.05

Sehingga dengan menggunakan Persamaan 8, 9, 10, 11 sebagai berikut dapat dihitung performa algoritma:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1\text{-}score = \frac{2 * presisi * recall}{presisi + recall} \quad (11)$$

Hasil pemodelan yang dibuat dari TFIDF dengan menggunakan algoritma Support Vector Machine (SVM) menunjukkan evaluasi performa sebagai berikut:

Untuk TFIDF dari n-gram n = 1 atau unigram diperoleh Akurasi Model sebesar 96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9607843137254902	1.0	0.98	686
Y	0.0	0.0	0.0	28
'accuracy':	0.9607843137254902			
'macro avg'	0.4803921568627451	0.5	0.49	714
'weighted avg'	0.9231064975009612	0.9607843137254902	0.9415686274509804	714

Untuk TFIDF dari n-gram n = 2 atau bigram diperoleh Akurasi Model sebesar 96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9607293127629734	1.0	0.979971387696 7096	685
Y	0.0	0.0	0.0	28
'accuracy':	0.9607293127629734			

'macro avg'	0.4803646563814867	0.5	0.4899856938483548	713
'weighted avg'	0.9230008124020151	0.9607293127629734	0.9414872378292372	713

Untuk TFIDF dari n-gram n = 3 atau trigram diperoleh Akurasi Model sebesar 96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9607293127629734	1.0	0.9799713876967096	685
Y	0.0	0.0	0.0	28
'accuracy':	0.9607293127629734			
'macro avg'	0.4803646563814867	0.5	0.4899856938483548	713
'weighted avg'	0.9230008124020151	0.9607293127629734	0.9414872378292372	713

Untuk TFIDF dari n-gram n = 4 atau fourgram diperoleh Akurasi Model sebesar 96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9607293127629734	1.0	0.9799713876967096	685
Y	0.0	0.0	0.0	28
'accuracy':	0.9607293127629734			
'macro avg'	0.4803646563814867	0.5	0.4899856938483548	713

'weighted avg'	0.9230008124020151	0.960729312762 9734	0.941487237829 2372	713

Untuk TFIDF dari n-gram n = 5 atau fivegram diperoleh Akurasi Model sebesar 96%. Evaluasi Performa Model berdasarkan label status hasil mediasi:

	precision	recall	'f1-score	support
T	0.9607293127629734	1.0	0.979971387696 7096	685
Y	0.0	0.0	0.0	28
'accuracy': 0.9607293127629734				
'macro avg'	0.4803646563814867	0.5	0.489985693848 3548	713
'weighted avg'	0.9230008124020151	0.960729312762 9734	0.941487237829 2372	713

Berdasarkan gambaran di atas maka Analisis Hasil Performa Model SVM dapt dijelaskan sebagai berikut:

1. Akurasi Model untuk semua file TF-IDF adalah 0.96%, yang berarti model berhasil mengklasifikasikan sekitar 96% dari semua kasus dengan benar. Meskipun akurasi tinggi, namun akurasi saja bukanlah metrik yang cukup untuk mengevaluasi performa model secara keseluruhan, terutama dalam kasus di mana terdapat ketidakseimbangan kelas.
2. Precision, Recall, dan F1-Score. Precision untuk kelas 'T' sangat tinggi, sekitar 0.96, menunjukkan bahwa ketika model memprediksi 'T', kemungkinan besar prediksi tersebut benar. Recall untuk kelas 'T' juga

tinggi, yaitu 1.0, yang berarti model berhasil menemukan semua instance dari kelas 'T'. Namun, precision dan recall untuk kelas 'Y' adalah 0.0, yang menunjukkan bahwa model tidak berhasil mengidentifikasi instance dari kelas 'Y'. Ini dapat terjadi karena ketidakseimbangan kelas yang ekstrem. F1-Score untuk kelas 'T' adalah sekitar 0.98, menunjukkan keseimbangan yang baik antara precision dan recall untuk kelas ini. F1-Score untuk kelas 'Y' adalah 0.0. mencerminkan kegagalan model untuk menangkap instance dari kelas ini.

3. Macro Average menunjukkan nilai rata-rata untuk precision, recall, dan f1-score tanpa mempertimbangkan ketidakseimbangan kelas. Precision dan recall rata-rata adalah 0.48 dan 0.5 masing-masing, sedangkan f1-score rata-rata adalah 0.49. Weighted Average menunjukkan nilai rata-rata tertimbang untuk precision, recall, dan f1-score, yang memperhitungkan kontribusi dari setiap kelas berdasarkan jumlah instance dalam kelas tersebut. Precision tertimbang adalah sekitar 0.92, recall tertimbang adalah sekitar 0.96, dan f1-score tertimbang adalah sekitar 0.94.
4. Ketidakseimbangan Kelas. Kelas 'T' (banyak instance) mendominasi data, sehingga model cenderung untuk memprediksi 'T' dengan akurasi tinggi tetapi gagal dalam memprediksi kelas 'Y' (sedikit instance).
5. Evaluasi Kelas 'Y'. Model tidak berhasil mengidentifikasi instance dari kelas 'Y', yang mengindikasikan
6. Akurasi Tidak Cukup terutama dalam mengenali kelas minoritas 'Y'.

4.9 Komparasi algoritma

Terdapat 4 kategori yang digunakan untuk perbandingan performa pemodelan algoritma yaitu:

TF-IDF n-gram n = 1, Kelas T

TF-IDF n-gram n = 1, Kelas Y

TF-IDF n-gram n > 1, Kelas T

TF-IDF n-gram n > 1, Kelas Y

Dengan metrik perfoma yang dibandingkan adalah Akurasi, Presisi, Recall dan F1-Score. Maka berdasarkan hasil pemodelan pada bagian sebelumnya, dapat diperoleh hasil perbandingan akurasi dari setiap algoritma sebagaimana ditunjukkan dalam tabel 4.11 sampai 4.14 di bawah ini.

Tabel 4.11. Perbandingan Akurasi algoritma TF-IDF n-gram n = 1, Kelas T

TF-IDF n-gram	Kelas	Algoritma	Akurasi	Presisi	Recall	F1 Score
n = 1 atau unigram	T	Naive Bayes (NB)	0.9565826330532213	0.9619181946403385	0.9941690962099126	0.9777777777777777
		Logistic Regression (LR)	0.9871988795518207	0.9870930232558140	1.0	0.9985443959243085
		Decision tree (DT)	0.9985994397759104	0.9985443959243085	1.0	0.9992716678805535
		Support Vector Machine (SVM)	0.9607843137254902	0.9607843137254902	1.0	0.98

Tabel 4.12. Perbandingan Akurasi algoritma TF-IDF n-gram n = 1, Kelas Y

TF-IDF n-gram	Kelas	Algoritma	Akurasi	Presisi	Recall	F1 Score
n = 1 atau unigram	Y	Naive Bayes (NB)	0.9565826330532213	0.2	0.03571428571428571	0.06060606060606061
		Logistic Regression (LR)	0.9985974754558204	1.0	0.9285714285714286	0.962962962962963
		Decision tree (DT)	0.9985994397759104	1.0	0.9642857142857143	0.9818181818181818
		Support Vector Machine (SVM)	0.9607843137254902	0.0	0.0	0.0

Tabel 4.13. Perbandingan Akurasi algoritma TF-IDF n-gram n > 1, Kelas T

TF-IDF n-gram	Kelas	Algoritma	Akurasi	Presisi	Recall	F1 Score
n > 1	T	Naïve Bayes (NB)	0.9584158415841584	0.9584158415841584	1.0	0.9787664307381193
		Logistic Regression (LR)	0.9985974754558204	0.9985422740524781	1.0	0.99927060539752
		Decision tree (DT)	0.9985974754558204	0.9985422740524781	1.0	0.99927060539752
		Support Vector Machine (SVM)	0.9607293127629734	0.9607293127629734	1.0	0.9799713876967096

Tabel 4.14. Perbandingan Akurasi algoritma TF-IDF n-gram n > 1, Kelas Y

TF-IDF n-gram	Kelas	Algoritma	Akurasi	Presisi	Recall	F1 Score
n > 1	Y	Naïve Bayes (NB)	0.9565826330532213	0.2	0.03571428571428571	0.06060606060606061
		Logistic Regression (LR)	0.9985974754558204	1.0	0.9642857142857143	0.9818181818181818
		Decision tree (DT)	0.9985974754558204	1.0	0.9642857142857143	0.9818181818181818
		Support Vector Machine (SVM)	0.9607293127629734	0.0	0.0	0.0

Dari tabel - tabel di atas, dapat dianalisis perbandingan akurasi dan metrik kinerja untuk berbagai algoritma klasifikasi dengan menggunakan fitur TF-IDF n-gram pada dua kelas, yaitu Kelas T dan Kelas Y sebagai berikut.

1) Kelas T (n = 1 atau unigram)

- Naïve Bayes (NB). Akurasi: 95.66%. Presisi: 96.19%, Recall: 99.42%, F1 Score: 97.78%. Naïve Bayes memiliki akurasi yang baik, namun tidak sebaik algoritma lainnya. Presisi dan recall-nya juga sangat baik, tetapi F1 Score-nya sedikit lebih rendah dibandingkan dengan algoritma lain.
- Logistic Regression (LR). Akurasi: 99.72%. Presisi: 99.71%. Recall: 100%. F1 Score: 99.85%. Analisis: Logistic Regression memiliki performa terbaik dengan akurasi dan F1 Score tertinggi, menunjukkan keseimbangan yang sangat baik antara presisi dan recall.
- Decision Tree (DT). Akurasi: 99.86%. Presisi: 99.85%. Recall: 100%. F1 Score: 99.93%. Decision Tree juga menunjukkan performa yang sangat baik, hampir identik dengan Logistic Regression dalam hal akurasi, presisi, recall, dan F1 Score.
- Support Vector Machine (SVM). Akurasi: 96.08%. Presisi: 96.08%. Recall: 100%. F1 Score: 98.00%. Analisis: SVM menunjukkan performa yang sangat baik dalam hal recall, tetapi akurasinya sedikit lebih rendah dibandingkan dengan Logistic Regression dan Decision Tree. F1 Score-nya juga tinggi.

2) Kelas Y (n = 1 atau unigram)

- Naïve Bayes (NB). Akurasi: 95.66%. Presisi: 20%. Recall: 3.57%. F1 Score: 6.06%. Naïve Bayes memiliki akurasi yang sama dengan Kelas T, tetapi metrik lainnya sangat buruk, menunjukkan bahwa model ini tidak efektif dalam mengklasifikasikan Kelas Y.
- Logistic Regression (LR). Akurasi: 99.86%. Presisi: 100%. Recall: 92.86%. F1 Score: 96.30%. Logistic Regression menunjukkan performa terbaik untuk Kelas Y dengan akurasi dan presisi yang sangat tinggi, serta recall yang juga sangat baik. F1 Score-nya juga tinggi.
- Decision Tree (DT). Akurasi: 99.86%. Presisi: 100%. Recall: 96.43%. F1 Score: 98.18%. Decision Tree juga sangat baik untuk Kelas Y dengan performa hampir identik dengan Logistic Regression dalam hal akurasi dan F1 Score.
- Support Vector Machine (SVM). Akurasi: 96.07%. Presisi: 0%. Recall: 0%. F1 Score: 0%. SVM gagal total dalam mengklasifikasikan Kelas Y, menunjukkan hasil yang sangat buruk di semua metrik.

3) Kelas T (n > 1)

- Naïve Bayes (NB). Akurasi: 95.84%. Presisi: 95.84%. Recall: 100%. F1 Score: 97.88%. Naïve Bayes mengalami penurunan akurasi sedikit dibandingkan dengan unigram, namun recall tetap tinggi. F1 Score menunjukkan peningkatan.

- Logistic Regression (LR). Akurasi: 99.86%. Presisi: 99.85%. Recall: 100%. F1 Score: 99.93%. Logistic Regression tetap menunjukkan performa terbaik dengan metrik yang hampir identik dengan Kelas T untuk unigram.
- Decision Tree (DT). Akurasi: 99.86%. Presisi: 99.85%. Recall: 100%. F1 Score: 99.93%. Decision Tree mempertahankan performa yang sangat baik mirip dengan Logistic Regression.
- Support Vector Machine (SVM). Akurasi: 96.07%. Presisi: 96.07%. Recall: 100%. F1 Score: 98.00%. SVM menunjukkan performa yang sedikit menurun tetapi tetap baik dengan recall tinggi.

4) Kelas Y (n > 1)

- Naïve Bayes (NB). Akurasi: 95.66%. Presisi: 20%. Recall: 3.57%. F1 Score: 6.06%. Naïve Bayes menunjukkan performa yang buruk dalam klasifikasi Kelas Y, mirip dengan hasil pada unigram.
- Logistic Regression (LR). Akurasi: 99.86%. Presisi: 100%. Recall: 96.43%. F1 Score: 98.18%. Logistic Regression menunjukkan performa terbaik dengan akurasi dan F1 Score yang tinggi, mirip dengan hasil pada unigram.
- Decision Tree (DT). Akurasi: 99.86%. Presisi: 100%. Recall: 96.43%. F1 Score: 98.18%. Decision Tree menunjukkan performa yang sangat baik dengan metrik yang sama dengan Logistic Regression.
- Support Vector Machine (SVM). Akurasi: 96.07%. Presisi: 0%. Recall: 0%. F1 Score: 0%. SVM tetap gagal dalam mengklasifikasikan Kelas Y pada n > 1, menunjukkan bahwa metode ini tidak cocok untuk kasus ini.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan penelitian yang dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Hasil mediasi terdahulu dapat digunakan untuk memprediksi hasil mediasi pada perkara baru. Prediksi hasil dilakukan dengan pemodelan menggunakan kolom petitum sebagai variabel sumber, dimana petitum berisi gugatan dari penggugat untuk memohon kepada pengadilan (Majelis Hakim) agar gugatannya dikabulkan. Hasil prediksi variabel target yang ditunjukkan dengan label Y untuk mediasi berhasil dan T untuk mediasi gagal pada kolom Status Hasil Mediasi dipengaruhi oleh isi petitum. Format isi petitum yang tidak seragam untuk setiap baris perkara akan menghasilkan nilai fitur yang berbeda – beda.
2. Hasil ekstraksi fitur dengan n-gram menunjukkan bahwa untuk n = 1 sampai dengan n = 5 memiliki pengaruh pada performa model yang dihasilkan. Namun, untuk n = 1 atau unigram pengaruhnya kurang optimal karena tidak dapat menangkap konteks yang sebenarnya apabila dibandingkan dengan n = 2 sampai n = 5.
3. Algoritma yang digunakan untuk klasifikasi teks dengan n-gram, menghasilkan tingkat akurasi yang berbeda – beda. Untuk Kelas T, Logistic Regression dan Decision Tree menunjukkan performa terbaik

dengan akurasi sangat tinggi dan metrik kinerja yang seimbang. Untuk Kelas Y, Logistic Regression dan Decision Tree juga menunjukkan performa yang sangat baik pada n-gram lebih dari 1, sementara Naïve Bayes dan SVM gagal dalam klasifikasi ini pada n-gram unigram maupun n > 1.

Terkait penggunaan n-gram, penggunaan n > 1 sedikit meningkatkan performa terutama pada Kelas T, tetapi tidak memberikan keuntungan signifikan untuk Kelas Y di Naïve Bayes dan SVM. Logistic Regression dan Decision Tree tetap menunjukkan performa terbaik secara konsisten pada semua konfigurasi n-gram.

5.2 Saran

Berdasarkan simpulan hasil penelitian, maka untuk penelitian lebih lanjut, disarankan hal – hal sebagai berikut:

- Gunakan dataset yang berimbang. Penyeimbangan dataset dapat dilakukan dengan memilih secara seksama dataset yang dapat digunakan dalam kelas status hasil mediasi "T" maupun "Y", misalnya dengan teknik oversampling maupun undersampling.
- Pada tahapan Preprocessing proses stemming tidak berhasil memisah kata bersambung. Untuk hal ini dapat diteliti lebih lanjut dikarenakan belum ada penelitian terbaru terkait pemisahan kata bersambung yang mengekstrak beberapa akar kata sehingga dapat menambah perbendaharaan kamus kata Library Sastrawi.

- Selanjutnya, untuk implementasi pembentukan n-gram perlu diteliti lebih lanjut maksimal nilai n yang dapat digunakan agar hasil pemodelan dapat memenuhi kriteria yang diinginkan. Algoritma klasifikasi yang berbeda selain yang digunakan dalam penelitian ini dapat pula diuji apakah menghasilkan performa yang lebih baik terutama untuk optimalisasi penentuan target dan fitur.

Untuk instansi, Penulis merekomendasikan hal – hal sebagai berikut:

- Agar dibuat format baku atau template untuk petitum sehingga dapat dihindari pengetikan yang tidak sesuai kaidah Bahasa Indonesia, termasuk typo, penggunaan spasi berlebih dan penggunaan symbol – symbol seperti dalam pengetikan dengan mesin ketik.
- Template petitum dibuat berdasdarkan klasifikasi perkara untuk memudahkan prediksi hasil mediasi
- Dibuat aturan terkait implementasi prediksi hasil mediasi berbasis NLP sebagai pengembangan aturan mediasi elektronik.

DAFTAR PUSTAKA

PUSTAKA BUKU

Garcia, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining. Intelligent Systems Reference Library. 2015* (Vol. 72). Springer International Publishing. http://www.springer.com/series/8578

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining* (T. A. Prabawati (ed.); I). ANDI OFFSET YOGYAKARTA.

Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Adrian, L. (2021). The new Normal: Online Dispute Resolution and Online Mediation. In *"Mediation Moves, Wolfgang Metzner Verlag GmbH (2021).* (pp. 175–193). https://ssrn.com/abstract=3891598

Alcántara Francia, O. A., Nunez-del-Prado, M., & Alatrista-Salas, H. (2022). Survey of Text Mining Techniques Applied to Judicial Decisions Prediction. *Applied Sciences (Switzerland)*, *12*(20). https://doi.org/10.3390/app122010200

Alghazzawi, D., Bamasag, O., Albeshri, A., Sana, I., & Ullah, H. (2022). Efficient Prediction of Court Judgments Using an LSTM + CNN Neural Network Model with an Optimal Feature Set. *Mathematics - MDPI*, *10*(5), 683. https://doi.org/https://doi.org/10.2290/math10050683

Avasthi, S., Chauhan, R., & Acharjya., and D. P. (2021). "Processing large text corpus using N-gram language modeling and smoothing." *Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI*. https://doi.org/DOI https://doi.org/10.1007/978-981-15-9689-6_3

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology*

Trends, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Cohen, M. C., Dahan, S., Rule, C., & Branting, L. K. (2022). Conflict Analytics: When Data Science Meets Dispute Resolution. *Manag Business Rev 2.2*, 86–93.

El Jelali, S., Fersini, E., & Messina, E. (2015). Legal retrieval as support to eMediation: matching disputant's case and court decisions. *Artificial Intelligence and Law*, *23*(1), 1–22. https://doi.org/10.1007/s10506-015-9162-1

Ferrario, A., & Naegelin, M. (2020). The Art of Natural Language Processing: Classical, Modern and Contemporary Approaches to Text Document Classification. *SSRN Electronic Journal*, 1–51. https://doi.org/10.2139/ssrn.3547887

Garcia, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining. Intelligent Systems Reference Library. 2015* (Vol. 10). Springer International Publishing. http://www.springer.com/series/8578

Georgieva-Trifonova, T., & Duraku, M. (2021). Research on N-grams feature selection methods for text classification. *IOP Conference Series: Materials Science and Engineering*, *1031*(1). https://doi.org/10.1088/1757-899X/1031/1/012048

Hsieh, H., Jiang, J., Yang, T.-H., Hu, R., & Wu, C.-L. (2022). Predicting the Success of Mediation Requests Using Case Properties and Textual Information for Reducing the Burden on the Court. *Digital Government: Research and Practice*, *2*(4), 1–18. https://doi.org/10.1145/3469233

Khoirunnisa, F., Yusliani, N., Rodiah, D., Bachelor, R., & Ilir, O. (2020). Effect of N-Gram on Document Classification on the Naïve Bayes Classifier Algorithm. In *Sriwijaya Journal of Informatic and Applications* (Vol. 01, Issue 01). https://doi.org/https://doi.org/10.36706/sjia.v1i1.13

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are n-gram categories helpful in text classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12138 LNCS*, 524–537. https://doi.org/10.1007/978-3-030-50417-5_39

Locke, D., & Zuccon, G. (2022). Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. *ArXiv Preprint ArXiv:2202.07209*, *1*(1). http://arxiv.org/abs/2202.07209

Lumbantoruan, P., Mawuntu, R., Waha, C. J. J., & Tangkere, C. (2021). E-Mediation in E-Litigation Stages in Court. *Journal of Law, Policy and Organization*, *108*, 66. https://doi.org/10.7176/JLPG/108-0

Mandal, A., Ghosh, K., Ghosh, S., & Mandal, S. (2022). A sequence labeling model for catchphrase identification from legal case documents. *Artificial Intelligence and Law*, *30*(3), 325–358. https://doi.org/10.1007/s10506-021-09296-2

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*, 237–266. https://doi.org/https://doi.org/10.1007/s10506-019-09255-y

Nurhidayat, R., & Dewi, K. E. (2023). Penerapan Algoritma K-Nearest Neighbor Dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *12*(1), 91–100. https://doi.org/10.34010/komputa.v12i1.9458

Park, S. H., Lee, D. G., Park, J. S., & Kim, J. W. (2021). A survey of research on data analytics-based legal tech. *Sustainability (Switzerland)*, *13*(14). https://doi.org/10.3390/su13148085

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 0–6. https://doi.org/10.1088/1757-899X/874/1/012017

Sengupta, S., & Dave, V. (2022). Predicting applicable law sections from judicial case reports using legislative text analysis with machine learning. *Journal of Computational Social Science*, *5*(1), 503–516. https://doi.org/10.1007/s42001-

021-00135-7

Shaikh, R. A., Sahu, T. P., & Anand, V. (2020). Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers. *Procedia Computer Science*, *167*, 2393–2402. https://doi.org/10.1016/j.procs.2020.03.292

Strickson, B., & De La Iglesia, B. (2020). Legal Judgement Prediction for UK Courts. *ACM International Conference Proceeding Series*, 204–209. https://doi.org/10.1145/3388176.3388183

Sueno, H. T., Gerardo, B. D., & Medina, R. P. (2020). Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique. *International Journal of Advanced Trends in Computer Science and Engineeringfile*, *9*(3), 3937–3944. https://doi.org/10.30534/ijatcse/2020/216932020

Sullivan, C. O., & Beel, J. (2019). Predicting the Outcome of Judicial Decisions made by the European Court of Human Rights. *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Scienc*. https://doi.org/https://doi.org/10.48550/arXiv.1912.10819

Tabassum, A., & Patil, R. R. (2020). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology*, *07*(06), 4865–4867. www.irjet.net

Vernanda, Y., Hansun, S., & Kristanda, M. B. (2020). Indonesian language email spam detection using n-gram and naïve bayes algorithm. *Bulletin of Electrical Engineering and Informatics*, *9*(5), 2012–2019. https://doi.org/10.11591/eei.v9i5.2444

Zadgaonkar, A. V., & Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical and Computer Engineering*, *11*(6), 5450–5457. https://doi.org/10.11591/ijece.v11i6.pp5450-5457

Zeleznikow, J. (2021). Using Artificial Intelligence to provide Intelligent Dispute Resolution Support. *Group Decision and Negotiation*, *30*(4), 789–812. https://doi.org/10.1007/s10726-021-09734-1

LAMPIRAN (TERSEDIA DALAM LINK : BIT.LY/CSV_KOMPARASI)

1 File datasetcollab_awal.csv

2 File 1_filtered_dataset.csv:

3 File 2_dataset_labelubah.csv

4 FILE 3_preprocessed_casefolding.csv

5 FILE 5_preprocessed_tokenized_nltk.csv

6 File long words.txt

7 FILE 6_preprocessed_no_long_words.csv

8 File 7_preprocessed_stopwords_removed.csv

9 File 8_data_latih.csv

10 File 9_data_uji.csv

11 File 10_data_latih_with_ngram.csv

12 FILE 11_data_uji_with_ngram.csv

13 File tfidf_results_1gram.csv

14 File tfidf_results_2gram.csv

15 File tfidf_results_3gram.csv

16 File tfidf_results_4gram.csv

17 File tfidf_results_5gram.csv

18 File tfidf_results_1gram_UJI.csv

19 File tfidf_results_2gram_UJI.csv

20 File tfidf_results_3gram_UJI.csv

21 File tfidf_results_4gram_UJI.csv

22 File tfidf_results_5gram_UJI.csv

23 File hasil_prediksi_NB_tfidf_results_1gram.csv

24 File hasil_prediksi_NB_tfidf_results_2gram.csv

25 File hasil_prediksi_NB_tfidf_results_3gram.csv

26 File hasil_prediksi_NB_tfidf_results_4gram.csv

27 File hasil_prediksi_NB_tfidf_results_5gram.csv

28 File hasil_prediksi_LR_tfidf_results_1gram.csv

29 File hasil_prediksi_LR_tfidf_results_2gram.csv

30 File hasil_prediksi_LR_tfidf_results_3gram.csv

31 File hasil_prediksi_LR_tfidf_results_4gram.csv

32 File hasil_prediksi_LR_tfidf_results_5gram.csv

33 File hasil_prediksi_DT_tfidf_results_1gram.csv

34 File hasil_prediksi_DT_tfidf_results_2gram.csv

35 File hasil_prediksi_DT_tfidf_results_3gram.csv

36 File hasil_prediksi_DT_tfidf_results_4gram.csv

37 File hasil_prediksi_DT_tfidf_results_5gram.csv

38 File hasil_prediksi_SVM_tfidf_results_1gram.csv

39 File hasil_prediksi_SVM_tfidf_results_2gram.csv

40 File hasil_prediksi_SVM_tfidf_results_3gram.csv

41 File hasil_prediksi_SVM_tfidf_results_4gram.csv

42 File hasil_prediksi_SVM_tfidf_results_5gram.csv