PENGGUNAAN ALGORITMA RANDOM FOREST DALAM MENILAI SENTIMEN PENGGUNA TWITTER TERKAIT ISU PERCERAIAN

Muhammad Rafli Aulia Rojani Lutfi[1]

[1]informatika, Universitas Pembangunan Nasional Veteran Jatim, Indonesia
20081010061@student.upnjatim.ac.id

ABSTRACT

Perceraian adalah isu yang kompleks dan sensitive dalam Masyarakat Indonesia. Berbagai factor yang mempengharuhi terjadi pereraian antara lain perselisihan yang terus berulang, kekerasan dalam rumah tanga, tidak adanya keharmonisan, permasalahan ekonomi, dll. Pada Platform Twitter banyak sekali orang yang memberikan pendapatnya mengenai perceraian. Hal ini menimbulkan beragam sentimen masyarakt yang dapat menjadi tolak ukur seseorang menilai suati fenomena yang sedang berlangsung.
Oleh karena itu, digunakan algoritma Random Forest untuk klasifikasi sentiment Masyarakat terhadap isu perceraian. Tujuan utamanya adalah mengukur Tingkat performansi dari algoritma Random Forest dalam mengklasifikasikan sentiment negatif, netral, dan positif. Menggunakan Tingkat accuracy dan nilai F1-score yang didaptkan dari confusion matrix untuk menilai seberapa baik model dalam melakukan klasifikasi. Dataset yang digunakan sekitar 1500 data tweet dari hasil scrapping melalui platform Twitter.
Dari hasil pengujian didapatkan Tingkat akurasi daari Algoritma Random Forest sebesar 70%. Diketahui bahwa Tingkat akurasi yang cukup rendah di sebabkan karena komposisi jumlah kelas yang sangat berbeda. Sehingga model lemah dalam memprediksi kelas positif.

KEYWORDS

Analisis Sentimen, Perceraian, Random Forest, Confusion Matrix

1. INTRODUCTION

Perceraian adalah isu yang kompleks dan sensitif dalam masyarakat Indonesia. Menurut [1] kasus perceraian di Indonesia dipicu oleh berbagai macam faktor, antara lain perselisihan yang terus berulang, kekerasan dalam rumah tangga, tidak adanya keharmonisan antara suami dan istri, permasalahan ekonomi, salah satu pihak pergi meninggalkan pihak lainnya, kecemburuan berlebih terhadap pasangan, dan adanya campur tangan pihak ketiga. Berdasarkan data dari laporan Statistik Indonesia 2023, jumlah kasus perceraian di Indonesia diperkirakan mencapai 516.334 pada tahun 2022. Angka ini mengalami peningkatan sebesar 15,31% dibandingkan dengan tahun 2021 yang mencatat 447.743 kasus. Angka tersebut merupakan jumlah kasus perceraian tertinggi dalam enam tahun terakhir di Indonesia.

Dalam era digital saat ini, media sosial seperti Twitter menjadi platform yang populer untuk berbagi pendapat dan pengalaman terkait perceraian. Hal ini dapat menimbulkan beragam sentiment. Sentimen masyarakat yang beragam dapat menjadi tolak ukur seseorang meniliai suatu fenomena yang sedang berlangsung.

Analisis sentimen atau opinion mining adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi masyarakat terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atribut [2]. Analisis sentimen diperlukan guna mengetahui opini masyarakat terhadap kebijakan pemerintah. Dataset adalah hal yang sangat penting dalam analisis sentimen. Analisis sentimen dilakukan terhadap ratusan ribu

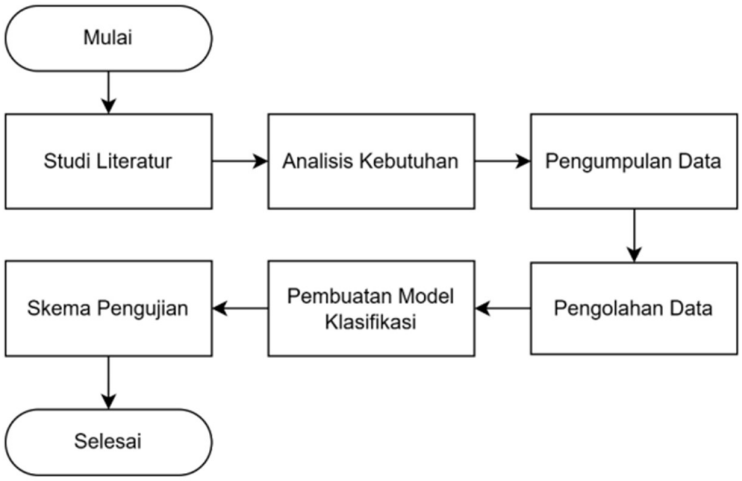
data tweet Twitter yang didapatkan melalui TwitterAPI untuk mengetahui kecenderungan masyarakat Indonesia terhadap isu perceraian, apakah ke arah positif, negatif, atau netral. Analisis sentimen terhadap dataset yang besar atau big data memiliki karakteristik yaitu ukuran data, variasi fitur, dan frekuensi kemunculan [3]. Jumlah dan variasi dataset mempengaruhi hasil klasifikasi sentimen, semakin besar data yang digunakan pada proses training semakin bagus performa model klasifikasi yang dihasilkan [4]. Penggunaan dataset yang besar juga memastikan hasil klasifikasi sentimen valid karena telah mewakili suara setiap orang.

Random Forest juga tergolong memiliki akurasi yang tinggi jika di banding algoritma klasifikasi lainnya karena memiliki karakteristik ensemble learning. Dibuktikan dengan penelitian yang telah dilakukan oleh [5] terhadap sentimen penilaian kustomer pada produk-produk media sosial, dari hasil perbandingan diketahui bahwa Random Forest paling unggul. Penelitian lain juga dilakukan oleh [6] guna meningkatkan performa hasil klasifikasi sentimen sebelumnya terhadap ulasan produk seluler di Kanada, diketahui bahwa Random Forest mampu meningkatkan akurasi. Melalui beberapa penelitian tersebut, maka Random Forest dipilih karena dapat menghasilkan akurasi yang tinggi dengan karakteristiknya yang ensemble learning.

2. METODELOGI

Pada bab Metodologi, penelitian ini menjelaskan langkah-langkah yang digunakan selama penelitian analisis sentimen. Diagram alir digunakan untuk menggambarkan terjadinya suatu proses sehingga mempermudah memahami proses yang sedang terjadi. Penelitian ini dapat dikategorikan dalam tipe penelitian implementatif karena mengimplementasi algoritma Random Forest untuk mengetahui sentimen masyarakat terhadap isu perceraian di Indonesia.

Langkah-langkah metodologi yang dilakukan untuk mencapai tujuan dari penelitian ini digambarkan melalui diagram alir di bawah ini.



Gambar 1. Langkah-langkah penelitian

Tahapan penelitian ini dimulai dengan melakukan studi literatur untuk mengidentifikasi permasalahan yang hendak diselesaikan menggunakan teknologi. Kemudian dilajutkan dengan analisis kebutuhan selama penelitian, lalu melakukan pengumpulan data berdasarkan topik yang diangkat. Pada tahap pembuatan model terlebih dahulu dilakukan preprocessing data yang kemudian dilanjutkan dengan pembuatan model klasifikasi menggunakan algoritma machine learning. Di akhir penelitian akan dilakukan pengujian terhadap hasil model berdasarkan rancangan skenario.

2.1. Studi Literatur

Studi literatur dilakukan dengan mengumpulkan sumber informasi yang relevan berdasarkan topik yang diangkat. Sumber referensi yang digunakan sebagai dasar acuan penelitian meliputi buku, skripsi, penelitian, jurnal, karya ilmiah, tesis, internet. Topik-topik referensi yang digunakan berkaitan dengan analisis sentimen, penggunaan algoritma Random Forest, dan topik lainnya yang mendukung tujuan penelitian ini.

2.2. Analisis Kebutuhan

Pada tahap analisis kebutuhan ini meliputi analisis kebutuhan data dan kebutuhan software & hardware agar penelitian berjalan sesuai harapan.

2.3. Kebutuhan Data

Data yang dibutuhkan dalam penelitian ini adalah data tweet pada media sosial Twitter yang berhubungan dengan isu perceraian di indonesia, dengan batasan data yang diambil mulai dari tanggal 29 November 2023 sampai 13 Desember 2023.

2.4. Kebutuhan Hardware dan Software

Pada penelitian ini diperlukan perangkat software dan hardware yang dapat menunjang keberhasilan penelitian. penelitian ini menggunakan Laptop Asus Vivobook M1403QA dengan kapasitas RAM sebesar 16GB dan CPU mencapai 8 core. Sedangkan software yang akan digunakan meliputi web browser untuk menjalankan program, jupyter notebook, dan bahasa pemrograman Python.

2.5. Pengumpulan Data

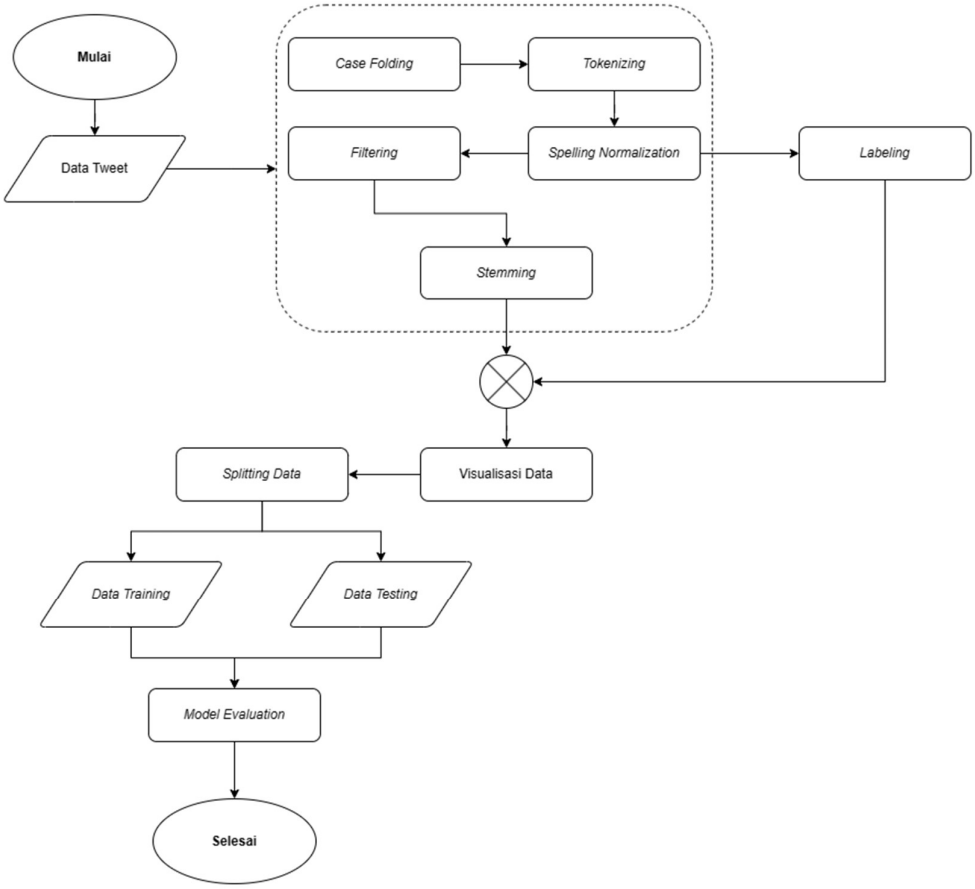
Tahap pengumpulan data bertujuan mendapatkan data mentah sebanyak mungkin dengan memperhatikan kata kunci yang digunakan dan batasan waktu. Tahapan ini meliputi penjelasan sumber data dan rancangan awal struktur data.

2.5.1 Sumber Data

Penelitian ini menggunakan sumber data sekunder yang diperoleh dari media sosial Twitter melalui TwitterAPI dengan mengambil seluruh data tweet yang berkaitan dengan opini atau pendapat masyarakat tentang isu perceraian di Indonesia. Data tweet diambil dengan ketententuan berbahasa Indonesia dan menggunakan kata kunci “perceraian”. Data tweet yang diambil adalah tweet yang dibuat sejak tanggal 29 Januari 2023 sampai 13 Desember 2023.

2.6. Pengolahan Data

Data set tweet pada Twitter yang telah dikumpulkan mengenai isu perceraian diolah menggunakan software Python 3.11 (64-bit). Sebelum dilakukan klasifikasi menggunakan algoritma machine learning, terlebih dahulu dilakukan preprocessing guna menghilangkan beberapa permasalahan yang bisa mengganggu tahap pemrosesan data menggunakan algoritma klasifikasi. Selanjutnya dilakukan pelabelan secara otomatis untuk mengetahui apakah sentimen tweet termasuk positif, netral, atau negatif. Dilanjutkan dengan melakukan visualisasi terhadap dataset dan perhitungan menggunakan metode feature extraction.



Gambar 2. Diagram alir metode pengolahan data

1. Menggunakan data tweet yang didapatkan melalaui crawling dan scraping pada platform media sosial Twitter melalui fasilitas terbuka Twitter API (Application Programming Interface) berdasarkan topik permasalahan.
2. Data Preprocessing
    a. Melakukan case folding, yaitu mengubah semua kata dengan huruf besar (upper case) dalam dokumen menjadi huruf kecil (lower case) seluruhnya serta menghilangkan tanda baca, link, nomor, dan character.
    b. Melakukan tokenizing yang bertujuan memecah kalimat menjadi kumpulan kata yang biasa disebut token.
    c. Melakukan spelling normalization untuk mengubah data tweet yang mengandung kata tidak baku menjadi kata baku. Melakukan filtering atau stopword removal adalah tahapan dengan menghilangkan kata yang dianggap tidak penting (stoplist) dan menyimpan kata yang bermakna penting (wordlist).
    d. Melakukan stemming, yaitu mengonversi kata yang berimbuhan menjadi bentuk kata dasar.
3. Melabeli setiap data tweet dengan label positif, negatif, dan netral menggunakan package Python yaitu transformers dengan model fine-tuned Indonesian RoBERTa
4. Melakukan visualisasi data tweet berupa grafik dan word cloud menggunakan package matplotlib, seaborn dan wordcloud yang tersedia di Python.

5. Membagi data tweet menjadi data training dan testing dengan rasio tertentu yang nantinya digunakan sebagai masukkan pada tahap pembuatan model klasifikasi.
6. Melakukan evaluasi model untuk menilai seberapa baik model dalam memprediksi sentiment pada dataset isu perceraian.

2.7. Pembuatan Model Klasifikasi

Setelah dataset dilakukan preprocessing dan dilakukan lebeling selanjutkan dilakukan klasifikasi menggunakan algoritma machine learning Random Forest guna mengetahui performa model. Setelah itu dipilih model terbaik untuk mengklasifikasi data tweet baru. Langkah-langkah klasifikasi data antara lain.

1. Menggunakan data tweet hasil ekstraksi fitur meliputi data training dan data testing yang dipanggil dari file data berformat pickle.
2. Membuat model klasifikasi Random Forest
   a. Menentukan banyak tree untuk setiap model pohon klasifikasi dari $b = 1$ sampai B sehingga untuk setiap tree dapat dituliskan Tb
   b. Membuat sampel dataset untuk setiap tree Tb dari N total data training dengan cara bootstrapping atau diacak
   c. Menentukan variabel m yaitu banyak fitur atau atribut secara acak sebagai splitting point sebanyak $\sqrt{p}$ total fitur.
   d. Melakukan perhitungan klasifikasi untuk setiap tree Tb diawali dengan menentukan root node dengan mencari kriteria pemisah yang memiliki nilai gini-split paling kecil pada data training.
   e. Mengulangi perhitungan entorpy-split untuk membentuk internal node sampai data dari sebuah node memiliki kelas homogen atau sudah tidak terdapat atribut beserta kriteria pemisahan, proses ini dilakukan pada setiap Tb. Jadi setiap Tb memiliki internal node pemisah dan kedalaman yang berbeda-beda.
   f. Melakukan majority voting terhadap hasil klasifikasi model setiap Tb sehingga didapatkan hasil akhir klasifikasi yaitu kelas prediksi. Proses ini disebut *debugging*.

2.8. Model Evaluation

Tahapan skema evaluasi model meliputi beberapa Langkah antara lain.

1. Confusion matrix adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi pada set data pengujian, membandingkan prediksi model dengan nilai sebenarnya. Matriks ini terdiri dari empat sel: True Positive (TP) yaitu jumlah kasus yang benar diprediksi sebagai positif, True Negative (TN) yaitu jumlah kasus yang benar diprediksi sebagai negatif, False Positive (FP) yaitu jumlah kasus yang seharusnya negatif, tetapi diprediksi sebagai positif, dan False Negative (FN) yaitu jumlah kasus yang seharusnya positif, tetapi diprediksi sebagai negatif.
2. Precision (presisi) adalah ukuran dari sejauh mana model benar dalam memprediksi kasus positif. Precision memberikan informasi tentang akurasi prediksi positif model dan mengukur tingkat kepercayaan terhadap kelas positif.
3. Recall (recall atau sensitivitas) adalah ukuran dari sejauh mana model mampu menemukan semua kasus positif yang seharusnya ditemukan. Recall memberikan pandangan tentang seberapa baik model dapat mengidentifikasi semua kasus positif yang ada.
4. F1 score adalah metrik yang menggabungkan precision dan recall menjadi satu angka yang menyediakan ukuran komprehensif tentang kinerja model. F1 score berguna ketika terdapat trade-off antara precision dan recall; misalnya, ketika keseimbangan antara kedua metrik itu penting.

3. HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang hasil dan pembahasan terkait implementasi program analisis sentimen terhadap data tweet platform media sosial Twitter berkaitan tentang isu perceraian di Indonesia menggunakan algoritma machine learning Random Forest.

3.1 Pengumpulan Data

Tahap ini merupakan tahapan awal penelitian yang bertujuan mendapatkan dataset sebanyak mungkin yang merupakan data teks tweet Twitter yang berkaitan isu perceraian di Indonesia. Berdasarkan penjelasan pada sub-bab 2.5, pengumpulan data terdiri dari crawling data dan scraping data.

Tahap crawling data bertujuan mendapatkan data tautan tweet pada setiap index halaman media sosial Twitter berdasarkan paramater kata kunci. Crawling data memanfaatkan github repository milik Helmi Satria. Masukkan kata kunci yang digunakan berupa kombinasi query, dapat meliputi gabungan satu frasa atau kemunculan satuan kata dalam satu kalimat. Penelitian ini menggunakan kata kunci frasa yaitu "perceraian".

Data yang dicari merupakan data tweet berbahasa Indonesia dengan rentang waktu tertentu. Didapatkan total data link tweet sekitar 1500 data yang kemudian disimpan dalam bentuk file berformat csv. Proses crawling data menjadi file berbentuk csv menggunakan library Pandas Python.

3.2 Preprocessing

Data tweet yang didapatkan dari platform Twitter merupakan data yang tidak terstruktur sehingga memerlukan suatu proses dalam text mining yaitu preprocessing data agar dapat meningkatkan ketepatan klasifikasi pada data. Tahap yang dilakukan pada preprocessing meliputi case folding, tokenizing, spelling normalization, filtering, dan stemming.

3.2.1 Case Folding

Case folding bertujuan menyeragamkan bentuk huruf serta penghapusan angka dan tanda baca yang tidak diperlukan. Tahapan ini terdiri dari beberapa bagian, yaitu:

1. Mengubah uppercase menjadi lowercase seluruhnya
2. Menghapus spasi di awal dan di akhir kalimat
3. Menghapus tab, baris baru, dan back slice
4. Menghapus non-ASCII
5. Menghapus mention dan hashtag
6. Menghapus URL
7. Menghapus tanda baca
8. Menghapus angka
9. Mengubah spasi ganda menjadi spasi tunggal
10. Menghapus karakter Tunggal

Tabel 1. Contoh hasil *case folding* data tweet

Original Tweet	Clean Text
— kehidupan pasca perceraian https://t.co/9TJhxA1n4d	kehidupan pasca perceraian

@tanyarlfes Polisi Benarkan Ammar Zoni Dtangkap Lagi Terkait Narkoba" untung gue ngga nonton ampe abis podcast'a dia sama deddy, tadi'a gue kasian pas d cerai Irish Bella. Tp seperti'a keputusan irish tepat, atau gara2 perceraian'a maka'a ammar "make" lagi? Ah gue nonton debat capres ajalah"	polisi benarkan ammar zoni dtangkap lagi terkait narkoba untung gue ngga nonton ampe abis podcast dia sama deddy tadi gue kasian pas cerai irish bella tp seperti keputusan irish tepat atau gara perceraian maka ammar make lagi ah gue nonton debat capres ajalah
@milkywaygyal @18fesss Sekarang kl kasus si sender, dia rugi ga menurutmu? Dan kl ga ada jalan keluar (si sender sudah mengajak ngobrol tp ga ada ngaruhnya), dan berujung ke perceraian...rugi ga menurutmu?	sekarang kl kasus si sender dia rugi ga menurutmu dan kl ga ada jalan keluar si sender sudah mengajak ngobrol tp ga ada ngaruhnya dan berujung ke perceraian rugi ga menurutmu

3.2.2 Tokenizing

Proses tokenizing dilakukan setelah data tweet memiliki bentuk yang seragam pada proses case folding sebelumnya. Tokenizing bertujuan memisah setiap kata dengan spasi sebagai indikator pemisah (delimiter). Setiap kata hasil pemisahan disebut sebagai token. Token akan menjadi variabel yang memiliki nilai tertentu. Berikut ini contoh hasil tokenizing.

Clean Text	Tokenize Text
kehidupan pasca perceraian	['kehidupan', 'pasca', 'perceraian']
polisi benarkan ammar zoni dtangkap lagi terkait narkoba untung gue ngga nonton ampe abis podcast dia sama deddy tadi gue kasian pas cerai irish bella tp seperti keputusan irish tepat atau gara perceraian maka ammar make lagi ah gue nonton debat capres ajalah	['polisi', 'benarkan', 'ammar', 'zoni', 'dtangkap', 'lagi', 'terkait', 'narkoba', 'untung', 'gue', 'ngga', 'nonton', 'ampe', 'abis', 'podcast', 'dia', 'sama', 'deddy', 'tadi', 'gue', 'kasian', 'pas', 'cerai', 'irish', 'bella', 'tp', 'seperti', 'keputusan', 'irish', 'tepat', 'atau', 'gara', 'perceraian', 'maka', 'ammar', 'make', 'lagi', 'ah', 'gue', 'nonton', 'debat', 'capres', 'ajalah']
sekarang kl kasus si sender dia rugi ga menurutmu dan kl ga ada jalan keluar si sender sudah mengajak ngobrol tp ga ada ngaruhnya dan berujung ke perceraian rugi ga menurutmu	['sekarang', 'kl', 'kasus', 'si', 'sender', 'dia', 'rugi', 'ga', 'menurutmu', 'dan', 'kl', 'ga', 'ada', 'jalan', 'keluar', 'si', 'sender', 'sudah', 'mengajak', 'ngobrol', 'tp', 'ga', 'ada', 'ngaruhnya', 'dan', 'berujung', 'ke', 'perceraian', 'rugi', 'ga', 'menurutmu']

3.2.3 Spelling Normalization

Spelling normalization adalah tahapan untuk memperbaiki pengejaan kata yang tidak tepat atau tidak baku menjadi kata baku berdasarkan KBBI. Untuk memperbaiki ejaan kata dibutuhkan koleksi data atau dictionary yang berisi kumpulan kata salah dan pembenarannya. File normalization_resource2.csv berisi ribuan kata perbaikan ejaan bahasa indonesia, dapat dilihat pada halaman Lampiran. Selanjutnya dilakukan iterasi setiap kalimat tweet agar setiap kata dapat

dilakukan pencocokan dengan dictionary spelling normalization. Berikut ini hasil dari implementasi normalisasi.

Tokenize Text	Normalize Text
['kehidupan', 'pasca', 'perceraian']	['kehidupan', 'pasca', 'perceraian']
['polisi', 'benarkan', 'ammar', 'zoni', 'dtangkap', 'lagi', 'terkait', 'narkoba', 'untung', 'gue', 'ngga', 'nonton', 'ampe', 'abis', 'podcast', 'dia', 'sama', 'deddy', 'tadi', 'gue', 'kasian', 'pas', 'cerai', 'irish', 'bella', 'tp', 'seperti', 'keputusan', 'irish', 'tepat', 'atau', 'gara', 'perceraian', 'maka', 'ammar', 'make', 'lagi', 'ah', 'gue', 'nonton', 'debat', 'capres', 'ajalah']	['polisi', 'benarkan', 'ammar', 'zoni', 'ditangkap', 'lagi', 'terkait', 'narkoba', 'untung', 'gue', 'enggak', 'menonton', 'sampai', 'habis', 'podcast', 'dia', 'sama', 'deddy', 'tadi', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'tapi', 'seperti', 'keputusan', 'irish', 'tepat', 'atau', 'gara', 'perceraian', 'maka', 'ammar', 'memakai', 'lagi', 'ah', 'gue', 'menonton', 'debat', 'capres', 'ajalah']
['sekarang', 'kl', 'kasus', 'si', 'sender', 'dia', 'rugi', 'ga', 'menurutmu', 'dan', 'kl', 'ga', 'ada', 'jalan', 'keluar', 'si', 'sender', 'sudah', 'mengajak', 'ngobrol', 'tp', 'ga', 'ada', 'ngaruhnya', 'dan', 'berujung', 'ke', 'perceraian', 'rugi', 'ga', 'menurutmu']	['sekarang', 'kalo', 'kasus', 'sih', 'sender', 'dia', 'rugi', 'enggak', 'menurutmu', 'dan', 'kalo', 'enggak', 'ada', 'jalan', 'keluar', 'sih', 'sender', 'sudah', 'mengajak', 'mengobrol', 'tapi', 'enggak', 'ada', 'ngaruhnya', 'dan', 'berujung', 'ke', 'perceraian', 'rugi', 'enggak', 'menurutmu']

3.2.4 Filtering

Filtering dilakukan dengan menyimpan token hasil spelling normalization yang memiliki makna penting dan membuang makna yang tidak penting. filtering dilakukan dengan menghapus stopword seperti kata untuk, juga, perlu, di, dan lain-lain. Berikut ini hasil dari penerapan filtering.

Normalize Text	Filtered Text
['kehidupan', 'pasca', 'perceraian']	['kehidupan', 'pasca', 'perceraian']
['polisi', 'benarkan', 'ammar', 'zoni', 'ditangkap', 'lagi', 'terkait', 'narkoba', 'untung', 'gue', 'enggak', 'menonton', 'sampai', 'habis', 'podcast', 'dia', 'sama', 'deddy', 'tadi', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'tapi', 'seperti', 'keputusan', 'irish', 'tepat', 'atau', 'gara', 'perceraian', 'maka', 'ammar', 'memakai', 'lagi', 'ah', 'gue', 'menonton', 'debat', 'capres', 'ajalah']	['polisi', 'benarkan', 'ammar', 'zoni', 'dtangkap', 'terkait', 'narkoba', 'untung', 'gue', 'menonton', 'habis', 'podcast', 'deddy', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'keputusan', 'irish', 'gara', 'perceraian', 'ammar', 'memakai', 'ah', 'gue', 'menonton', 'debat', 'capres', 'ajalah']
['sekarang', 'kalo', 'kasus', 'sih', 'sender', 'dia', 'rugi', 'enggak', 'menurutmu', 'dan', 'kalo', 'enggak', 'ada', 'jalan', 'keluar', 'sih', 'sender', 'sudah', 'mengajak', 'mengobrol', 'tapi', 'enggak', 'ada', 'ngaruhnya', 'dan', 'berujung', 'ke', 'perceraian', 'rugi', 'enggak', 'menurutmu']	['kalo', 'sih', 'sender', 'rugi', 'menurutmu', 'kalo', 'jalan', 'sih', 'sender', 'mengajak', 'mengobrol', 'ngaruhnya', 'berujung', 'perceraian', 'rugi', 'menurutmu']

3.2.5 Stemming

Pada tahap ini bertujuan untuk mengubah kata imbuhan menjadi bentuk kata dasar. Imbuhan yang dihilangkan berupa awalan dan akhiran dari sebuah kata. Berikut ini hasil dari penerapan stemming.

Filtered Text	Stemmed Text
['kehidupan', 'pasca', 'perceraian']	['hidup', 'pasca', 'cerai']
['polisi', 'benarkan', 'ammar', 'zoni', 'dtangkap', 'terkait', 'narkoba', 'untung', 'gue', 'menonton', 'habis', 'podcast', 'deddy', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'keputusan', 'irish', 'gara', 'perceraian', 'ammar', 'memakai', 'ah', 'gue', 'menonton', 'debat', 'capres', 'ajalah']	['polisi', 'benar', 'ammar', 'zoni', 'dtangkap', 'kait', 'narkoba', 'untung', 'gue', 'tonton', 'habis', 'podcast', 'deddy', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'putus', 'irish', 'gara', 'cerai', 'ammar', 'pakai', 'ah', 'gue', 'tonton', 'debat', 'capres', 'aja']
['kalo', 'sih', 'sender', 'rugi', 'menurutmu', 'kalo', 'jalan', 'sih', 'sender', 'mengajak', 'mengobrol', 'ngaruhnya', 'berujung', 'perceraian', 'rugi', 'menurutmu']	['kalo', 'sih', 'sender', 'rugi', 'turut', 'kalo', 'jalan', 'sih', 'sender', 'ajak', 'obrol', 'ngaruhnya', 'ujung', 'cerai', 'rugi', 'turut']

3.3 Labeling

Pelabelan dilakukan pada data hasil spelling normalization sebelum dilakukan proses filtering, hal ini guna mempertahankan konteks kalimat. Pada proses pelabelan 1500 data tidak dilakukan secara manual, namun menggunakan fine-tuned model dari pretrained model Indonesian RoBERTa Base melalui library Transformers. Pelabelan menghasilkan kelas positif, negatif, dan netral. Berikut ini hasil pelabelan dataset.

Normalize Text	Sentiment
['kehidupan', 'pasca', 'perceraian']	Positive
['polisi', 'benarkan', 'ammar', 'zoni', 'ditangkap', 'lagi', 'terkait', 'narkoba', 'untung', 'gue', 'enggak', 'menonton', 'sampai', 'habis', 'podcast', 'dia', 'sama', 'deddy', 'tadi', 'gue', 'kasihan', 'pas', 'cerai', 'irish', 'bella', 'tapi', 'seperti', 'keputusan', 'irish', 'tepat', 'atau', 'gara', 'perceraian', 'maka', 'ammar', 'memakai', 'lagi', 'ah', 'gue', 'menonton', 'debat', 'capres', 'ajalah']	negative
['sekarang', 'kalo', 'kasus', 'sih', 'sender', 'dia', 'rugi', 'enggak', 'menurutmu', 'dan', 'kalo', 'enggak', 'ada', 'jalan', 'keluar', 'sih', 'sender', 'sudah', 'mengajak', 'mengobrol', 'tapi', 'enggak', 'ada', 'ngaruhnya', 'dan', 'berujung', 'ke', 'perceraian', 'rugi', 'enggak', 'menurutmu']	negative

3.4 Visualisasi Dataset

Pada dataset dapat dilakukan juga pengelompokkan kata yang sering muncul yang digambarkan melalui visualisasi word cloud. Visualisasi kemunculan setiap kata digunakan agar

mempermudah analisis terhadap fitur atau kata yang dapat mempengaruhi hasil klasifikasi. Berikut visualisasi word cloud pada dataset.

Gambar 3. Word Cloud Frekuensi data

3.5 Pembagian Data

Sebelum menuju proses pembuatan model klasifikasi menggunakan algoritma machine learning terlebih dahulu dilakukan splitting atau pembagian dataset menjadi data training dan data testing berdasarkan rasio yang ditetapkan. Pembagian data dilakukan menggunakan library yang menyediakan fungsi split data, yaitu library sklearn. Pembagian rasio terbesar digunakan untuk proses pelatihan model sedangkan sisanya digunakan untuk proses pengujian. Untuk penelitian kali ini penulis menggunakan komposisi 70:30, 70 persen untuk data training dan 30 persen untuk data testing.

3.6 Feature Extraction

Setelah dataset selesai dilakukan preprocessing dan splitting menjadi data train dan test, selanjutnya akan dilakukan pembobotan pada masing-masing kata sehingga menghasilkan fitur kata. Proses ini dikenal dengan sebutan word embedding atau dapat dikatakan juga features extraction. Metode yang digunakan untuk ekstraksi fitur adalah Document-term Matrix dengan fungsi CountVectorizer() yang terdapat pada library Sklearn.

Nilai ekstraksi fitur merupakan angka diskrit frekuensi kemunculan kata pada setiap baris data atau kalimat tweet. Nilai tersebut digunakan sebagai masukkan fungsi pembuatan model klasifikasi yang akan menghasilkan kelompok klasifikasi sentimen yaitu positif, negatif, dan netral. Kemudian untuk menentukan kelas-kelas sentimen tersebut akan dilihat berdasarkan bobot yang diperoleh atau dari seberapa besar kemampuan setiap variabel.

3.7 Klasifikasi Data

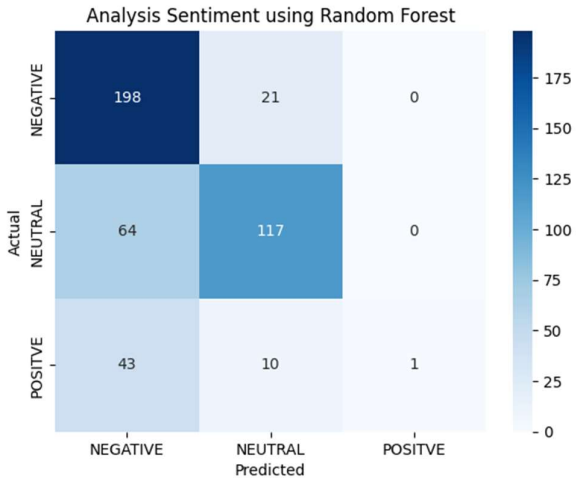
Setelah melalui tahap preprocessing, pembagian data, dan word embedding, selanjutnya dataset akan melalui tahap klasifikasi data. Data yang digunakan dalam klasifikasi akan dibedakan menjadi dua, yaitu data training dan data testing. Penelitian ini akan dilakukan pengujian terhadap ketepatan klasifikasi model berdasarkan variasi rasio pembagian data pada tahap skenario pengujian. Sebagai contoh jika menggunakan rasio 70:30 maka 70% adalah data training yaitu sebanyak 1059 tweet dan data testing sebanyak 454 tweet.

3.7.1 Random Forest

Pada tahap klasifikasi menggunakan Random Forest akan ditampilkan juga rancangan kode program berdasarkan langkah-langkah dan persamaan yang telah dijelaskan pada sub-bab 2.11 dan sub-bab 3.5. Dalam langkah awal perhitungan Random Forest diperlukan banyak jumlah tree ($n\_estimator$) yang akan digunakan untuk membuat perhitungan klasifikasi. Banyak pohon keputusan yang akan digunakan pada penelitian ini adalah 150 tree. Setelah itu menentukan banyak fitur yang digunakan masing-masing tree. Karena jenis model adalah model klasifikasi, menentukan banyak fitur dengan cara $\sqrt{p}$ dengan p adalah total fitur. Hal ini dilakukan guna mengurangi tingkat kesamaan tiap model tree dan meminimalisir terjadinya bias. Setiap pohon keputusan dalam Random Forest menggunakan data train sejumlah 80% dari total data train asal atau masukkan awal di kelas 'RandomForest'. Proses ini dilakukan agar setiap keputusan klasifikasi memiliki hasil klasifikasi yang beragam sebelum dilakukan pengambilan keputusan keseluruhan atau majority voting.

3.8 Model Evaluation

Pada tahap ini, akurasi dari penggunaan random forest untuk data training sebesar 80% dan data testing sebesar 70%. Untuk meniliai lebih lanjut kemampuan model dalam mengklasifikasikan data, penulis menggunakan confusion matrix untuk menilai performa model. Berikut ini confusion matrix yang didapatkan.



Gambar 4. Confusion matrix pada model *random forest*

Dari gambar diatas dapat kita ketehaui bahwa model cenderung berhasil memprediksi data dengan kelas negative dan cukup baik dalam memprediksi data dengan kelas neutral, namun tidak dengan data dengan kelas positif. Untuk data yang memiliki kelas positif model malah cenderung memprediksi sebagai kelas negatif, dibuktikan

dengan jumlah data sebesar 43 data. Dari confusion matrix diatas kita bisa mendapatkan beberapa informasi antara lain nilai precision, recall, dan f1-score. berikut ini nilai-nilai yang berhasil penulis dapatkan antara lain.

Kelas	Precision	Recall	F1 Score
NEGATIVE	0.65	0.90	0.76
NEUTRAL	0.80	0.66	0.73
POSITVE	1.00	0.02	0.04

Dari table diatas kita dapat simpulkan bahwa model cenderung dapat memprediksi kelas negative dan neutral dengan nilai 0,76 dan 0,73. Namun tidak dengan kelas positif dengan nilai sangat rendah sebesar 0,04.

Setelah penulis cari tahu lebih dalam ternyata komposisi dari jumlah kelas sangat berpengaruh terhadap kemampuan model dalam memprediksi data. Jumlah kelas negatif yang berhasil ditemui sebesar 743 data, untuk kelas netral berjumlah 578 data, dan untuk kelas positif sebesar 192 data. Perbedaan jumlah kelas inilah yang menyebabkan performa model rendah.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan terhadap analisis sentimen kebijakan kenaikan harga BBM di Indonesia, dapat disimpulkan sebagai berikut ini:

1. Algoritma Random Forest Berhasil diimplementasikan untuk klasifikasi sentiment twitter terkait isu perceraian yang ada di Indonesia dengan jumlah data sebesar 1500 data.
2. Besar akurasi model random forest adalah 70%. Hal ini dapat terjadi dikarenakan terdapat ketimpangan pada jumlah komposisi tiap kelas.

REFERENCES

A. H. Najmuddin, N. Khamimah, dan N. S. Ufaira, "PERCERAIAN DI ERA DIGITAL: PENGARUH MEDIA SOSIAL DAN TEKNOLOGI," Jurnal Hukum dan Kewarganegaraan, vol. 1, 2023, doi: 10.3783/causa.v1i1.571.

B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2012.

P. Zikopoulos dan C. Eaton, 2011. "Understanding Big Dataa: Analytics for enterprise class hadoop and streaming data McGraw-Hill Osborne Media"

T. Pranckevičius dan V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," Baltic Journal of Modern Computing, vol. 5, no. 2, 2017, doi: 10.22364/bjmc.2017.5.2.05.

Kalasalingam Academy of Research and Education. IEEE Student Branch., Institute of Electrical and Electronics Engineers, dan IEEE Power & Energy Society, IEEE International Conference on Intelligent Techniques in Control, Optimization & Signal Processing : INCOS-'19 : 11th-13th April 2019.

Hegde, Y., and Padma, S. K., 2017. "Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada". International Advance Computing Conference 7