KLASIFIKASI OPINI PUBLIK BEASISWA PRAKERJA MENGGUNAKAN METODE MIX DISTRIBUSI NAÏVE BAYES

SKRIPSI

Oleh :
ZAKY FATHURRAHMAN
NIM. 200605110017

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

KLASIFIKASI OPINI PUBLIK BEASISWA PRAKERJA MENGGUNAKAN METODE MIX DISTRIBUSI NAÏVE BAYES

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
ZAKY FATHURRAHMAN
NIM. 200605110017

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

HALAMAN PERSETUJUAN

KLASIFIKASI OPINI PUBLIK BEASISWA PRAKERJA MENGGUNAKAN METODE MIX DISTRIBUSI NAÏVE BAYES

SKRIPSI

Oleh :
ZAKY FATHURRAHMAN
NIM. 200605110017

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 12 MARET 2026

Pembimbing I,	Pembimbing II,
Dr. Irwan Budi Santoso, M.Kom	Dr. Cahyo Crysdian, M.Cs
NIP. 19770103 201101 1 004	NIP. 19740424 200901 1 008

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

KLASIFIKASI OPINI PUBLIK BEASISWA PRAKERJA MENGGUNAKAN METODE MIX DISTRIBUSI NAÏVE BAYES

SKRIPSI

Oleh :
ZAKY FATHURRAHMAN
NIM. 200605110017

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 12 MARET 2026

Susunan Dewan Penguji

Ketua Penguji	: Dr. Totok Chamidy, M.Kom NIP. 19691222 200604 1 001	( )
Anggota Penguji I	: Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )
Anggota Penguji II	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Anggota Penguji III	: Dr. Cahyo Crysdian, M.Cs NIP. 19740424 200901 1 008	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M. Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Zaky Fathurrahman
NIM	: 200605110017
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Klasifikasi Opini Publik Beasiswa Prakerja Menggunakan Metode Mix Distribusi Naïve Bayes

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 12 Maret 2026
Yang membuat pernyataan,

Zaky Fathurrahman
NIM. 200605110017

MOTTO

"Pelaut yang tangguh tidak lahir di laut yang tenang."

HALAMAN PERSEMBAHAN

Puji syukur kehadirat ALLAH SWT yang selalu memberikan rahmat dan
Hidayah-Nya sehingga penulis diberi kemudahan dan semangat dalam
Menyelesaikan penulisan skripsi ini dengan baik.
Karya ini penulis persembahkan kepada:

Ibu dan Ayah tercinta, Rita dan Joni,
Yang selalu memberikan dukungan, semangat dan do'a yang tak pernah berhenti
Meski jarak memisahkan, dukungan dan kehangatan
kalian yang membuat penulis sampai pada titik ini.

Seluruh kerabat yang telah membantu dan memotivasi penulis
Hingga akhirnya penulis mampu menyelesaikan penelitian ini.

Dan teruntuk diri sendiri,
Terimakasih telah berjuang sampai titik tuju.

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

Segala puji hanya milik Allah Subhanahu Wa Ta'ala atas segala nikmat dan kasih sayang-Nya yang telah memudahkan penulis untuk menyelesaikan skripsi yang berjudul "Klasifikasi Opini Publik Beasiswa Prakerja Menggunakan Metode Mix Distribusi Naïve Bayes". Semoga shalawat dan salam senantiasa terlimpah kepada Nabi Muhammad Sallalahu 'Alaihi wa Sallam. Dan semoga kita semua mendapat syafaatnya di hari kiamat nanti, Aamiin.

Penulis mengucapkan rasa syukur dan terima kasih yang tak terhingga kepada semua pihak-pihak yang selalu memberikan bantuan dan motivasi kepada penulis untuk dapat menyelesaikan skripsi ini. Ucapan ini penulis sampaikan kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Supriyono, M. Kom., selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Irwan Budi Santoso, M.Kom selaku dosen pembimbing I dan Dr. Cahyo Crysdian, M.Cs selaku dosen pembimbing II yang telah memberikan bantuan dan arahan kepada penulis, sehingga bisa menuntaskan skripsi ini.

5. Dr. Totok Chamidy, M.Kom selaku dosen penguji I dan Okta Qomaruddin Aziz, M.Kom selaku dosen penguji II yang telah menguji serta memberikan masukan sehingga penulis dapat menuntaskan skripsi dengan baik.
6. Segenap Dosen, Admin, Laboran dan Mahasiswa Program Studi Teknik Informatika yang telah memberikan banyak dukungan, arahan dan bimbingan selama pengerjaan skripsi ini.
7. Kedua orang tua dan ketiga adik tercinta, Ayah Joni Putra Erika, Ibu Rita Leswarni, Fazil Azmi, Rizky Fahri, dan Alina Cahya Kirani untuk segenap cinta, doa, dan kasih sayang serta support yang tiada henti mengiringi segenap langkah penulis dalam menyelesaikan penelitian ini. Terima kasih selalu menjadi kekuatan terbesar bagi penulis untuk melakukan dan menuntaskan segala hal. Semoga Allah selalu melimpahkan segalanya untuk kalian.
8. Khalid, Shofey, Imam, Dani, teman-teman Pemuda Sinarmas yang sudah menjadi saudara saya selama berada di perantauan, dan teman-teman Angkatan 2020 Teknik Informatika “INTEGER” yang telah menemani dan memberi banyak bantuan intelektual, motivasi, dan semangat kepada penulis untuk menyelesaikan penelitian ini.
9. Alm. Helmi Noor Hafidz dan Keluarga, saudara seperjuangan yang akan selalu abadi dalam hidup penulis, dan Terima Kasih sudah menjadi keluarga penulis selama diperantauan.
10. Semua pihak yang tidak dapat disebutkan satu per satu, yang terlibat dan memberi dukungan dalam perjalanan penulisan skripsi ini.

Akhir kata, penulis mengakui bahwa penulisan pada skripsi ini masih banyak kekurangan. Saya berharap semoga skripsi ini diterima sebagai amal ibadah yang tulus dan bermanfaat di sisi Allah Subhanahu Wa Ta'ala. Semoga karya ini menjadi bagian dari kontribusi yang tak terputus dalam rangka memperkuat dan mengembangkan ilmu pengetahuan, serta melaksanakan tugas sebagai hamba Allah yang berkomitmen.

Wassalamualaikum Warahmatullahi Wabarakatuh.

Malang, 09 April 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Fathurrahman, Zaky. 2026. ***Klasifikasi Opini Publik Beasiswa Prakerja Menggunakan Metode Mix Distribusi Naïve Bayes***. Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Irwan Budi Santoso, M.Kom, (II) Dr. Cahyo Crysdian.

Kata kunci: *Beasiswa Prakerja, Klasifikasi Teks, Mix Distribusi Naïve Bayes, Analisis Sentimen, Twitter.*

Program beasiswa Prakerja yang diluncurkan pemerintah Indonesia untuk menekan angka pengangguran pasca pandemi Covid-19 menimbulkan beragam opini publik di media sosial, khususnya Twitter. Meskipun antusiasme masyarakat tinggi, muncul berbagai kritik terkait efektivitas, transparansi sistem seleksi, hingga kendala teknis operasional. Penelitian ini bertujuan untuk mengklasifikasikan opini publik mengenai program beasiswa Prakerja di media sosial Twitter guna memberikan evaluasi objektif bagi pemerintah terhadap efektivitas dan transparansi sistem tersebut. Menggunakan pendekatan *Natural Language Processing* (NLP), penelitian ini menerapkan metode *Mix Distribusi Naïve Bayes* yang mengombinasikan *Multinomial* dan *Gaussian Naïve Bayes* untuk menangani karakteristik data teks yang beragam. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, ekstraksi fitur menggunakan TF-IDF, serta reduksi dimensi melalui *Truncated SVD*. Untuk mengatasi ketidakseimbangan data, teknik *RandomOverSampler* diterapkan, sementara optimasi model dilakukan menggunakan *GridSearchCV*. Pengujian dilakukan dengan tiga skenario pembagian data (70:30, 80:20, dan 90:10) dan diukur melalui parameter *Confusion Matrix* seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Hasil klasifikasi ini diharapkan dapat menjadi rujukan bagi instansi terkait dalam memahami sentimen masyarakat serta melakukan perbaikan kebijakan layanan publik di masa mendatang.

ABSTRACT

Fathurrahman, Zaky. 2026. **Public Opinion Classification of Prakerja Scholarship Using Mix Distribution Naïve Bayes Method.** Thesis. Informatics Engineering Department, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Advisors: (I) Dr. Irwan Budi Santoso, M.Kom, (II) Dr. Cahyo Crysdian.

Keywords: *Prakerja Scholarship, Text Classification, Mix Distribution Naïve Bayes, Sentiment Analysis, Twitter*.

The Prakerja scholarship program, launched by the Indonesian government to reduce unemployment rates following the Covid-19 pandemic, has generated diverse public opinions on social media, particularly on Twitter. Despite high public enthusiasm, various criticisms have emerged regarding its effectiveness, selection system transparency, and technical operational constraints. This study aims to classify public opinion regarding the Prakerja scholarship program on Twitter to provide an objective evaluation for the government concerning the system's effectiveness and transparency. Utilizing a Natural Language Processing (NLP) approach, this research implements the *Mix Distribution Naïve Bayes* method, which combines *Multinomial* and *Gaussian Naïve Bayes* to handle diverse text data characteristics. The research stages include data collection, preprocessing, feature extraction using TF-IDF, and dimensionality reduction through *Truncated SVD*. To address data imbalance, the *RandomOverSampler* technique is applied, while model optimization is conducted using *GridSearchCV*. Testing was performed using three data splitting scenarios (70:30, 80:20, and 90:10) and measured through Confusion Matrix parameters such as *Accuracy*, *Precision*, *Recall*, and *F1-Score*. The results of this classification are expected to serve as a reference for relevant institutions in understanding public sentiment and improving public service policies in the future.

الملخص

فتح الرحمن، زكي. ٢٠٢٦. **باستخدام طريقة خليط توزيع نافي بايز (Prakerja) "تصنيف الرأي العام لمنحة "براكيرجا (Mix Distribusi Naïve Bayes)**. بحث جامعي. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف: (١) الدكتور إيروان بودي سانتوسو، الماجستير، (٢) الدكتور كاحيو

الكلمات الرئيسية: منحة براكيرجا، تصنيف النصوص، خليط توزيع نافي بايز، تحليل المشاعر، تويتر.

جائحة كوفيد-١٩ الذي أطلقته الحكومة الإندونيسية لتقليل معدلات البطالة بعد (Prakerja) "أثار برنامج منحة "براكيرجا آراءً عامة متنوعة على وسائل التواصل الاجتماعي، وخاصة تويتر. وعلى الرغم من الحماس الشعبي العالي، ظهرت انتقادات مختلفة تتعلق بالفعالية، وشفافية نظام الاختيار، والعقبات التقنية التشغيلية. تهدف هذه الدراسة إلى تصنيف الرأي العام بشأن برنامج منحة براكيرجا" على تويتر لتقديم تقييم موضوعي للحكومة حول فعالية وشفافية النظام. وباستخدام نهج معالجة اللغات الطبيعية" (NLP)، التي تجمع بين توزيعي (Mix Distribusi Naïve Bayes) يطبق هذا البحث طريقة خليط توزيع نافي بايز للتعامل مع خصائص البيانات النصية المتنوعة. تشمل مراحل البحث جمع البيانات (Gaussian) و (Multinomial)، وتقليل الأبعاد من خلال (TF-IDF)، واستخراج الميزات باستخدام (preprocessing)، والمعالجة المسبقة (Truncated SVD). بينما تم تحسين النموذج باستخدام (RandomOverSampler)، ولعلاج عدم توازن البيانات، تم تطبيق تقنية (GridSearchCV). تم إجراء الاختبار بثلاثة سيناريوهات لتقسيم البيانات (٧٠:٣٠، ٨٠:٢٠، و٩٠:١٠) وقياسها من خلال معايير مصفوفة الارتباك (Confusion Matrix) مثل الدقة (Accuracy)، والضبط (Precision)، والاستدعاء (Recall)، ومقياس (F1-Score). ومن المتوقع أن تكون نتائج هذا التصنيف مرجعاً للجهات المعنية في فهم مشاعر المجتمع وتحسين سياسات الخدمة العامة في المستقبل.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Indonesia, sebagai negara berkembang, menghadapi berbagai tantangan kompleks, di antaranya adalah tingginya tingkat pengangguran. Kondisi ini terjadi ketika penawaran tenaga kerja melebihi permintaan, menciptakan surplus tenaga kerja yang tidak terserap di pasar kerja (Suhandi et al., 2021). Ketidakseimbangan ini diperparah oleh wabah Covid-19, yang memaksa banyak perusahaan melakukan Pemutusan Hubungan Kerja (PHK), merugikan lebih banyak pekerja. Dalam upaya menangani masalah pengangguran, pemerintah meluncurkan program beasiswa Prakerja, sebagai langkah untuk mendukung pemulihan ekonomi dan memberdayakan masyarakat. Program ini menawarkan pelatihan dan pembinaan kompetensi kepada individu yang mencari pekerjaan, mereka yang terkena dampak PHK, serta pelaku usaha mikro yang terdampak penurunan daya beli (Prakerja.go.id). Tujuan utamanya adalah untuk meningkatkan daya saing tenaga kerja dan memperkuat ekonomi nasional.

Saat program beasiswa Prakerja diperkenalkan, muncul sejumlah polemik dan kontroversi karena dianggap tidak efektif dalam menangani tantangan ketenagakerjaan yang timbul selama pandemi Covid-19. Kritik ini muncul karena kurangnya perencanaan yang matang dalam implementasi kebijakan ini. Masalah-masalah seperti kurangnya transparansi dalam sistem seleksi menyebabkan alokasi beasiswa yang tidak tepat sasaran, yang pada gilirannya dapat memperburuk situasi keuangan negara. Meskipun kontroversial, tingginya antusiasme masyarakat

terhadap program ini menunjukkan bahwa ada kebutuhan yang diakui oleh publik. Tantangan komunikasi juga menjadi kendala, karena batasan antara masyarakat dan pemerintah membuat sulit bagi individu untuk menyampaikan aspirasi mereka secara langsung. Namun, dengan perkembangan teknologi informasi dan media sosial, masyarakat kini dapat mengungkapkan pendapat dan aspirasi mereka dengan lebih mudah, termasuk melalui platform seperti Twitter. Ini memungkinkan partisipasi yang lebih luas dalam diskusi dan evaluasi terhadap kebijakan seperti beasiswa Prakerja. Pendapat atau komentar yang diberikan masyarakat pada media sosial Twitter ini dapat kita kelompokkan menjadi dua kategori utama, yaitu komentar positif dan komentar negatif. Dari banyaknya komentar serta pendapat dari masyarakat, dibutuhkan sebuah klasifikasi atau pengelompokan yang dapat digunakan sebagai gambaran bagi pemerintah untuk mengkaji kembali atau memperbaiki kebijakan beasiswa prakerja, mulai dari server yang sering bermasalah hingga pencairan intensif yang sering telat.

Klasifikasi teks merupakan salah satu dari bidang *Natural Language Processing* (NLP) yang mengelompokkan teks ke satu atau lebih kategori yang diinginkan pada dokumen, opini, atau kalimat yang bertujuan untuk mengetahui apakah mereka termasuk kategori positif, negatif, atau netral. Oleh karena itu melalui penerapan klasifikasi ini sebuah lembaga atau perusahaan dapat mengetahui tentang opini masyarakat terhadap suatu layanan atau produk secara spesifik. Klasifikasi teks perlu dilakukan agar suatu lembaga dapat melakukan evaluasi kepada kinerja, layanan atau produknya.

Semua inovasi dan kebijakan yang sudah dilakukan oleh pemerintah semata-mata untuk memperbaiki perekonomian, serta untuk memakmurkan kembali perekonomian masyarakat secara adil. Karena pemerintahan yang baik dan adil adalah pemerintahan yang setiap keputusannya bertujuan untuk membuat masyarakatnya menjadi lebih makmur dan sejahtera, dimana hal ini dijelaskan dalam Al-Qur'an ayat 25 Surat Al-Hadid:

لَقَدْ أَرْسَلْنَا رُسُلَنَا بِالْبَيِّنَاتِ وَأَنْزَلْنَا مَعَهُمُ الْكِتَابَ وَالْمِيزَانَ لِيَقُومَ النَّاسُ بِالْقِسْطِ ۖ وَأَنْزَلْنَاالْحَدِيدَ فِيهِ بَأْسٌ شَدِيدٌ وَمَنَافِعُ

لِلنَّاسِ وَلِيَعْلَمَ اللَّهُ مَنْ يَنْصُرُهُ وَرُسُلَهُ بِالْغَيْبِ ۚ إِنَّ اللَّهَ قَوِيٌّ عَزِيزٌ

"Sungguh, Kami benar-benar telah mengutus rasul-rasul Kami dengan bukti-bukti yang nyata dan Kami menurunkan bersama mereka kitab dan neraca (keadilan) agar manusia dapat berlaku adil. Kami menurunkan besi yang mempunyai kekuatan hebat dan berbagai manfaat bagi manusia agar Allah mengetahui siapa yang menolong (agama)-Nya dan rasul-rasul-Nya walaupun (Allah) tidak dilihatnya. Sesungguhnya Allah Maha Kuat lagi Maha Perkasa." (QS. Al-Hadid: 25).

Menurut Ustadz Marwan Hadidi bin Musa, M.Pd.I pada Hidayatul Insan bi Tafsiril Qur'an makna dari ayat tersebut yaitu keadilan baik dalam ucapan maupun dalam perbuatan. Agama yang yang dibawa para rasul berisi keadilan dalam perintah dan larangan, dan dalam bermu'amalah dengan makhluk, dalam jinayat, qishas, hudud, mawarits, dan lain-lain. Yakni dapat menegakkan agama Allah dan mewujudkan maslahat mereka yang begitu banyak. Ayat ini merupakan dalil bahwa para rasul semuanya sepakat dalam kaidah syara', yaitu menegakkan keadilan meskipun berbeda-beda gambaran keadilan itu sesuai situasi, kondisi dan zaman. Sehingga dapat disimpulkan bahwa pemerintah sebagai penentu keputusan dan

pemegang kuasa harus bersikap adil dan menegakkan kebenaran walaupun berbeda pandangan, karena keadilan itu harus sesuai situasi, kondisi dan zaman.

Pada dasarnya, klasifikasi komentar positif dan komentar negatif bisa saja dilakukan secara manual. Namun, ketika jumlah komentar yang akan dikelompokkan menjadi banyak dan bertambah terus menerus, maka waktu dan usaha yang dikeluarkan akan menjadi lebih banyak sehingga menjadi tidak efisien. Maka dari itu pada studi ini, peneliti akan memanfaatkan metode *Machine Learning* untuk mengelompokkan komentar agar menjadi lebih efisien. *Machine Learning* memiliki berbagai macam metode klasifikasi yang biasa digunakan, seperti *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Decision Tree*. Salah satu metode yang sering digunakan untuk mengklasifikasi adalah *Naïve Bayes Classifier*.

Metode *Naïve Bayes Classifier*, sebuah algoritma kategorisasi sederhana yang berbasis pada probabilitas, bekerja dengan menambahkan frekuensi kombinasi nilai pada set data untuk memperoleh probabilitas. Algoritma ini melakukan klasifikasi melalui dua proses yaitu training dan testing. Metode *Naïve Bayes* memiliki keunggulan yaitu mampu melakukan klasifikasi meskipun memiliki data training yang sedikit untuk estimasi parameternya, dan juga kecepatan dan akurasi yang tinggi ketika diterapkan pada dataset yang besar dan beragam (Fauzan & Hikmah, 2022). Jika dibandingkan dengan metode lainnya, *Naïve Bayes* termasuk kedalam metode yang cukup mudah dan tidak memerlukan banyak data training (Septianingrum & Irawan, 2021).

Pada penelitian sebelumnya yang dilakukan oleh Anggraini & Utami (2021) mengenai Klasifikasi Sentimen Masyarakat Terhadap Kebijakan Kartu Prakerja di

Indonesia Menggunakan Metode *Naïve Bayes*. Dataset yang digunakan sebanyak 2646 komentar dan dikelompokkan menjadi kelas positif, kelas netral, dan kelas negatif. Pada penelitian ini, didapatkan nilai akurasi sebesar 91.06%. Penelitian juga pernah dilakukan oleh Hendrastuty et al., (2021) mengenai Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode *Support Vector Machine* dan dikelompokkan menjadi positif, negatif, dan netral. Pengujian dilakukan menggunakan *confusion matrix* melalui library SVM, dari data uji sebanyak 302 data yang sebelumnya telah dilakukan proses klasifikasi dan menghasilkan akurasi 98.67%.

Berdasarkan pertimbangan masalah dan konteks yang ada, penelitian ini akan menggunakan metode *Mix Distribusi Naïve Bayes* dalam klasifikasi opini publik beasiswa prakerja menggunakan metode *Mix Distribusi Naïve Bayes*. Dengan adanya penelitian ini adalah untuk mengukur opini masyarakat Indonesia terkait kebijakan tersebut. Tujuan dari penelitian ini adalah untuk mempelajari penerapan metode *Mix Distribusi Naive Bayes* terhadap kebijakan kartu prakerja, serta mengevaluasi sejauh mana tingkat keakuratan klasifikasi yang dihasilkan oleh metode *Mix Distribusi Naive Bayes*. Karena tidak semua data berupa *Gaussian* saja tapi bisa saja berupa *Multinomial*, *Binomial*, dan sebagainya.

1.2 Rumusan Masalah

Bagaimana kinerja metode *Mix Distribution Naïve Bayes* dalam klasifikasi opini terhadap kebijakan Program Kartu Prakerja, ditinjau dari nilai *confusion matrix* (TP, TN, FP, FN) serta metrik evaluasi akurasi, presisi, *recall*, dan *F1-score*?

1.3 Batasan Masalah

1. Penggunaan data adalah data primer yaitu teks komentar pada media sosial Twitter pada keyword “prakerja”.
2. Data komentar yang diimplementasikan pada studi ini terdiri dari 1934 data komentar berbahasa Indonesia.

1.4 Tujuan Penelitian

Mengukur dan menganalisis kinerja klasifikasi yang dihasilkan oleh metode tersebut menggunakan *confusion matrix* (TP, TN, FP, FN) serta metrik evaluasi yaitu akurasi, presisi, *recall*, dan *F1-score*).

1.5 Manfaat Penelitian

Bagi pihak pemerintah yaitu Kementrian Ketenagakerjaan Republik Indonesia (KEMENAKER), diharapkan penelitian ini dapat dijadikan referensi untuk menjadi bahan evaluasi program beasiswa prakerja ini agar bisa ditingkatkan lagi menjadi lebih baik.

BAB II

STUDI PUSTAKA

2.1 Analisis Sentimen Opini Publik

Hasil dari penelitian sebelumnya yang pernah dilakukan oleh Sanusi et al., (2021) yang membahas tentang Prakerja juga menggunakan Algoritma *Recurrent Neural Network* dan mengambil data dari Twitter untuk melakukan pengelompokan data pengguna twitter menjadi positif, negatif dan netral. Pada penelitian ini dataset yang digunakan sebanyak 1934 tweet dan penelitian ini dimulai dari *crawling* data pada twitter setelah itu pelabelan data, lalu masuk pada tahap *preprocessing* dan pembagian data menjadi data training dan data testing, setelah itu masuk ke tahap pembuatan model, pelatihan model, pengetesan model dan terakhir impor model. Pada penelitian ini diperoleh nilai akurasi 95,66%.

Selanjutnya penelitian juga dilakukan oleh Hendrastuty et al., (2021). Hendrastuty juga meganalisis sentimen masyarakat terhadap Program Kartu Prakerja dengan mengambil data pada Twitter dengan metode *Support Vector Machine* (SVM). Tahapan pertama pada penelitian ini yaitu pengambilan data dari Twitter dengan kata kunci "Prakerja" kemudian akan diperoleh data komentar masyarakat terkait program kartu Prakerja yang rilis di masa pandemi. Setelah itu data yang diperoleh dilakukan *preprocessing* data. Setelah melalui pra proses data, kemudian data diberi label secara manual. Setelah itu dilakukan klasifikasi data menggunakan algoritma *Support Vector Machine* (SVM). Data yang sudah didapatkan akan diklasifikasikan menjadi tiga kategori kelas yaitu positif, negatif, dan netral. Pada penelitian ini dilakukan perbandingan dua kernel yaitu linear

dengan RBF. Hasil evaluasi yang dilakukan pada nilai akurasi kernel linear 98.67%, *precission* 98%, *recall* 99%, dan F1-Score 98%, sedangkan pada nilai akurasi kernel RBF 98.34%, *precission* 97%, *recall* 98%, F1-Score 98%, dapat disimpulkan bahwa sentimen masyarakat dari pengguna twitter terhadap program kartu prakerja dimasa pandemi lebih condong ke netral sebesar 98,34%.

Selanjutnya terdapat penelitian yang dilakukan oleh Anggraini & Utami (2021) mereka juga meneliti melakukan klasifikasi sentimen masyarakat terhadap program dan kebijakan Kartu Prakerja yang dilakukan pemerintah di Indonesia. Ada 5 langkah yang dilakukan pada penelitian ini yaitu langkah pertama melakukan pengumpulan data sekunder pada twitter. Selanjutnya melakukan tahap *preprocessing* dan pelabelan data. Setelah itu masuk pada tahap pemodelan, klasifikasi dan terakhir evaluasi model. Pada klasifikasi *Naïve Bayes*, digunakan partisi data menjadi 75% data training dan 25% data testing dari keseluruhan data sebanyak 2646 tweet. Pada penelitian ini diperoleh nilai akurasi sebesar 91,06%.

Pada penelitian yang pernah dilakukan juga oleh Novianti & Wibowo (2022) yang membahas tentang analisis sentimen masyarakat Indonesia pengguna Twitter terhadap kebijakan Program Kartu Prakerja yang dilakukan pemerintah ditengah masa pandemi Covid-19 dengan menggunakan Metode *Naïve Bayes Classifier*. Data yang telah dikumpulkan dari tweet pengguna twitter mengenai kartu prakerja yang diambil selama 3 bulan pada bulan November 2020 – bulan Januari 2021 didapatkan sebanyak 3.890 data tweet. Hasil klasifikasi menggunakan metode *Naïve Bayes Classifier* didapatkan akurasi klasifikasi pada data training nilai Gmean sebesar 80,1% dan nilai AUC sebesar 81,2%.

Pada penelitian yang dilakukan oleh Dey et al (2020) mereka melakukan perbandingan antara *Support Vector Machine* dan *Naïve Bayes Classifier* untuk melakukan analisa sentimen untuk ulasan pada produk yang ada pada website jual beli Amazon. Data yang digunakan adalah ulasan dari salah satu produk popular yang ada pada Amazon dan diperoleh sekitar 4000 dataset. Hasil dari perbandingan antara *Support Vector Machine* dan *Naïve Bayes Classifier* didapatkan hasil akurasi sebesar 84% untuk *Support Vector Machin*e dan 82,875% untuk *Naïve Bayes Classifier*.

Pada penelitian lain juga dilakukan oleh Farisi et al., (2019) yang melakukan analisis sentimen pada ulasan hotel dengan menggunakan dataset yang diambil dari *Data finitis's Business Database*. Dataset yang digunakan pada penelitian ini berjumlah 5000 kalimat, setelah itu kata dilabel manual yang menghasilkan 3946 kalimat berlabel 1 dan 1053 kalimat berlabel 0. Data yang sudah dilabeli akan masuk ke tahap *preprocessing*, *Multinomial Naïve Bayes*, dan dievaluasi menggunakan *K-Fold Cross Validation*. Pada penelitian ini didapatkan hasil akurasi paling optimal yaitu 91,4%.

Penelitian yang menggunakan metode *Multinomial Naïve Bayes Classifier* juga dilakukan oleh Ayogu (2020). Pada penelitiannya Ayogu membahas tentang penerapan *Multinomial Naïve Bayes* untuk melakukan klasifikasi pada dokumen *Yoruba Text*. Pada penelitian ini data yang digunakan sebanyak 500 contoh yang didapatkan dari sumber online dan *the bag-of-words* (BoW). Hasil akhir akurasi yang diperoleh pada penelitian ini yaitu 82%.

Selanjutnya Dewi et al., (2023) juga melakukan penelitian menggunakan metode *Multinomial Naïve Bayes* pada penelitiannya. Penelitian yang dilakukan oleh Dewi ini bertujuan untuk menganalisis teks untuk mengidentifikasi apakah teks tersebut merupakan sentimen positif atau negatif. Eksperimen ini didasarkan pada kumpulan data *Internet Movie Database* (IMDB), yang terdiri dari ulasan film dan label positif atau negatif yang terkait. Pada penelitian ini memperoleh akurasi sebesar 87,64%.

Pada studi selanjutnya dilakukan oleh Ramadhan & Adhinata, (2022). Studi yang dilakukan Ramadhan & Adhinata ini bertujuan untuk menganalisis sentimen publik terhadap vaksin COVID-19 dan mengambil dataset dari Twitter dalam bentuk cuitan dan retweet. Studi ini menggunakan metode *Gaussian Naïve Bayes* untuk melihat hasil klasifikasi analisis sentimen. Hasil yang diperoleh berdasarkan eksperimen membuktikan bahwa metode *Gaussian Naïve Bayes* dapat menghasilkan akurasi rata-rata sebesar 97,48% untuk setiap dataset vaksin yang digunakan.

Selanjutnya untuk metode yang menggunakan *Multinomial Naïve Bayes* juga dilakukan oleh Zulfikar et al., (2023). Tujuan dari penelitian ini adalah untuk menganalisis dokumen teks dari Twitter mengenai kebijakan publik dalam penanganan COVID19 yang saat itu sedang dilakukan. Dokumen teks diklasifikasikan ke dalam sentimen positif dan negatif dengan menggunakan *Multinomial Naive Bayes*. Hasil penelitian yang dilakukan oleh Zulfikar et al., (2023) difokuskan pada model atau pola yang dihasilkan oleh Algoritma *Multinomial Naive Bayes*. Hasil klasifikasi cuitan pengguna media sosial terhadap

kebijakan *new normal* menunjukkan hasil yang baik dengan nilai akurasi 90,25%. Setelah mengklasifikasikan cuitan pengguna media sosial mengenai kebijakan *new normal*, hasilnya menunjukkan bahwa lebih dari 70% setuju dan mendukung kebijakan *new normal*.

Berikut peneliti paparkan penelitian-penelitian terkait yang dijadikan sumber acuan peneliti untuk melakukan penelitian mengenai Klasifikasi Opini Publik Beasiswa Prakerja Menggunakan *Metode Mix Distribusi Naive Bayes*.

Tabel 2. 1 Penelitian Sebelumnya

NO	Peneliti	Metode	Hasil
1	(Novianti & Wibowo, 2022)	*Naïve Bayes Classifier*	G-mean 80,1% dan AUC 81,2%
2	(Sanusi et al., 2021)	*Recurrent Neural Network*	95,66%
3	(Hendrastuty et al., 2021)	*Support Vector Machine*	98,67%
4	(Anggraini & Utami, 2021)	*Naïve Bayes Classifier*	91,06%
5	(Dey et al., 2020)	*Support Vector Machine* & *Naïve Bayes Classifier*	84% dan 82,875%
6	(Farisi et al., 2019)	*Multinomial Naïve Bayes*	91,4%
7	(Ayogu, 2020)	*Multinomial Naïve Bayes*	82%

8	(Dewi et al., 2023)	*Multinomial Naïve Bayes*	87,64%
9	(Ramadhan & Adhinata, 2022)	*Gaussian Naïve Bayes*	97,48%
10	(Zulfikar et al., 2023)	*Multinomial Naïve Bayes*	90,25%

2.2 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF (Term Frequency–Inverse Document Frequency) merupakan salah satu metode pembobotan fitur yang paling umum dan efektif dalam pengolahan teks, khususnya pada bidang *information retrieval* dan *text classification*. Metode ini banyak digunakan karena mampu memberikan representasi numerik teks yang informatif serta terbukti menghasilkan kinerja yang baik pada berbagai penelitian analisis sentimen, terutama dari sisi akurasi dan *recall*. *TF-IDF* bekerja dengan menilai tingkat kepentingan suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus dokumen, sehingga kata-kata yang bersifat informatif akan memperoleh bobot yang lebih besar dibandingkan kata-kata yang bersifat umum. Secara konseptual, *TF-IDF* tersusun atas dua komponen utama, yaitu *Term Frequency* (TF) *dan Inverse Document Frequency* (IDF). TF mengukur seberapa sering suatu kata muncul dalam sebuah dokumen; semakin sering kata tersebut muncul, semakin besar nilai TF yang diperoleh, yang menunjukkan tingkat relevansi kata tersebut terhadap dokumen. Namun, frekuensi kemunculan saja belum cukup untuk menentukan kepentingan suatu kata, karena kata yang sangat sering muncul di banyak dokumen cenderung kurang informatif. Oleh karena itu,

IDF digunakan untuk mengurangi bobot kata-kata yang muncul di banyak dokumen dan meningkatkan bobot kata-kata yang relatif jarang muncul dalam korpus.

Dengan mengalikan nilai TF dan IDF, metode *TF-IDF* mampu menyeimbangkan antara kepentingan lokal suatu kata dalam dokumen dan kepentingan global kata tersebut dalam seluruh kumpulan dokumen. Kata-kata yang sering muncul dalam satu dokumen tetapi jarang muncul pada dokumen lain akan memiliki bobot *TF-IDF* yang tinggi, sedangkan kata-kata yang sering muncul di hampir semua dokumen akan memiliki bobot yang lebih rendah. Hal ini menjadikan *TF-IDF* efektif dalam mengekstraksi kata-kata kunci yang representatif untuk proses klasifikasi teks (Jeremy Andre et al., 2019).

Dalam penelitian ini, *TF-IDF* digunakan sebagai metode utama untuk merepresentasikan teks komentar publik ke dalam bentuk vektor numerik sebelum diproses oleh model *Mix Distribusi Naive Bayes*. Representasi *TF-IDF* memungkinkan model *Multinomial Naive Bayes* memanfaatkan pola frekuensi kata secara langsung, serta menyediakan dasar fitur yang dapat direduksi menggunakan *TruncatedSVD* untuk kebutuhan *Gaussian Naive Bayes*. Dengan demikian, penggunaan *TF-IDF* menjadi komponen penting dalam keseluruhan sistem analisis sentimen yang dibangun. Nilai pembobotan *TF-IDF* dapat dihitung menggunakan persamaan 2.1.

$$\mathrm{TF{-}IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D) \tag{2.1}$$

Keterangan:
$\mathrm{TF{-}IDF}_{(t,d,D)}$ = Pembobotan *TF-IDF*
$\mathrm{TF}_{(t,d)}$ = Jumlah kemunculan term dalam dokumen
$\mathrm{IDF}_{(t,D)}$ = Bobot *inverse* dalam nilai df

2.3 Multinomial Naïve Bayes

Multinomial Naive Bayes (MNB) merupakan salah satu algoritma klasifikasi berbasis probabilistik yang banyak digunakan dalam *text classification* dan *information retrieval*. Metode ini didasarkan pada Teorema Bayes, dengan asumsi *naïve* bahwa setiap fitur (kata atau token) bersifat saling bebas (conditional independence) terhadap fitur lainnya jika kelas dokumen telah diketahui. Meskipun asumsi ini bersifat sederhana, *Multinomial Naive Bayes* terbukti efektif dan efisien dalam menangani data teks berdimensi tinggi, terutama ketika direpresentasikan dalam bentuk frekuensi kata atau bobot *TF-IDF*.

Berbeda dengan varian *Naive Bayes* lainnya, *Multinomial Naive Bayes* secara eksplisit memodelkan distribusi frekuensi kemunculan kata dalam suatu dokumen. Setiap dokumen dipandang sebagai hasil pengambilan kata-kata dari suatu distribusi *Multinomial* yang bergantung pada kelas tertentu. Probabilitas sebuah dokumen termasuk ke dalam suatu kelas dihitung berdasarkan peluang kemunculan setiap kata dalam dokumen tersebut pada kelas yang bersangkutan. Dengan demikian, kata-kata yang sering muncul dalam dokumen tertentu dan jarang muncul pada kelas lain akan memberikan kontribusi yang lebih besar terhadap keputusan klasifikasi.

Dalam praktiknya, *Multinomial Naive Bayes* menghitung probabilitas posterior suatu kelas berdasarkan hasil perkalian probabilitas kata-kata yang muncul dalam dokumen, yang kemudian dinormalisasi menggunakan prior kelas. Untuk mengatasi masalah probabilitas nol (zero probability problem), metode ini umumnya dilengkapi dengan teknik *smoothing*, seperti *Laplace smoothing*,

sehingga kata yang tidak pernah muncul pada data pelatihan tetap memiliki probabilitas yang kecil namun tidak bernilai nol (Winahyu & Suharjo, 2021).

Dalam penelitian ini, *Multinomial Naive Bayes* digunakan sebagai salah satu komponen utama dalam model *Mix Distribusi Naive Bayes*. MNB bekerja langsung pada representasi fitur *TF-IDF* untuk menangkap pola frekuensi kata dan *n-gram* yang mencerminkan kecenderungan sentimen pada komentar publik. Keunggulan MNB dalam memanfaatkan informasi berbasis kata menjadikannya sangat sesuai untuk analisis sentimen pada teks pendek dan informal, seperti komentar YouTube berbahasa Indonesia.

2.4 Gaussian Naïve Bayes

Gaussian Naive Bayes (GNB) merupakan salah satu algoritma klasifikasi berbasis probabilistik dalam *machine learning* yang didasarkan pada Teorema Bayes dengan asumsi *naïve* bahwa setiap fitur bersifat saling independen (conditional independence) terhadap fitur lainnya, jika kelas target telah diketahui. Berbeda dengan *Multinomial Naive Bayes* yang memodelkan data diskret, *Gaussian Naive Bayes* dirancang untuk menangani fitur bernilai kontinu dengan mengasumsikan bahwa distribusi setiap fitur pada masing-masing kelas mengikuti *Distribusi Gaussian* (Normal).

Pada *Gaussian Naive Bayes*, probabilitas kemunculan suatu fitur dihitung menggunakan fungsi kepadatan probabilitas distribusi normal yang ditentukan oleh dua parameter utama, yaitu rata-rata (mean) dan varians (variance) dari fitur tersebut pada setiap kelas. Dengan asumsi ini, setiap fitur berkontribusi terhadap probabilitas posterior kelas melalui distribusi *Gaussian*-nya masing-masing.

Probabilitas posterior suatu kelas kemudian diperoleh dari kombinasi probabilitas seluruh fitur dan probabilitas prior kelas, dan keputusan klasifikasi ditentukan berdasarkan kelas dengan nilai probabilitas posterior tertinggi.

Asumsi distribusi *Gaussian* memungkinkan *Gaussian Naive Bayes* bekerja secara efektif pada data yang memiliki karakteristik kontinu dan berskala numerik. Meskipun asumsi independensi dan normalitas fitur bersifat sederhana, *Gaussian Naive Bayes* dikenal memiliki efisiensi komputasi yang tinggi serta mampu memberikan performa yang cukup baik pada berbagai permasalahan klasifikasi, khususnya ketika jumlah data relatif terbatas (Hasibuan & Heriyanto, 2022).
Dalam penelitian ini, *Gaussian Naive Bayes* digunakan sebagai bagian dari model *Mix Distribusi Naive Bayes*. GNB tidak langsung diterapkan pada vektor *TF-IDF* yang bersifat jarang (sparse), melainkan pada representasi fitur padat (dense) hasil reduksi dimensi menggunakan *TruncatedSVD*. Reduksi dimensi ini bertujuan untuk menghasilkan fitur laten yang lebih sesuai dengan asumsi distribusi *Gaussian*, sehingga kinerja GNB menjadi lebih optimal. Dengan demikian, *Gaussian Naive Bayes* berperan dalam menangkap variasi semantik dan pola laten pada data teks yang tidak sepenuhnya dapat ditangkap oleh model berbasis frekuensi kata.

2.5 GridSearch CV

GridSearchCV merupakan metode optimasi hiperparameter yang digunakan untuk mencari kombinasi parameter terbaik dari suatu model machine learning secara sistematis. *GridSearchCV* bekerja dengan mendefinisikan sekumpulan nilai kandidat untuk setiap hiperparameter, kemudian mengevaluasi seluruh kombinasi parameter tersebut menggunakan teknik cross-validation. Pendekatan ini bertujuan

untuk memperoleh konfigurasi model yang memberikan kinerja paling optimal berdasarkan metrik evaluasi yang telah ditentukan.

Dalam *GridSearchCV*, data latih dibagi menjadi beberapa lipatan (k-fold), di mana pada setiap iterasi sebagian data digunakan untuk melatih model dan sebagian lainnya digunakan untuk memvalidasi kinerja model. Proses ini diulang hingga seluruh lipatan digunakan sebagai data validasi. Nilai kinerja dari setiap kombinasi hiperparameter kemudian dirata-ratakan untuk menghasilkan estimasi performa yang lebih stabil dan tidak bergantung pada satu pembagian data tertentu. Dengan demikian, *GridSearchCV* membantu mengurangi risiko overfitting terhadap data latih.

Pada penelitian ini, *GridSearchCV* diterapkan di dalam sebuah Pipeline yang mencakup seluruh tahapan pemodelan, mulai dari ekstraksi fitur *TF-IDF*, reduksi dimensi menggunakan *TruncatedSVD*, hingga klasifikasi menggunakan *Mix Distribusi Naive Bayes*. Hiperparameter yang dituning meliputi pengaturan *TF-IDF* (seperti min_df, max_df, max_features, dan ngram_range), jumlah komponen *SVD* (n_components), parameter smoothing pada *Multinomial Naive Bayes*, serta nilai α sebagai bobot pencampuran antara *Multinomial Naive Bayes* dan *Gaussian Naive Bayes*. Seluruh proses tuning dilakukan menggunakan skema *k-fold cross-validation* dengan metrik evaluasi utama F1-macro, sehingga kinerja model pada kedua kelas sentimen dapat diperhatikan secara seimbang.

Hasil dari *GridSearchCV* berupa kombinasi hiperparameter terbaik (best_params_) dan model terbaik (best_estimator_), yang kemudian digunakan sebagai model final pada tahap pengujian. Dengan penggunaan *GridSearchCV*,

pemilihan hiperparameter dalam penelitian ini dilakukan secara objektif, terkontrol, dan dapat direproduksi, sehingga hasil analisis sentimen yang diperoleh memiliki tingkat keandalan yang lebih tinggi (Prayoga et al., 2025).

2.6 Truncated Singular Value Decomposition

Truncated Singular Value Decomposition (TruncatedSVD) merupakan teknik reduksi dimensi yang digunakan untuk memetakan data berdimensi tinggi ke dalam ruang berdimensi lebih rendah dengan tetap mempertahankan informasi penting dari data asli. *TruncatedSVD* merupakan varian dari metode *Singular Value Decomposition* (SVD) yang hanya mempertahankan sejumlah komponen utama (top-k components), sehingga lebih efisien secara komputasi dan cocok untuk data berukuran besar.

Dalam konteks pengolahan teks, *TruncatedSVD* sering digunakan pada matriks fitur *TF-IDF* yang bersifat jarang (sparse) dan berdimensi sangat tinggi. Dengan menerapkan *TruncatedSVD*, matriks *TF-IDF* diproyeksikan ke dalam representasi fitur yang lebih padat (dense) dan berdimensi lebih rendah. Proses ini dikenal pula sebagai *Latent Semantic Analysis* (LSA), karena mampu menangkap struktur semantik laten di balik kemunculan kata-kata dalam dokumen.

Secara intuitif, *TruncatedSVD* mengidentifikasi pola utama atau "topik laten" yang merepresentasikan hubungan antar kata dan dokumen. Setiap komponen hasil reduksi dimensi merepresentasikan kombinasi linier dari fitur-fitur asli yang paling berkontribusi terhadap variasi data. Dengan demikian, *TruncatedSVD* tidak hanya berfungsi untuk mengurangi dimensi data, tetapi juga membantu menghilangkan noise dan redundansi fitur yang tidak relevan.

Dalam penelitian ini, *TruncatedSVD* digunakan sebagai tahap perantara sebelum penerapan *Gaussian Naive Bayes* (GNB). Hal ini dilakukan karena GNB mengasumsikan bahwa fitur bersifat kontinu dan mengikuti distribusi *Gaussian*, sedangkan fitur *TF-IDF* bersifat diskret dan jarang. Dengan mereduksi *TF-IDF* menjadi fitur laten yang padat dan kontinu, *TruncatedSVD* membuat data menjadi lebih sesuai dengan asumsi distribusi *Gaussian*, sehingga kinerja GNB dapat meningkat.

Jumlah komponen *TruncatedSVD* (n_components) ditentukan melalui proses *GridSearchCV* untuk memperoleh keseimbangan antara representasi informasi dan efisiensi komputasi. Dengan integrasi *TruncatedSVD* dalam model *Mix Distribusi Naive Bayes*, sistem mampu memanfaatkan informasi semantik laten dari teks sekaligus meningkatkan stabilitas dan akurasi klasifikasi sentimen.

BAB III

DESAIN DAN IMPLEMENTASI SISTEM

3.1 Tahapan Penelitian

Penelitian ini bertujuan mengklasifikasikan sentimen opini publik terkait program Prakerja menggunakan model *Mix Distribusi Naïve Bayes* yang menggabungkan *Multinomial Naive Bayes* (MNB) pada *TF-IDF* dan *Gaussian Naive Bayes* (GNB) pada fitur yang direduksi dengan *TruncatedSVD*.

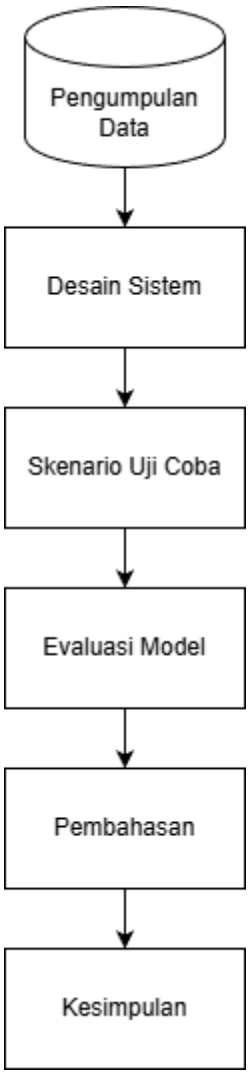
Prosedur penelitian dimulai dengan pengumpulan data, yaitu komentar-komentar publik yang diperoleh dari YouTube dan portal berita. Data kemudian dibersihkan dengan menghapus kolom yang tidak relevan, baris kosong, dan duplikat. Selanjutnya, dilakukan pra-proses teks khusus untuk Bahasa Indonesia yang mencakup normalisasi, penghapusan URL, angka, simbol, *stopword*, dan *stemming* menggunakan pustaka Sastrawi.

Pada tahap desain sistem, data yang sudah dibersihkan diubah menjadi representasi fitur menggunakan *TF-IDF*. Model *Mix Distribusi Naïve Bayes* dibangun dalam sebuah *pipeline* yang mencakup dua cabang: MNB yang bekerja pada *TF-IDF* dan GNB yang memanfaatkan representasi padat hasil reduksi dimensi dengan SVD.

Pada tahap skema uji coba, data dibagi menjadi dua bagian menggunakan *stratified split*, dan model kemudian diuji menggunakan *GridSearchCV* untuk menyesuaikan hiperparameter seperti α, jumlah komponen SVD, dan parameter *TF-IDF*. Skor F1-macro digunakan sebagai metrik evaluasi utama.

Selanjutnya, pada tahap evaluasi model, hasil prediksi diuji dengan menghitung *accuracy*, *precision*, *recall*, dan *F1-score* dalam skema mikro, makro, dan berbobot. *Confusion matrix* juga digunakan untuk menganalisis kesalahan klasifikasi.

Terakhir, pada pembahasan dan kesimpulan, hasil dari evaluasi model dibahas untuk mengidentifikasi performa model dan memberikan rekomendasi untuk pengembangan lebih lanjut.



Gambar 3. 1 Prosedur Penelitian

3.2 Pengumpulan Data

Pada penelitian ini menggunakan data yang berasal dari komentar Youtube dan portal berita online. Dataset ini terdiri dari 3 kolom yitu author, comments, dan published_at. Dataset ini kita perolah dalam format excel. Berikut merupakan 5 contoh dari 1934 data eksperimen yang digunakan pada tabel 3.1 dibawah ini.

Tabel 3. 1 Data Eksperimen

Author	Comments	Published_at
@rajakepo2867	Listrik pulsa pun mahal	3 tahun yang lalu
@elangsumatera1446	Jadi kalau seperti itu pendapat dari PLN kesempatan dong dengan covit koropsi	3 tahun yang lalu
@didikahmad9328	Jgn menghutang kalau rakyat kecil jadi tumbalnya	3 tahun yang lalu
@jhosjhos3606	Amburadul	3 tahun yang lalu
@abifatihsyakira478	Di rumah sy 2 bulan bertutut2 faktanya memang naik	3 tahun yang lalu
@usufas9809	Kabar prakerja 2023 mana?	1 tahun yang lalu
@saktiaja5109	Saya kena imbas pndemi ga dapat bantuan berbentuk apapun.pas pandemi berhenti kerja nyatanya ga dapat apapun	1 tahun yang lalu
@fahrunfth8206	mahasiswa yang dapat bidikmisi apakah bisa daftar ga ya? kira2 ada sanksi nya ga	3 tahun yang lalu
@jaqlienelsy	Udah 3 kali daftar gagal terus pada hal butuh banget saat hrs beerhe	3 tahun yang lalu
@tsuwaibahaslamiyah5417	akhirnya terjawab, sangat membantu ÄŸÅ¸Ëœâ€ž	3 tahun yang lalu
@tobirludy1603	Untuk apa pak pake platpon digital,mlah bikin tambah ribet	3 tahun yang lalu

3.3 Desain Sistem

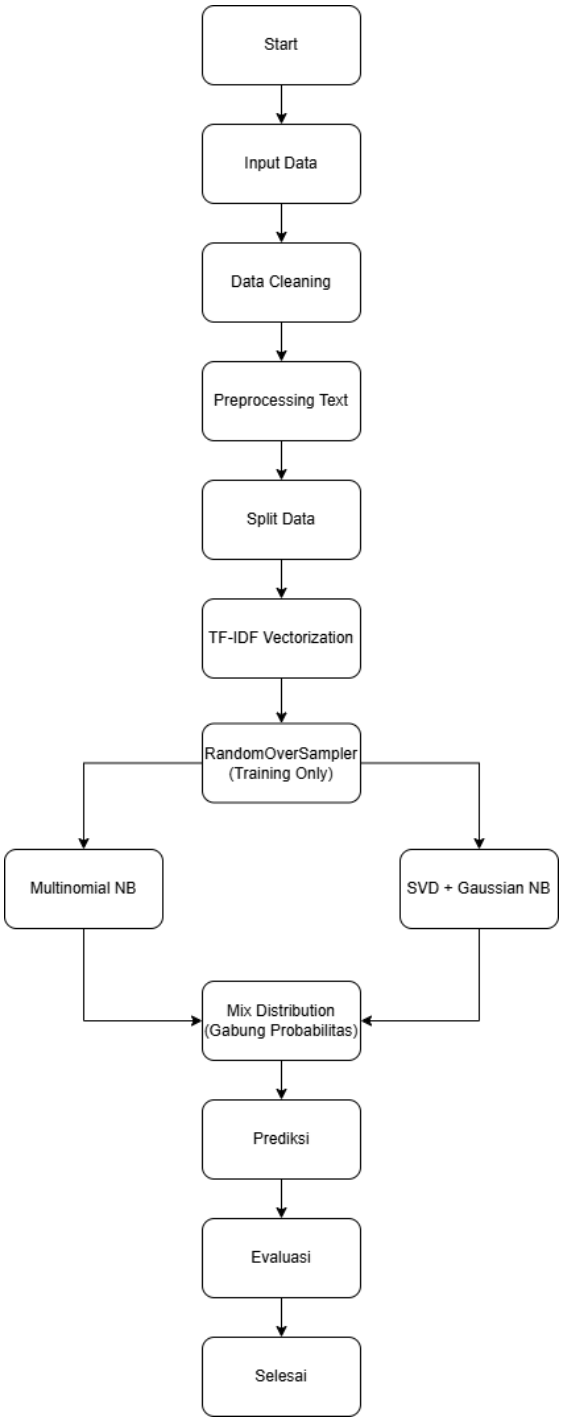
Penelitian ini bertujuan untuk mengklasifikasikan sentimen opini publik terkait program Prakerja menjadi dua kelas, yaitu Positif dan Negatif, menggunakan pendekatan *Mix Distribusi* yang menggabungkan *Multinomial Naive Bayes* (MNB) pada *TF-IDF* dan *Gaussian Naive Bayes* (GNB) pada representasi fitur yang direduksi menggunakan *TruncatedSVD*. Sistem dimulai dengan pengumpulan data berupa komentar-komentar YouTube yang sudah dilabeli sentimen. Data tersebut kemudian melalui tahapan *pre-processing* teks yang meliputi normalisasi huruf kecil, penghapusan URL, mention, angka, tanda baca, *stopwords*, dan *stemming* menggunakan pustaka Sastrawi.

Setelah data dibersihkan, representasi fitur dihasilkan menggunakan *TF-IDF Vectorizer*, yang mengubah teks menjadi vektor bobot berdasarkan frekuensi kata dan kelangkaan kata di dalam korpus. Pada tahap pemodelan, sistem membangun *Mix Distribusi* yang memiliki dua cabang klasifikasi. Cabang pertama adalah MNB, yang bekerja langsung pada representasi *TF-IDF* dan menangkap pola frekuensi kata serta *n-gram*. Cabang kedua adalah GNB, yang menerima representasi padat (dense) hasil reduksi dimensi menggunakan *TruncatedSVD*. GNB berfungsi untuk menangkap pola semantik atau tema laten dari teks. Kedua model ini menghasilkan probabilitas posterior yang kemudian digabungkan menggunakan parameter pencampur α, dengan rumus:

$$\log P_{mix}(y = c|x) = \alpha \log P_{MNB}(y = c|x) + (1 - \alpha) \log P_{GNB}(y = c|x) \quad (3.1)$$

Nilai α yang berkisar antara 0 hingga 1 ini mengatur kontribusi relatif dari kedua model. Pengaturan nilai α, parameter *TF-IDF*, dan jumlah komponen SVD

dilakukan menggunakan *GridSearchCV* dengan skema *k-fold cross-validation* untuk mencari kombinasi terbaik dari hiperparameter. Untuk visualisasinya bisa dilihat seperti pada gambar 3.2.



Gambar 3. 2 Blok Diagram Desain Sistem

3.4 Preprocessing

Tahap *preprocessing* pada program ini diawali dengan pemuatan berkas CSV komentar YouTube beserta label sentimennya. Saat membaca data, skrip menangani kemungkinan delimeter yang tidak konsisten serta menghapus kolom "Unnamed" (sisa indeks) dan baris kosong agar struktur dataset bersih. Kolom teks kemudian dinormalisasi menjadi huruf kecil (lowercase) dan dibersihkan dengan fungsi regex: menghapus URL, username/mention (@…), tagar (#…), angka, serta tanda baca/simbol yang tidak informatif dan mengompres spasi ganda. Hasilnya disimpan pada kolom comment_cleaned sebagai teks yang telah terstandardisasi.

Berikutnya, teks dibawa ke level token menggunakan tokenisasi, lalu dilakukan penghapusan *stopword* Bahasa Indonesia agar kata-kata fungsi yang tidak membawa makna sentimen (mis. yang, dan, di, ke, para) dihilangkan. Untuk konsistensi morfologi, program menerapkan *stemming* Sastrawi sehingga kata-kata kembali ke bentuk dasar (mis. berbantu → bantu, kebijakan → bijak). Rangkaian ini menghasilkan kolom *stemmed* yang berisi representasi kata dasar tanpa *stopword*; inilah yang menjadi masukan utama ke tahap ekstraksi fitur.

Sebagai langkah akhir sebelum pemodelan, label sentimen dipastikan konsisten (Positif/Negatif) dan dataset dipecah secara *stratified* menjadi data latih dan data uji agar proporsi kelas tetap seimbang. Dengan demikian, *pipeline preprocessing* menghasilkan teks yang bersih, terstandardisasi, dan tereduksi noise, sehingga fitur *TF-IDF* dan komponen laten (hasil *TruncatedSVD* untuk cabang *Gaussian*) yang dibentuk pada tahap berikutnya lebih representatif terhadap sinyal sentimen yang sebenarnya dengan menggunakan perintah berikut.

```
text = text.lower()
text = re.sub(r"http\S+|www\.\S+", " ", text)
text = re.sub(r"[@#]\w+", " ", text)
text = replace_emoji(text)
text = re.sub(r"[^0-9a-zA-ZÀ-ÿ_\s]", " ", text)
text = reduce_repeat(text)
text = normalize_slang(text)
text = re.sub(r"\s+", " ", text).strip()
text = add_negation_bigrams(text)
```

Pada tahap *preprocessing*, teks dibersihkan melalui fungsi clean_id(). Prosesnya dimulai dengan mengubah teks menjadi *lowercase*, kemudian menghapus URL, *mention*, dan *hashtag*. Setelah itu emoji diubah menjadi token sentimen, karakter khusus dibuang, huruf berulang dikurangi, kata slang dinormalisasi, spasi dirapikan, dan terakhir dibentuk negation bigram agar makna negasi tetap terjaga.

3.5 Ekstraksi Fitur Menggunakan TF-IDF

Setiap dokumen atau komentar d direpresentasikan sebagai vektor bobot kata menggunakan TF–IDF. Bobot TF–IDF menggabungkan *term frequency* (TF) sebagai ukuran intensitas kemunculan term dalam dokumen dan *inverse document frequency* (IDF) sebagai ukuran kelangkaan term pada seluruh korpus. Pendekatan ini membuat kata yang sering muncul pada satu dokumen tetapi jarang muncul pada dokumen lain menjadi lebih informatif untuk proses klasifikasi menggunakan code berikut.

```
vect = TfidfVectorizer(
    ngram_range=(1,2),
    min_df=3,
    max_df=0.85,
    max_features=5000,
    sublinear_tf=True
)
X_tfidf = vect.fit_transform(X_text)
```

Pada tahap ekstraksi fitur, teks hasil preprocessing diubah menjadi bentuk numerik menggunakan TF-IDF. Metode ini memberi bobot lebih tinggi pada kata yang penting dalam suatu dokumen, tetapi tidak terlalu umum di seluruh dokumen. Output TF-IDF berupa matriks fitur yang kemudian menjadi input model klasifikasi. Untuk rumus *TF-IDF* dapat dilihat pada persamaan 3.2.

- TF:

$$TF_{(t,d)} = \frac{f_{(t,d)}}{\sum_k f_{(k,d)}} \tag{3.2}$$

Keterangan:

t : kata (term) yang sedang dihitung

d : dokumen (kalimat/komentar)

$f_{(t,d)}$: frekuensi kemunculan kata t dalam d

$\sum_k f_{(k,d)}$: total semua kata dalam dokumen d

- IDF:

$$IDF_{(t)} = \log\left(\frac{N}{df_t}\right) \tag{3.3}$$

Keterangan:

N : jumlah total dokumen dalam dataset

df_t : jumlah dokumen yang mengandung kata t

- TF-IDF:

$$TFIDF_{(t,d)} = TF_{(t,d)} \times IDF_{(t)} \tag{3.4}$$

Matriks inilah yang kemudian digunakan sebagai fitur numerik untuk proses klasifikasi menggunakan Naive Bayes. Berdasarkan hasil perhitungan TF-IDF pada

Tabel 3.8, setiap kata pada dokumen memiliki bobot yang berbeda bergantung pada frekuensi kemunculannya dalam dokumen dan kelangkaannya pada seluruh dokumen. Kata yang muncul pada lebih sedikit dokumen memiliki nilai IDF yang lebih tinggi sehingga menghasilkan bobot TF-IDF yang lebih besar. Matriks TF-IDF ini kemudian digunakan sebagai representasi numerik dari data teks sebelum diproses oleh model klasifikasi sentimen pada tahap selanjutnya.

3.6 Reduksi Dimensi Menggunakan Truncated SVD

Vektor TF–IDF berdimensi tinggi dan bersifat sparse, sehingga dilakukan reduksi dimensi menggunakan Truncated Singular Value Decomposition (SVD). SVD memfaktorkan matriks TF–IDF A menjadi tiga matriks U, Σ, dan V^T. Selanjutnya dipilih k komponen singular terbesar (misalnya $k = 300$) untuk membentuk aproksimasi A_k yang mempertahankan informasi dominan namun berukuran lebih kecil. Hasil reduksi ini menghasilkan fitur kontinu berdimensi k yang lebih efisien untuk model berbasis distribusi kontinu seperti Gaussian Naïve Bayes. Berikut untuk persamaannya:

- SVD (Singular Value Decomposition)

$$A = U\Sigma V^T \tag{3.5}$$

- Reduksi dimensi (Truncated SVD):

$$A_k = U_k \Sigma_k V_k^T \tag{3.6}$$

Keterangan:

A : Matriks TF-IDF awal (dokumen × term)

U, V : matriks ortogonal (singular vectors)

Σ: matriks diagonal singular values

k : Jumlah komponen yang dipilih (disini 300)

3.7 Klasifikasi dengan Multinomial Naïve Bayes (MNB)

Untuk cabang pertama, digunakan Multinomial Naïve Bayes yang memodelkan peluang kemunculan kata w_i pada kelas c melalui estimasi likelihood dengan smoothing Laplace. Selanjutnya prediksi dilakukan dengan memilih kelas yang memaksimalkan jumlah skor log dari prior kelas dan kontribusi fitur (berbobot TF–IDF) terhadap likelihood kata pada kelas tersebut. Untuk persamaannya dapat dilihat sebagai berikut.

- Likelihood kata dengan Laplace Smoothing:

$$P(w_i|y = c) = \frac{N_{ic} + \alpha}{\sum_j N_{jc} + \alpha|V|} \tag{3.7}$$

Keterangan:

w_i : kata ke-I dalam vocabulary

$y = c$: kelas c (positif/negatif)

N_{ic} : jumlah kata w_i pada kelas c

α : parameter smoothing Laplace

$|V|$: jumlah fitur kata

$\sum_j N_{jc}$: total frekuensi semua kata dalam kelas c

- Prediksi kelas (skor log):

$$\hat{y} = argmax[\log P(y = c) + \sum_i x_i \log P(w_i|y = c)] \tag{3.8}$$

Keterangan:

$\hat{y}$: kelas hasil prediksi

$(y = c)$: probabilitas awal (prior) kelas c

x_i : bobot TF-IDF kata ke-i

$P(w_i|y = c)$: peluang kata pada kelas c

3.8 Klasifikasi dengan Gaussian Naïve Bayes (GNB)

Untuk cabang kedua, digunakan Gaussian Naïve Bayes pada fitur hasil SVD yang bersifat kontinu. Model mengasumsikan setiap fitur x_i pada kelas c mengikuti distribusi Normal dengan parameter rata-rata μ_{ic} dan varians σ_{ic}^2. Nilai probabilitas kondisional dihitung dengan fungsi densitas Gaussian. Berikut untuk bentuk persamaannya.

$$P(x_i|y = c) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left(-\frac{(x_i-\mu_{ic})^2}{2\sigma_{ic}^2}\right) \quad (3.9)$$

Keterangan:

x_i : nilai fitur ke-i hasil SVD

$y = c$: kelas c

μ_{ic} : rata-rata fitur i pada kelas c

σ_{ic}^2 : varians fitur i pada kelas c

π : konstanta 3.1415926

exp (): fungsi eksponensial

Pada penelitian ini, Gaussian Naive Bayes digunakan setelah proses ekstraksi fitur TF-IDF dan reduksi dimensi menggunakan TruncatedSVD. Oleh karena itu, fitur yang diproses oleh Gaussian Naive Bayes berupa data kontinu hasil transformasi SVD. Probabilitas setiap fitur pada suatu kelas dihitung menggunakan distribusi Gaussian berdasarkan parameter mean dan varians masing-masing kelas. Selanjutnya, skor log posterior dihitung dengan menjumlahkan log prior probability dan log likelihood dari seluruh fitur. Kelas dengan skor log posterior terbesar dipilih sebagai hasil klasifikasi akhir.

Pada penelitian ini, Gaussian Naive Bayes digunakan setelah proses ekstraksi fitur TF-IDF dan reduksi dimensi menggunakan TruncatedSVD. Oleh karena itu, fitur yang diproses oleh Gaussian Naive Bayes berupa data kontinu hasil transformasi SVD. Probabilitas setiap fitur pada suatu kelas dihitung menggunakan distribusi Gaussian berdasarkan parameter mean dan varians masing-masing kelas. Selanjutnya, skor log posterior dihitung dengan menjumlahkan log prior probability dan log likelihood dari seluruh fitur. Kelas dengan skor log posterior terbesar dipilih sebagai hasil klasifikasi akhir.

3.9 Mix Distribusi Naïve Bayes

Pada penelitian ini, tugas klasifikasi sentimen difokuskan pada dua kelas, yaitu positif dan negatif. Untuk meningkatkan reliabilitas prediksi, penelitian tidak hanya menggunakan satu model tunggal, melainkan menggabungkan dua pendekatan Naïve Bayes yang bekerja pada representasi fitur berbeda. Strategi ini didasarkan pada pertimbangan bahwa teks ulasan/komentar sering mengandung variasi bahasa, sinonim, serta gaya penulisan yang menyebabkan satu jenis representasi fitur saja tidak selalu cukup kuat menjelaskan seluruh pola data.

Cabang pertama menggunakan Multinomial Naïve Bayes (MNB) pada fitur TF–IDF. Pada konteks klasifikasi biner, MNB cenderung efektif ketika terdapat kata-kata yang menjadi penanda kuat sentimen (misalnya kata yang konsisten muncul pada komentar positif atau negatif). TF–IDF menonjolkan term yang relatif khas pada dokumen tertentu dan tidak terlalu umum di seluruh korpus. Dengan demikian, MNB berbasis TF–IDF dapat menangkap sinyal leksikal yang eksplisit,

yaitu “kata apa yang muncul” dan seberapa penting kata tersebut pada sebuah komentar.

Namun, TF–IDF menghasilkan vektor yang berdimensi tinggi dan bersifat sparse. Kondisi ini membuat prediksi bisa sensitif terhadap perbedaan pemilihan kata, penggunaan sinonim, serta variasi cara menyampaikan opini. Untuk mengurangi dampak variasi tersebut, penelitian ini juga menggunakan cabang kedua, yaitu Gaussian Naïve Bayes (GNB) pada fitur hasil Truncated SVD. Truncated SVD merangkum informasi TF–IDF menjadi fitur yang lebih padat (dense) dan kontinu, yang dapat merepresentasikan struktur laten dari komentar, misalnya kecenderungan topik atau pola umum yang tidak selalu muncul dalam bentuk kata yang sama. Pada klasifikasi biner, representasi ini berguna ketika sentimen tersirat melalui rangkaian kata yang berbeda-beda namun memiliki makna serupa, sehingga pendekatan berbasis fitur laten dapat membantu menjaga stabilitas prediksi.

Karena kedua model memproses informasi dari sudut pandang yang berbeda, penelitian ini menerapkan pendekatan Mix Distribusi dengan menggabungkan skor prediksi dari MNB dan GNB. Penggabungan tidak dilakukan secara langsung pada probabilitas mentah, melainkan menggunakan ruang log-probabilitas. Pertimbangan penggunaan log-probabilitas adalah untuk menjaga stabilitas perhitungan dan memastikan skor dari dua model dapat dibandingkan dan digabungkan secara lebih konsisten, terutama pada kasus klasifikasi teks yang sering menghasilkan nilai probabilitas sangat kecil.

Dalam proses penggabungan, diperkenalkan parameter bobot α dengan rentang [0,1] untuk mengatur kontribusi masing-masing model. Secara interpretatif, nilai α menunjukkan tingkat dominansi MNB terhadap keputusan akhir. Jika α mendekati 1, keputusan akhir akan lebih dipengaruhi oleh MNB (TF–IDF). Sebaliknya, jika α mendekati 0, keputusan akhir akan lebih dipengaruhi oleh GNB (SVD). Nilai α yang berada di antara keduanya mencerminkan kompromi yang menggabungkan kekuatan sinyal kata-kata spesifik (MNB) dan kekuatan representasi pola global yang lebih padat (GNB). Dengan demikian, prediksi akhir diharapkan lebih kuat karena tidak bergantung pada satu jenis fitur saja.

Secara operasional pada klasifikasi biner, untuk setiap komentar masukan, sistem menghitung dua skor yaitu, skor untuk kelas positif dan skor untuk kelas negatif. Skor dari MNB dan GNB kemudian digabungkan menggunakan skema pembobotan berbasis log-probabilitas (lihat Persamaan 3.9). Setelah skor gabungan untuk kedua kelas diperoleh, keputusan akhir ditentukan dengan memilih kelas yang memiliki skor gabungan lebih tinggi. Mekanisme ini memastikan bahwa keputusan positif atau negatif merupakan hasil pertimbangan gabungan dari dua representasi fitur yang berbeda.

Agar hasil penelitian dapat dipertanggungjawabkan secara ilmiah, nilai α tidak dipilih secara subjektif. Parameter ini diperlakukan sebagai hiperparameter yang ditentukan melalui proses validasi, yaitu dengan menguji beberapa kandidat nilai α dan memilih nilai yang menghasilkan performa terbaik pada data validasi berdasarkan metrik evaluasi (misalnya F1-score atau accuracy). Dengan cara ini,

pembobotan Mix Distribusi bukan sekadar keputusan intuitif, melainkan keputusan yang didukung hasil evaluasi empiris.

Adapun ketentuan yang dijaga pada tahap penggabungan ini adalah: (1) kedua model harus menghasilkan skor untuk dua kelas yang sama (positif dan negatif) pada komentar input yang sama; (2) α berada pada interval [0,1] agar interpretasi bobot tetap konsisten; dan (3) penggabungan dilakukan di ruang log untuk menjaga kestabilan numerik dan konsistensi pembobotan. Dengan rancangan tersebut, model Mix Distribusi diharapkan mampu meningkatkan kualitas klasifikasi biner dengan memanfaatkan kekuatan MNB dalam menangkap kata-kata indikator sentimen dan kekuatan GNB dalam menangkap representasi pola laten yang lebih stabil.

Setelah prediksi akhir diperoleh melalui penggabungan model, penelitian ini memastikan bahwa proses pelatihan tidak bias akibat ketidakseimbangan jumlah data positif dan negatif. Oleh karena itu, tahap berikutnya menerapkan penyeimbangan kelas pada data latih menggunakan RandomOverSampler sebelum evaluasi dilakukan. Secara formal, kombinasi campuran didefinisikan sebagai:

$$\log P_{mix}(y = c|x) = \alpha \log P_{MNB}(y = c|x) + (1 - \alpha) \log P_{GNB}(y = c|x),\ 0 \leq \alpha \leq 1 \quad (3.10)$$

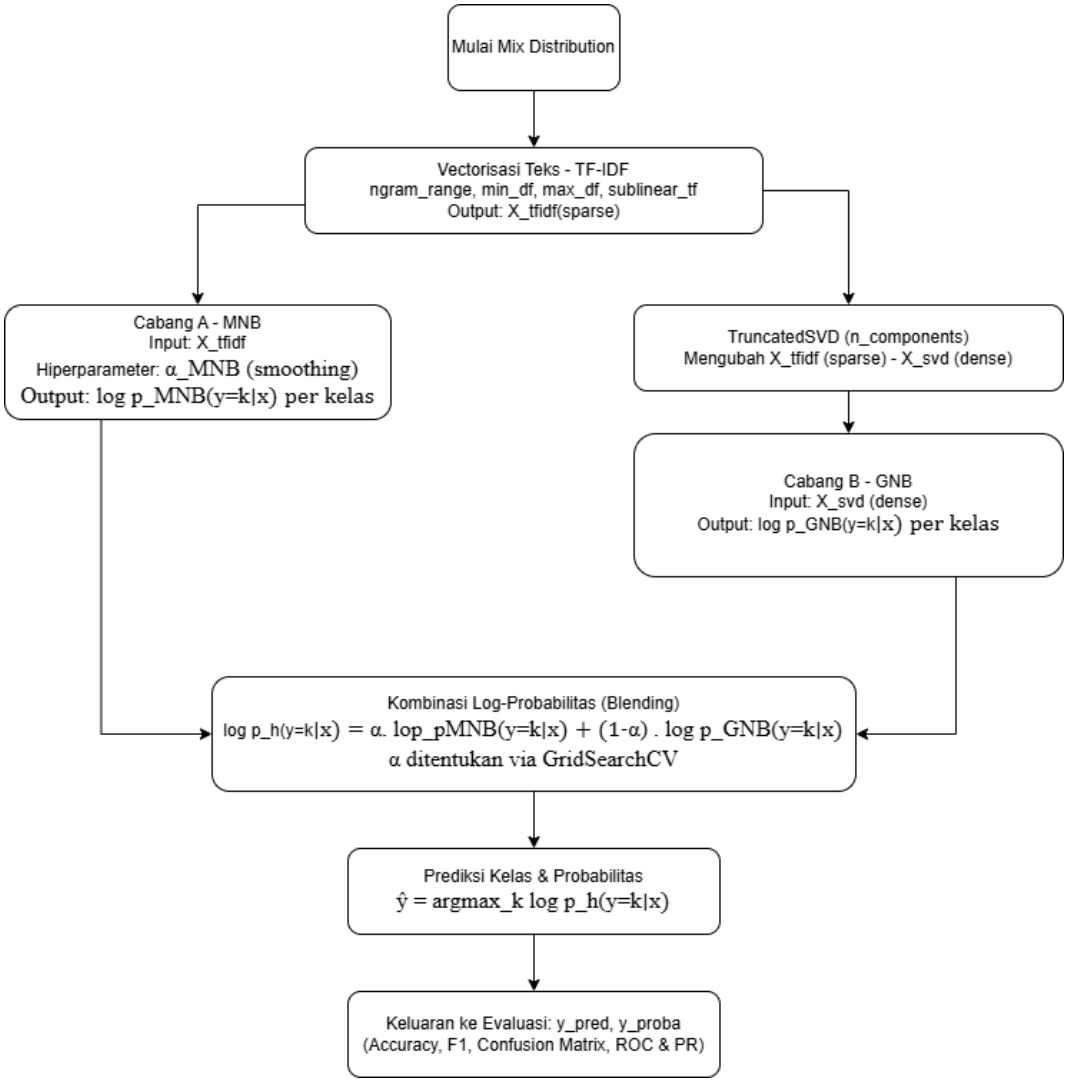
Keterangan:

P_{mix} : probabilitas gabungan model

P_{MNB} : probabilitas dari MultinoialNB

P_{GNB} : probabilitas dari GaussianNB

Pada penelitian ini, Mix Distribusi Naive Bayes dibangun dengan menggabungkan probabilitas posterior yang dihasilkan oleh Multinomial Naive Bayes dan Gaussian Naive Bayes. Multinomial Naive Bayes memproses bobot fitur hasil TF-IDF, sedangkan Gaussian Naive Bayes memproses fitur kontinu hasil reduksi dimensi TruncatedSVD. Penggabungan dilakukan pada ruang log menggunakan bobot α, sehingga probabilitas akhir diperoleh dari kombinasi kontribusi kedua model. Setelah log-probability digabungkan, nilainya dikembalikan ke bentuk probabilitas melalui fungsi eksponensial dan dinormalisasi agar total probabilitas sama dengan satu. Kelas dengan probabilitas akhir terbesar dipilih sebagai hasil klasifikasi akhir.



Gambar 3. 3 Blok Diagram Mix Distribusi

Gambar 3.3 menggambarkan arsitektur *mix distribution* yang digunakan pada penelitian ini untuk melakukan klasifikasi sentimen biner (positif dan negatif). Proses dimulai dari tahap vektorisasi teks menggunakan *TF–IDF*, yang menghasilkan representasi fitur X_{tfidf} dalam bentuk matriks berdimensi tinggi dan bersifat *sparse*. Representasi ini digunakan langsung oleh Cabang A (*Multinomial Naïve Bayes*/*MNB*) karena model tersebut sesuai untuk fitur berbasis kata yang

menggambarkan kecenderungan kemunculan term pada masing-masing kelas sentimen. Pada cabang ini, model *MNB* mempelajari parameter likelihood serta prior kelas dari data latih, lalu menghasilkan skor log-probabilitas untuk tiap kelas (positif/negatif) pada setiap komentar masukan.

Secara paralel, keluaran *TF–IDF* yang sama juga diproses oleh *Truncated SVD* untuk melakukan reduksi dimensi sehingga diperoleh representasi fitur baru X_{svd} yang lebih padat (dense) dan bersifat kontinu. Representasi ini kemudian menjadi masukan bagi Cabang B (*Gaussian Naïve Bayes*/*GNB*). *GNB* dipilih karena bekerja efektif pada fitur kontinu hasil reduksi dimensi, sehingga model mampu menangkap pola laten yang lebih global dari teks dan mengurangi sensitivitas terhadap variasi pemilihan kata. Dari cabang ini, *GNB* juga menghasilkan skor log-probabilitas untuk dua kelas yang sama.

Tahap inti pada arsitektur *Mix Distribusi* ditunjukkan pada blok kombinasi log-probabilitas (blending). Pada tahap ini, skor dari *MNB* dan *GNB* digabungkan menggunakan parameter bobot α untuk mengatur kontribusi relatif masing-masing model. Mekanisme penggabungan pada ruang log dipilih untuk menjaga stabilitas numerik dan memastikan skor dapat digabungkan secara konsisten. Nilai α diperlakukan sebagai hiperparameter dan ditentukan secara objektif melalui proses validasi (misalnya *GridSearchCV*) agar kombinasi yang digunakan benar-benar menghasilkan performa terbaik pada data penelitian, bukan berdasarkan preferensi subjektif.

Setelah skor gabungan diperoleh untuk kelas positif dan negatif, sistem menentukan prediksi akhir dengan memilih kelas yang memiliki skor gabungan tertinggi. Output prediksi kemudian digunakan pada tahap evaluasi melalui perhitungan metrik kinerja seperti *Accuracy*, *F1-score*, *Confusion Matrix*, serta kurva evaluasi seperti *ROC* dan *Precision–Recall*. Dengan rancangan alur pada Gambar 3.3, penelitian ini berupaya memanfaatkan kekuatan *MNB* dalam menangkap sinyal kata-kata spesifik sekaligus memanfaatkan keunggulan *GNB* pada fitur hasil *SVD* yang lebih ringkas, sehingga prediksi sentimen diharapkan menjadi lebih robust dan stabil pada variasi bahasa di data komentar.

Untuk menghindari bias evaluasi, penyeimbangan kelas (RandomOverSampler) diterapkan hanya pada data latih, sedangkan pembentukan *TF–IDF* dan *SVD* dilakukan dengan prinsip *fit* pada data latih dan *transform* pada data uji.

3.10 GridSearch CV

Setelah data dibagi menjadi data latih dan data uji, proses pemodelan dilakukan menggunakan sebuah pipeline yang berisi model *Mix Distribution Naive Bayes*. Model ini terdiri dari beberapa komponen utama, yaitu *TF-IDF* untuk mengubah teks menjadi data numerik, *Multinomial Naive Bayes* (MNB) untuk menangkap pola frekuensi kata, serta *Gaussian Naive Bayes* (GNB) yang menggunakan hasil reduksi dimensi dari *TruncatedSVD* agar lebih sesuai dengan data berbentuk kontinu.

Untuk mendapatkan performa terbaik, dilakukan proses penalaan hiperparameter menggunakan *GridSearchCV*. Metode ini bekerja dengan cara mencoba berbagai kombinasi parameter secara sistematis, kemudian memilih kombinasi terbaik berdasarkan hasil evaluasi. Proses ini menggunakan *stratified k-fold cross-validation*, sehingga pembagian data pada setiap lipatan tetap menjaga proporsi kelas. Selain itu, pada setiap lipatan, proses pembentukan fitur (TF-IDF dan SVD) hanya dilakukan pada data latih, sehingga tidak terjadi kebocoran data (data leakage).

Parameter yang diuji dalam *GridSearchCV* meliputi:

- α (alpha) untuk mengatur bobot gabungan antara MNB dan GNB
- alpha_mnb sebagai parameter smoothing pada MNB
- n_components untuk menentukan jumlah dimensi hasil reduksi SVD
- Parameter TF-IDF seperti min_df, max_df, max_features, ngram_range, dan sublinear_tf

Seluruh parameter tersebut dituliskan menggunakan format *pipeline* (misalnya clf__alpha, clf__n_components, dan sebagainya), sehingga *GridSearchCV* dapat mengatur setiap bagian model dengan tepat.

Dalam penelitian ini, proses evaluasi menggunakan *F1-score* (macro) karena metrik ini lebih sesuai untuk data yang tidak seimbang. Nilai *cross-validation* (CV) yang digunakan adalah 5, dan proses komputasi dijalankan secara paralel untuk mempercepat waktu pencarian.

Ketika *GridSearchCV* dijalankan, sistem akan mencoba seluruh kombinasi parameter yang tersedia (misalnya ribuan kombinasi), kemudian menghitung rata-rata skor pada setiap kombinasi. Kombinasi dengan nilai terbaik akan dipilih sebagai best_params_, dan model akan dilatih ulang menggunakan seluruh data latih untuk menghasilkan best_estimator_.

Hasil pengujian menunjukkan bahwa kombinasi parameter terbaik umumnya berada pada nilai α sekitar 0.6, alpha_mnb sekitar 0.5, serta jumlah komponen SVD sebesar 200, dengan pengaturan *TF-IDF* yang cukup selektif. Kombinasi ini mampu menghasilkan keseimbangan yang baik antara informasi kata dan efisiensi model.

Model terbaik yang dihasilkan dari *GridSearchCV* kemudian digunakan untuk melakukan prediksi pada data uji dan dievaluasi menggunakan metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.

Dengan penggunaan *GridSearchCV*, proses pemilihan parameter menjadi lebih sistematis, objektif, dan dapat dipertanggungjawabkan, sehingga hasil model yang diperoleh lebih optimal dan sesuai dengan tujuan penelitian.

3.11 RandomOverSampler

Pada penelitian klasifikasi sentimen biner, salah satu tantangan yang sering muncul adalah ketidakseimbangan jumlah data antara kelas positif dan kelas negatif. Ketidakseimbangan ini dapat terjadi karena kecenderungan pengguna lebih banyak menuliskan komentar dengan polaritas tertentu, atau karena proses pengambilan data yang tidak sepenuhnya merata. Apabila kondisi ini dibiarkan,

model klasifikasi beresiko mengalami bias terhadap kelas mayoritas, yakni cenderung memprediksi kelas yang jumlahnya lebih banyak. Dampaknya, nilai akurasi bisa terlihat tinggi, tetapi kemampuan model mengenali kelas minoritas menjadi rendah, sehingga model tidak representatif untuk digunakan sebagai alat analisis sentimen yang objektif.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik *oversampling* menggunakan *RandomOverSampler*. *RandomOverSampler* bekerja dengan cara menggandakan sampel secara acak dari kelas minoritas hingga jumlah sampel pada kedua kelas menjadi seimbang. Dengan demikian, proses pelatihan model memperoleh paparan yang lebih setara terhadap pola-pola pada kelas positif dan negatif. Implementasi *oversampling* dipilih karena sederhana, efektif dalam banyak kasus klasifikasi biner, serta tidak mengubah informasi asli dari kelas mayoritas melainkan hanya menambah representasi kelas minoritas melalui duplikasi sampel.

Secara operasional, jumlah sampel setiap kelas setelah *oversampling* disamakan dengan jumlah sampel pada kelas yang paling besar (kelas mayoritas). Target jumlah sampel tersebut dapat dinyatakan sebagai:

$$N_{baru} = \max\left(N_{positif}, N_{negatif}\right) \tag{3.11}$$

Keterangan:

N_{baru} **:** jumlah sampel telah disamakan

$N_{positif}, N_{negatif}$: jumlah data per kelas

Di mana N_{positif} dan N_{negatif} adalah jumlah sampel awal pada masing-masing kelas, sedangkan N_{baru} adalah jumlah sampel target per kelas setelah dilakukan penyeimbangan. Setelah *oversampling* diterapkan, dataset latih menjadi lebih seimbang sehingga model tidak "terdorong" untuk selalu memilih kelas dominan.

Ketentuan penting dalam penerapan *RandomOverSampler* adalah bahwa *oversampling* dilakukan hanya pada data latih (training set), bukan pada data uji (test set). Hal ini bertujuan untuk menjaga objektivitas evaluasi. Jika *oversampling* dilakukan pada data uji, maka akan terjadi kebocoran informasi dan hasil evaluasi menjadi bias karena data uji tidak lagi merepresentasikan distribusi asli. Dengan demikian, penelitian ini menerapkan *oversampling* setelah proses pemisahan data dan hanya pada bagian data yang digunakan untuk pelatihan model.

Sebagai catatan, penggunaan *oversampling* dapat meningkatkan kemampuan model dalam mengenali kelas minoritas, namun juga berpotensi meningkatkan risiko *overfitting* karena adanya duplikasi data. Oleh sebab itu, efektivitas *RandomOverSampler* tetap dievaluasi secara empiris melalui perbandingan metrik performa, terutama metrik yang sensitif terhadap ketidakseimbangan kelas seperti *F1-score*, *precision*, dan *recall*.

Evaluasi performa dilakukan untuk mengukur sejauh mana model *Mix Distribution* mampu mengklasifikasikan komentar ke dalam dua kelas sentimen (positif dan negatif) secara konsisten dan dapat diandalkan. Pada penelitian ini, evaluasi tidak hanya berfokus pada satu metrik, karena pada klasifikasi biner terutama ketika data berpotensi tidak seimbang akurasi saja dapat memberikan

gambaran yang menyesatkan. Oleh karena itu, penelitian menggunakan beberapa metrik yang saling melengkapi, yaitu *Confusion Matrix*, *Accuracy*, *Precision*, *Recall*, dan *F1-score*, serta kurva evaluasi seperti *ROC* dan *Precision–Recall* (PR) untuk memberikan interpretasi yang lebih komprehensif.

3.12Confusion Matrix

Confusion matrix digunakan untuk merangkum hasil prediksi model terhadap label aktual. Dalam klasifikasi biner, *confusion matrix* terdiri dari empat komponen utama:

- True Positive (TP): jumlah komentar yang sebenarnya positif dan diprediksi positif.
- True Negative (TN): jumlah komentar yang sebenarnya negatif dan diprediksi negatif.
- False Positive (FP): jumlah komentar yang sebenarnya negatif tetapi diprediksi positif.
- False Negative (FN): jumlah komentar yang sebenarnya positif tetapi diprediksi negatif.

Keempat komponen ini menjadi dasar untuk menghitung metrik evaluasi berikutnya. Selain itu, *confusion matrix* juga membantu peneliti memahami pola kesalahan model, misalnya apakah model lebih sering “menganggap negatif sebagai positif” atau sebaliknya.

3.13 Accuracy

Accuracy menggambarkan proporsi prediksi yang benar dibandingkan seluruh data uji. Metrik ini memberikan ringkasan umum kinerja model, tetapi pada kondisi ketidakseimbangan kelas, *accuracy* dapat tetap tinggi meskipun model lemah pada kelas minoritas. Oleh karena itu, *accuracy* digunakan sebagai indikator awal dan harus dibaca bersama metrik lain. Dengan menggunakan persamaan 3.12.

$$\frac{TP+TN}{TP+FN+FP+TN} \tag{3.12}$$

3.14 Precision

Precision menunjukkan tingkat ketepatan prediksi pada kelas positif, yaitu seberapa banyak prediksi positif model yang benar-benar positif dengan persamaan 3.13. Metrik ini penting ketika konsekuensi *false positive* perlu diminimalkan, misalnya ketika komentar negatif yang diprediksi positif dapat mengaburkan interpretasi evaluasi sentimen.

$$\frac{TP}{TP+FP} \tag{3.13}$$

3.15 F1-Score

F1-score merupakan rata-rata harmonik dari *precision* dan *recall*, sehingga memberikan keseimbangan antara kedua metrik tersebut. Dalam klasifikasi biner, *F1-score* sering dijadikan metrik utama ketika data tidak seimbang karena mempertimbangkan kesalahan FP dan FN secara simultan menggunakan persamaan 3.14.

$$\frac{TP}{TP+FN} \tag{3.14}$$

3.16 Recall

Recall mengukur kemampuan model menangkap seluruh komentar yang benar-benar positif, yaitu seberapa banyak kasus positif yang berhasil terdeteksi dengan menggunakan persamaan 3.15. Metrik ini penting ketika konsekuensi *false negative* lebih kritis, misalnya jika komentar positif yang tidak terdeteksi menyebabkan analisis sentimen menjadi bias negatif.

$$2\frac{Precision\ x\ recall}{Precision\ x\ recall} \tag{3.15}$$

BAB IV

SKENARIO UJI COBA DAN PEMBAHASAN

4.1 Hasil Uji Coba

Proses pelatihan model dilakukan setelah data melalui tahap *pre-processing* dan representasi fitur menggunakan *TF-IDF Vectorizer*. Pada tahap ini, model *Mix Distribusi* dilatih menggunakan data latih sebanyak 1547 sampel, yang diperoleh dari proses *stratified split* dengan proporsi 80:20. Pembagian *stratified* memastikan bahwa distribusi kedua kelas (Positif dan Negatif) tetap seimbang antara data latih maupun data uji. Pada data latih, distribusi kelas adalah 902 komentar Negatif dan 645 komentar Positif, sehingga pelatihan dilakukan pada komposisi data yang konsisten dengan distribusi kelas keseluruhan.

Pelatihan model dilakukan menggunakan *GridSearchCV*, yang berfungsi untuk mencari kombinasi representative terbaik secara sistematis melalui proses *k-fold cross-validation* sebanyak 5 lipatan. Seluruh komponen *TF-IDF*, *TruncatedSVD*, dan *Mix Distribusi* dibungkus dalam sebuah *pipeline* sehingga proses pelatihan berlangsung konsisten, tidak terjadi kebocoran data, dan setiap tahap transformasi hanya dipelajari dari data pada lipatan pelatihan.

Selama proses *GridSearchCV*, sistem menelusuri ratusan hingga ribuan kombinasi representative dan menghitung skor validasi silang menggunakan metrik *F1-macro*, yang dipilih agar performa kedua kelas diperhatikan secara seimbang, terutama karena data memiliki sedikit ketidakseimbangan label. Setelah melakukan *GridSearch*, kita mendapat konfigurasi yang menunjukkan kombinasi yang seimbang antara representasi frekuensi kata (TF-IDF untuk MNB) dan representasi

represen laten (SVD untuk GNB). Nilai $\alpha \approx 0.6$ menunjukkan bahwa kontribusi MNB sedikit lebih dominan dibandingkan GNB, namun tetap mempertahankan informasi fitur dense yang ditangkap oleh GNB.

Setelah proses pencarian representative selesai, *GridSearchCV* menghasilkan best_estimator_ yang kemudian digunakan sebagai model final untuk proses evaluasi pada data uji. Model ini merupakan hasil pelatihan penuh menggunakan seluruh data latih (1547 sampel) dengan konfigurasi terbaik dari *GridSearch*.

Secara keseluruhan, tahap pelatihan ini memastikan bahwa model yang digunakan pada proses prediksi adalah model yang telah dioptimalkan berdasarkan kombinasi representative terbaik serta diuji dan divalidasi secara ketat melalui *cross-validation*. Hal ini memberikan keyakinan bahwa model memiliki kinerja yang stabil dan mampu melakukan generalisasi dengan baik terhadap data baru. Optimisasi model: dilakukan dengan *GridSearchCV* untuk mencari kombinasi parameter terbaik seperti alpha, max_df, dan ngram_range.

Seluruh proses dalam penelitian ini mulai dari pemuatan data, *pre-processing* teks, ekstraksi fitur *TF-IDF*, pelatihan model *Mix Distribusi Naive Bayes*, hingga evaluasi—dijalankan menggunakan Jupyter Notebook. Pemilihan Jupyter Notebook didasarkan pada kemampuannya mengeksekusi kode secara bertahap, menampilkan output secara langsung, serta memudahkan dokumentasi dan pengujian model dalam satu lingkungan kerja.

Pembagian data ini dilakukan menggunakan beberapa represen, seperti 80:20, 70:30, dan 90:10, untuk melihat stabilitas performa model pada berbagai

tingkat ketersediaan data pelatihan. Melalui pembagian data yang sistematis dan terkontrol, penelitian ini memastikan bahwa model yang dievaluasi merupakan model yang benar-benar teruji dan memiliki kemampuan generalisasi yang baik. Pembagian data (train–test split) dilakukan untuk memisahkan data yang digunakan untuk melatih model dan data yang digunakan untuk menguji performanya. Pemisahan ini penting agar model dapat diuji pada data yang tidak pernah dilihat sebelumnya sehingga menghindari *overfitting* dan memastikan kemampuan generalisasi. Pembagian dilakukan secara *stratified* agar proporsi kelas tetap seimbang pada data latih maupun data uji, sehingga evaluasi model menjadi lebih objektif dan tidak bias. Pada penelitian ini kita menjalankan 3 skenario pengujian dan untuk rincian pembagian datanya dapat dilihat pada tabel 4.1 dibawah ini.

Tabel 4. 1 Pembagian Data Training dan Data Testing

Pembagian	Train	Test
80:20	1547	387
70:30	1354	580
90:10	1741	193

Tahap *pre-processing* dilakukan untuk membersihkan dan menormalkan data teks agar siap digunakan dalam proses pemodelan. *Pre-processing* ini sangat penting karena data komentar YouTube umumnya bersifat tidak terstruktur, mengandung noise, dan memiliki variasi penulisan yang tinggi. Proses *pre-processing* diawali dengan normalisasi teks, yaitu mengubah seluruh karakter menjadi huruf kecil (lowercase) untuk menghindari perbedaan makna akibat variasi kapitalisasi. Selanjutnya dilakukan pembersihan teks dengan menghapus URL,

mention, hashtag, angka, serta tanda baca atau repres yang tidak memiliki kontribusi represen terhadap analisis sentimen.

Setelah pembersihan dasar, dilakukan penghapusan *stopword* menggunakan daftar *stopword* Bahasa Indonesia untuk menghilangkan kata-kata umum yang sering muncul tetapi tidak memiliki nilai diskriminatif terhadap sentimen, seperti kata hubung dan kata depan. Tahap berikutnya adalah *stemming* menggunakan Sastrawi, yang bertujuan mengembalikan kata ke bentuk dasarnya sehingga variasi kata berimbuhan dapat direduksi menjadi satu representasi yang sama. Proses ini membantu mengurangi dimensi fitur dan meningkatkan konsistensi representasi teks.

Hasil dari seluruh tahapan *pre-processing* disimpan dalam kolom teks bersih (misalnya *stemmed*), yang kemudian digunakan sebagai masukan pada tahap ekstraksi fitur menggunakan *TF-IDF*. Dengan menerapkan *pre-processing* secara sistematis, data yang digunakan dalam pelatihan model menjadi lebih bersih, representative, dan relevan, sehingga dapat meningkatkan kinerja dan stabilitas model klasifikasi sentimen yang dibangun. Untuk melihat contoh proses yg dilakukan pada tahap ini dapat dilihat pada tabel 4.2 berikut.

Tabel 4. 2 Contoh Proses Pre-Processing

Tahap	Teks
Teks Mentah	Program Prakerja ini BAGUS banget!!! 😍
Lowercase	program prakerja ini bagus banget!!!

Cleaning	program prakerja ini bagus banget
Stopword Removal	program prakerja bagus banget
Stemming	program prakerja bagus banget

Pada penelitian ini dilakukan pengujian terhadap tiga model klasifikasi, yaitu *Multinomial Naive Bayes* (MNB-only), *Gaussian Naive Bayes* (GNB-only), dan *Mix Distribution Naive Bayes*, dengan menggunakan tiga skenario pembagian data, yaitu 70:30, 80:20, dan 90:10 antara data latih dan data uji.

Evaluasi kinerja model dilakukan menggunakan confusion matrix yang terdiri dari *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), dan *True Positive* (TP), serta metrik evaluasi berupa *accuracy*, *precision*, *recall*, dan *F1-score*.

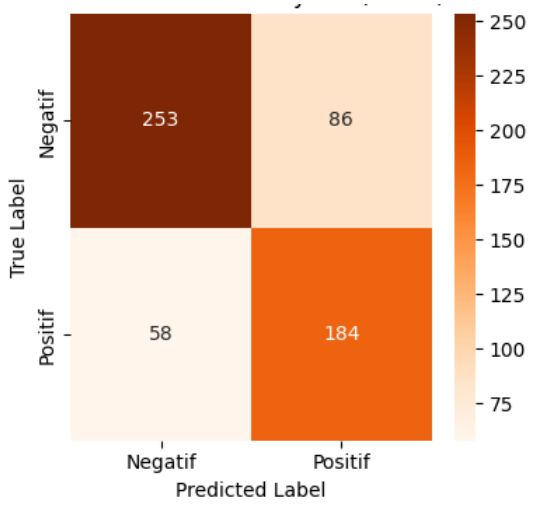
4.1.1 Skenario 1 (Data *Training* 70% dan Data *Testing* 30%)

Pada skenario pengujian pertama, dataset dibagi dengan rasio 70:30, sehingga sebanyak 581 data digunakan sebagai data uji. Hasil pengujian model *Mix Distribusi* pada skenario ini ditunjukkan melalui *confusion matrix* dan metrik evaluasi pada Gambar 4.1.

Berdasarkan *confusion matrix*, model berhasil mengklasifikasikan 253 data berlabel *True Negatif* (TN) dan 184 data berlabel *True Positif* (TP) secara benar. Namun demikian, terdapat 86 data Negatif yang salah diprediksi sebagai Positif (FP) dan 58 data Positif yang salah diprediksi sebagai Negatif (FN). Hasil ini

menunjukkan bahwa jumlah prediksi benar lebih dominan dibandingkan prediksi salah, sehingga model mampu mengenali pola sentimen dengan cukup baik.

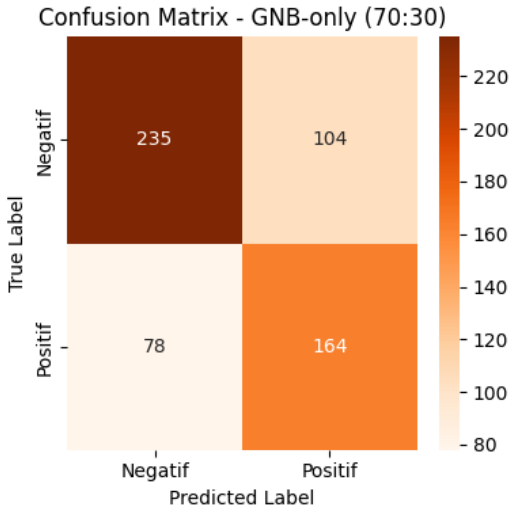
Secara kuantitatif, model mencapai akurasi sebesar 75,22%, yang menunjukkan bahwa sekitar tiga perempat data uji dapat diklasifikasikan dengan benar. Nilai *F1-score* mikro sebesar 0,752 dan *F1-score* makro sebesar 0,747 menandakan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.



Gambar 4. 1 Matrix Mixed Distribution Skenario 1

Berdasarkan *confusion matrix* pada gambar 4.2, model berhasil mengklasifikasikan 235 data berlabel *True Negatif* (TN) dan 162 data berlabel *True Positif* (TP) secara benar. Namun demikian, terdapat 104 data Negatif yang salah diprediksi sebagai Positif (FP) dan 80 data Positif yang salah diprediksi sebagai Negatif (FN). Hasil ini menunjukkan bahwa jumlah prediksi benar lebih dominan dibandingkan prediksi salah, sehingga model mampu mengenali pola sentimen dengan cukup baik.

Secara kuantitatif, model mencapai akurasi sebesar 68,33%, yang menunjukkan bahwa sekitar hampir tiga perempat data uji dapat diklasifikasikan dengan benar. Nilai *F1-score* mikro sebesar 0,683 dan *F1-score* makro sebesar 0,677 menandakan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.

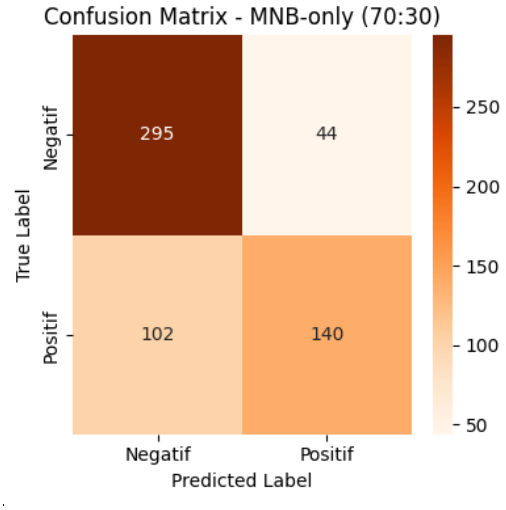


Gambar 4. 2 Matrix Gaussian Naïve Bayes Skenario 1

Berdasarkan *confusion matrix* pada gambar 4.3, model berhasil mengklasifikasikan 295 data berlabel *True Negatif* (TN) dan 140 data berlabel *True Positif* (TP) secara benar. Namun demikian, terdapat 44 data Negatif yang salah diprediksi sebagai Positif (FP) dan 102 data Positif yang salah diprediksi sebagai Negatif (FN). Hasil ini menunjukkan bahwa model mampu mengenali pola sentimen dengan cukup baik.

Secara kuantitatif, model mencapai akurasi sebesar 74,87%, yang menunjukkan bahwa sekitar hampir tiga perempat data uji dapat diklasifikasikan dengan benar.

Nilai *F1-score* mikro sebesar 0,748 dan *F1-score* makro sebesar 0,752 menandakan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.



Gambar 4. 3 Matrix Multinomial Naïve Bayes Skenario 1

Secara keseluruhan, hasil skenario pengujian pertama dengan pembagian data 70:30 menunjukkan bahwa model *Mix Distribution* memiliki performa yang cukup stabil, dengan *accuracy* tertinggi dan mampu melakukan klasifikasi sentimen dengan baik. Model *Mix Distribution* memberikan performa terbaik berdasarkan F1-score. MNB memiliki precision tinggi, namun recall rendah karena banyak data positif yang tidak terdeteksi (FN tinggi) seperti pada tabel 4.3.

Tabel 4. 3 Split Data 70:30

Model	Accuracy	Precision	Recall	F1-Score
GNB	68.33%	60.90%	66.94%	63.78%
MIX	75.22%	68.15%	76.03%	71.88%
MNB	74.87%	76.09%	57.85%	65.73%

Dengan jumlah data training yang relatif lebih sedikit dibandingkan skenario lainnya, model memiliki informasi yang lebih terbatas untuk mempelajari pola sentimen pada data. Akibatnya, performa model masih menunjukkan beberapa kesalahan klasifikasi, terutama pada kelas positif yang terlihat dari nilai False Negative yang relatif lebih tinggi pada beberapa model. Meskipun demikian, skenario ini memberikan evaluasi yang cukup ketat, karena jumlah data testing yang lebih besar dapat menunjukkan kemampuan generalisasi model terhadap data baru.

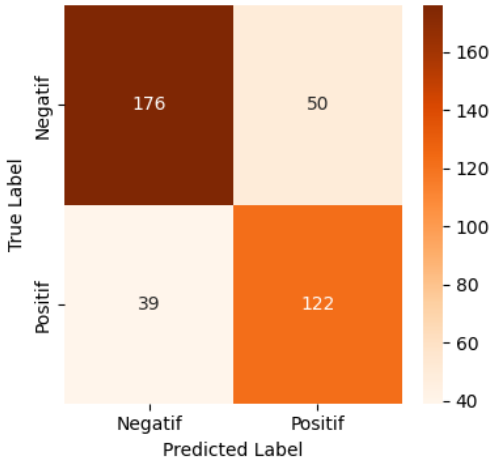
4.1.2 Skenario 2 (Data *Training* 80% dan Data *Testing* 20%)

Pada skenario pengujian ini, data dibagi dengan rasio 80:20, sehingga sebanyak 387 data digunakan sebagai data uji. Kinerja model *Mix Distribusi* pada skenario ini ditunjukkan melalui *confusion matrix* dan metrik evaluasi sebagaimana ditampilkan pada Gambar 4.4.

Berdasarkan confusion matrix, model berhasil mengklasifikasikan 176 data berlabel *True Negatif* (TN) dan 122 data berlabel *True Positif* (TP). Di sisi lain, terdapat 50 data Negatif yang salah diprediksi sebagai Positif (FP) serta 39 data Positif yang salah diprediksi sebagai Negatif (FN). Dengan demikian, jumlah prediksi benar lebih besar dibandingkan prediksi salah, yang menunjukkan bahwa model mampu mengenali pola sentimen dengan baik pada skenario ini.

Secara kuantitatif, model mencapai akurasi sebesar 77,00%, yang menunjukkan peningkatan dibandingkan skenario pembagian data 70:30. *Nilai F1-*

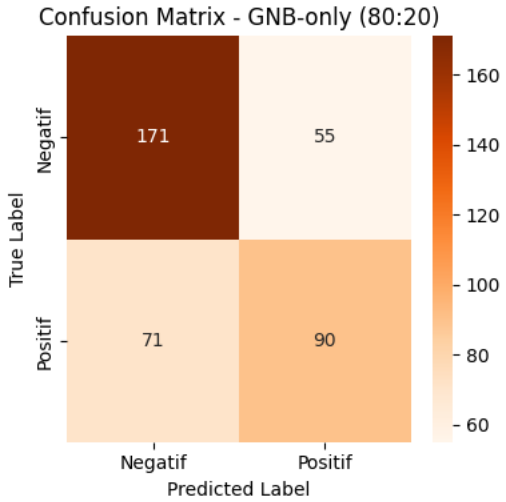
score mikro sebesar 0,770 dan *F1-score* makro sebesar 0,764 mengindikasikan bahwa performa model relatif seimbang antara kedua kelas sentimen.



Gambar 4. 4 Matrix Mix Distribution Skenario 2

Berdasarkan confusion matrix pada gambar 4.5, model berhasil mengklasifikasikan 171 data berlabel *True Negatif* (TN) dan 90 data berlabel *True Positif* (TP). Di sisi lain, terdapat 55 data Negatif yang salah diprediksi sebagai Positif (FP) serta 71 data Positif yang salah diprediksi sebagai Negatif (FN). Dengan demikian, jumlah prediksi benar lebih besar dibandingkan prediksi salah, yang menunjukkan bahwa model mampu mengenali pola sentimen dengan baik pada skenario ini.

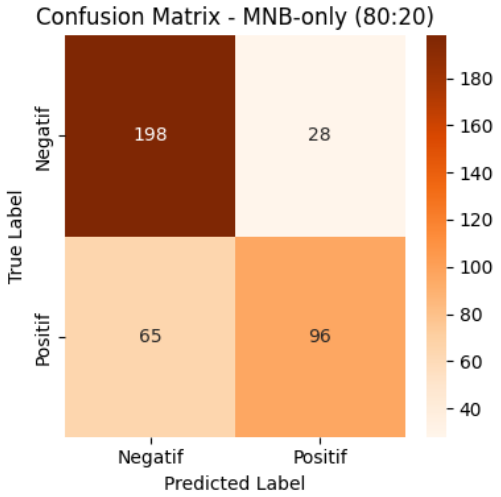
Secara kuantitatif, model mencapai akurasi sebesar 67,44%. *Nilai F1-score* mikro sebesar 0,674 dan *F1-score* makro sebesar 0,663 mengindikasikan bahwa performa model relatif seimbang antara kedua kelas sentimen.



Gambar 4. 5 Matrix Gaussian Naïve Bayes Skenario 2

Berdasarkan confusion matrix pada gambar 4.6, model berhasil mengklasifikasikan 198 data berlabel *True Negatif* (TN) dan 96 data berlabel *True Positif* (TP). Di sisi lain, terdapat 28 data Negatif yang salah diprediksi sebagai Positif (FP) serta 65 data Positif yang salah diprediksi sebagai Negatif (FN). Dengan demikian, jumlah prediksi benar lebih besar dibandingkan prediksi salah, yang menunjukkan bahwa model mampu mengenali pola sentimen dengan baik pada skenario ini.

Secara kuantitatif, model mencapai akurasi sebesar 75,97%. *Nilai F1-score* mikro sebesar 0,759 dan *F1-score* makro sebesar 0,763 mengindikasikan bahwa performa model relatif seimbang antara kedua kelas sentimen.



Gambar 4. 6 Matrix Multinomial Naïve Bayes Skenario 2

Secara keseluruhan, hasil pengujian dengan pembagian data 80:20 menunjukkan bahwa peningkatan proporsi data latih memberikan dampak positif terhadap kinerja model. Model Mix Distribution kembali menunjukkan performa terbaik dengan nilai F1-score tertinggi serta nilai FN yang lebih kecil, yang berarti lebih baik dalam menangkap sentimen positif.

Tabel 4. 4 Split Data 80:20

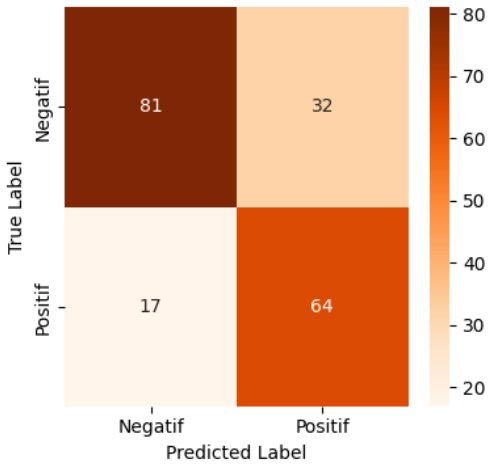
Model	Accuracy	Precision	Recall	F1-Score
GNB	67.44%	62.07%	55.90%	58.82%
MIX	77.00%	70.93%	75.78%	73.27%
MNB	75.97%	77.42%	59.63%	67.37%

Pada skenario 80:20, jumlah data training meningkat sehingga model memiliki lebih banyak data untuk mempelajari pola sentimen. Dengan bertambahnya data training, model menjadi lebih mampu mengenali hubungan antara kata-kata dalam teks dan label sentimen yang sesuai. Hal ini terlihat dari

peningkatan jumlah prediksi yang benar pada beberapa model, serta penurunan kesalahan klasifikasi dibandingkan dengan skenario sebelumnya. Skenario ini biasanya dianggap sebagai pembagian data yang cukup seimbang, karena masih menyediakan jumlah data testing yang cukup untuk melakukan evaluasi performa model.

4.1.3 Skenario 3 (Data *Training* 90% dan Data *Testing* 10%)

Pada skenario pengujian ini, dataset dibagi dengan rasio 90:10, sehingga sebanyak 194 data digunakan sebagai data uji. Skenario ini bertujuan untuk mengamati kinerja model *Mix Distribusi* ketika proporsi data latih diperbesar, sehingga model memiliki kesempatan yang lebih luas untuk mempelajari pola sentimen dari data.

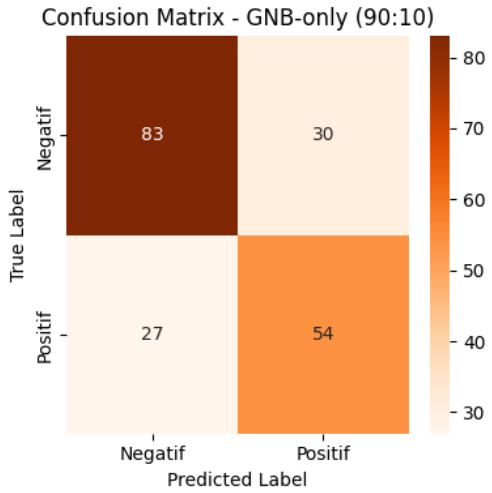


Gambar 4. 7 Matrix Mix Distribution Skenario 3

Berdasarkan *confusion matrix* pada Gambar 4.7, model berhasil mengklasifikasikan 81 data berlabel *True Negatif* (TN) dan 64 data berlabel *True Positif* (TP). Namun demikian, masih terdapat 32 data Negatif yang salah diprediksi

sebagai Positif (FP) serta 17 data Positif yang salah diprediksi sebagai Negatif (FN). Secara keseluruhan, jumlah prediksi benar (145 data) masih lebih dominan dibandingkan prediksi salah (49 data).

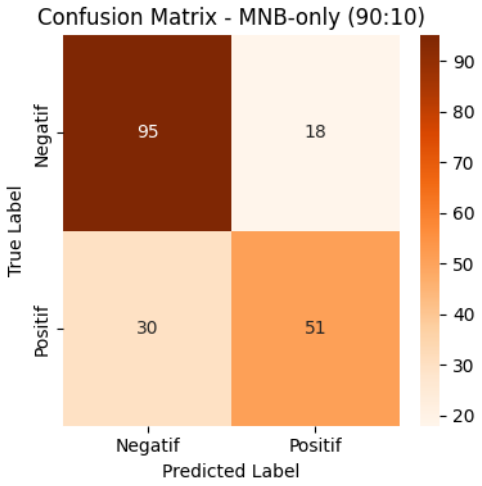
Secara kuantitatif, model mencapai akurasi sebesar 74,74%, yang menunjukkan bahwa sekitar tiga perempat data uji dapat diklasifikasikan dengan benar. Nilai *F1-score* mikro sebesar 0,747 dan *F1-score* makro sebesar 0,746 mengindikasikan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.



Gambar 4. 8 Matrix Gaussian Naïve Bayes Skenario 3

Berdasarkan *confusion matrix* pada Gambar 4.8, model berhasil mengklasifikasikan 80 data berlabel *True Negatif* (TN) dan 55 data berlabel *True Positif* (TP). Namun demikian, masih terdapat 33 data Negatif yang salah diprediksi sebagai Positif (FP) serta 26 data Positif yang salah diprediksi sebagai Negatif (FN). Secara keseluruhan, jumlah prediksi benar (135 data) masih lebih dominan dibandingkan prediksi salah (59 data).

Secara kuantitatif, model mencapai akurasi sebesar 69,59%, yang menunjukkan bahwa sekitar tiga perempat data uji dapat diklasifikasikan dengan benar. Nilai *F1-score* mikro sebesar 0,695 dan *F1-score* makro sebesar 0,689 mengindikasikan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.



Gambar 4. 9 Matrix Multiomial Naïve Bayes Skenario 3

Berdasarkan *confusion matrix* pada Gambar 4.9, model berhasil mengklasifikasikan 95 data berlabel *True Negatif* (TN) dan 51 data berlabel *True Positif* (TP). Namun demikian, masih terdapat 18 data Negatif yang salah diprediksi sebagai Positif (FP) serta 30 data Positif yang salah diprediksi sebagai Negatif (FN). Secara keseluruhan, jumlah prediksi benar (146 data) masih lebih dominan dibandingkan prediksi salah (48 data).

Secara kuantitatif, model mencapai akurasi sebesar 75,26%, yang menunjukkan bahwa sekitar tiga perempat data uji dapat diklasifikasikan dengan benar. Nilai *F1-score* mikro sebesar 0,752 dan *F1-score* makro sebesar 0,749 mengindikasikan bahwa kinerja model relatif seimbang pada kedua kelas sentimen.

Secara keseluruhan, hasil pengujian dengan pembagian data 90:10, MNB memiliki *accuracy* tertinggi, namun *Mix Distribution* memiliki *F1-score* lebih tinggi karena *recall* yang jauh lebih baik. Hal ini menunjukkan bahwa *Mix Distribution* lebih efektif dalam mengidentifikasi data positif.

Tabel 4. 5 Split Data 90:10

Model	Accuracy	Precision	Recall	F1-Score
GNB	69.59%	62.50%	67.90%	65.09%
MIX	74.74%	66.67%	79.01%	72.32%
MNB	75.26%	73.91%	62.96%	68.00%

Pada skenario 90:10, sebagian besar data digunakan sebagai data training, sehingga model memiliki lebih banyak contoh untuk mempelajari pola data. Dengan jumlah data training yang lebih besar, model dapat mempelajari karakteristik kata dan hubungan antar fitur dengan lebih baik, sehingga kemampuan prediksi model cenderung meningkat. Hal ini terlihat dari peningkatan nilai True Positive dan penurunan False Negative pada beberapa model, khususnya pada model Mix. Namun, jumlah data testing yang lebih sedikit membuat proses evaluasi menjadi lebih sensitif terhadap variasi data, sehingga hasil evaluasi mungkin terlihat lebih tinggi tetapi kurang merepresentasikan performa model pada data yang lebih beragam.

4.2 Pembahasan

Berdasarkan hasil pengujian pada tiga skenario pembagian data, model *Mix Distribution Naive Bayes* secara konsisten menunjukkan performa terbaik dalam

hal keseimbangan antara *precision* dan *recall*, yang dapat kita lihat dari nilai *F1-score* yang lebih tinggi dibandingkan model lainnya.

Model MNB-only cenderung memiliki *precision* tinggi, yang berarti prediksi positif yang dihasilkan cukup akurat. Namun, nilai *recall* yang rendah menunjukkan bahwa model ini kurang mampu menangkap seluruh data positif, sehingga menghasilkan nilai FN yang tinggi.

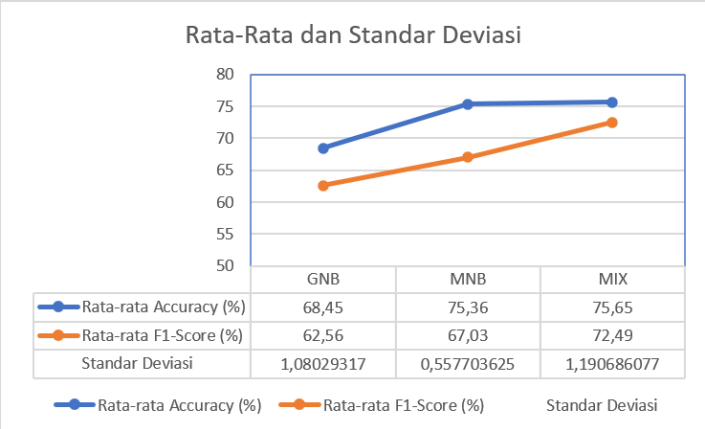
Sementara itu, model GNB-only memiliki performa yang relatif lebih rendah dibandingkan kedua model lainnya, terutama pada nilai *F1-score*, yang menunjukkan bahwa pendekatan Gaussian kurang optimal untuk data berbasis teks.

Penggunaan metode *Mix Distribution Naive Bayes* yang menggabungkan probabilitas dari *Multinomial Naive Bayes* dan *Gaussian Naive Bayes* terbukti mampu meningkatkan performa model secara keseluruhan. Hal ini disebabkan karena *Mix Distribution* mampu memanfaatkan keunggulan kedua metode, sehingga menghasilkan prediksi yang lebih seimbang.

Secara umum, semakin besar proporsi data training, maka model memiliki lebih banyak informasi untuk mempelajari pola pada data sehingga performa prediksi cenderung meningkat. Sebaliknya, semakin besar proporsi data testing, maka evaluasi model menjadi lebih ketat karena model diuji pada data yang lebih banyak. Perbedaan pembagian data mempengaruhi jumlah informasi yang dipelajari model. Semakin besar data training, model dapat mempelajari pola sentimen dengan lebih baik sehingga prediksi cenderung lebih akurat. Namun, jika

data testing terlalu sedikit, evaluasi model menjadi kurang representatif terhadap data baru.

Dalam penelitian ini, skenario 90:10 memberikan performa prediksi yang lebih baik, karena model memperoleh data training yang lebih banyak. Namun, skenario 80:20 dianggap sebagai pembagian yang paling seimbang, karena memberikan keseimbangan antara proses pelatihan model dan evaluasi performa pada data uji.



Gambar 4. 10 Rata-Rata dan Standar Deviasi

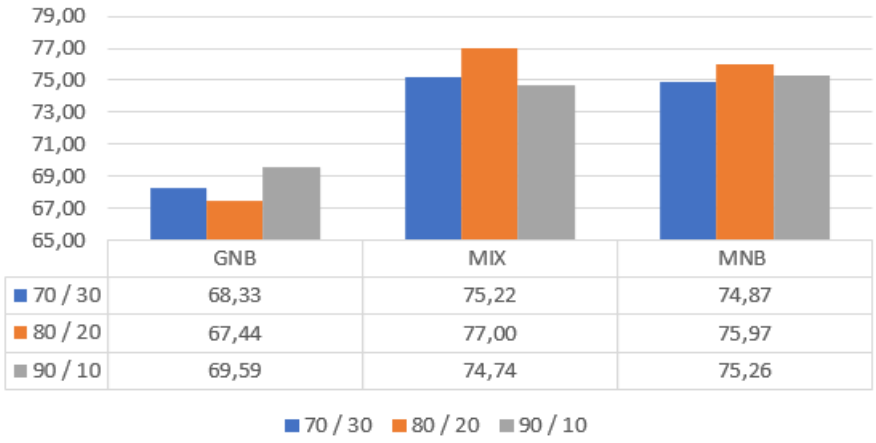
Berdasarkan hasil visualisasi pada Gambar (4.10), terlihat adanya tren peningkatan performa yang signifikan dari model Gaussian Naïve Bayes (GNB) menuju metode usulan Mix Distribution Naïve Bayes (MIX). Model GNB secara konsisten menempati posisi terendah dengan rata-rata akurasi sebesar 68,45% dan F1-Score sebesar 62,56%. Rendahnya performa ini mengindikasikan bahwa asumsi distribusi normal pada GNB kurang mampu merepresentasikan pola persebaran kata dalam data opini publik yang cenderung bersifat diskrit. Di sisi lain, model Multinomial Naïve Bayes (MNB) menunjukkan perbaikan performa dengan

akurasi yang meningkat tajam menjadi 75,36%, meskipun peningkatan ini tidak diikuti dengan kenaikan F1-Score yang setara, yang hanya mencapai 67,03%. Hal ini menunjukkan bahwa walaupun MNB mampu memprediksi jawaban benar secara keseluruhan dengan baik, model tersebut masih memiliki kendala dalam menjaga keseimbangan antara ketepatan (precision) dan jangkauan (recall) pada kelas-kelas opini yang diuji.

Metode usulan Mix Distribution Naïve Bayes (MIX) terbukti memberikan hasil terbaik dan paling stabil di antara ketiga model tersebut. Secara angka, model MIX mencapai puncak rata-rata akurasi sebesar 75,65% dan rata-rata F1-Score yang unggul jauh di angka 72,49%. Keunggulan yang mencolok pada metrik F1-Score ini menjadi poin krusial yang menunjukkan bahwa integrasi berbagai distribusi probabilitas dalam satu model mampu menangkap karakteristik fitur teks secara lebih fleksibel dan akurat. Selain itu, secara visual pada grafik, jarak (gap) antara garis akurasi dan F1-Score pada model MIX terlihat paling rapat dibandingkan model lainnya. Kerapatan ini membuktikan bahwa metode MIX berhasil meminimalisir kesalahan klasifikasi pada kelas tertentu dan memberikan hasil prediksi yang lebih berkualitas. Meskipun memiliki standar deviasi sebesar 1,190 yang sedikit lebih tinggi akibat lonjakan performa pada skenario tertentu, model MIX tetap menjadi model yang paling direkomendasikan karena konsistensi keunggulannya dalam menyeimbangkan aspek kuantitas dan kualitas klasifikasi opini publik.

4.2.1 Pengaruh Split Data

Pengaturan rasio split data memberikan pengaruh yang terukur terhadap performa model klasifikasi opini publik ini. Berdasarkan hasil pengujian, skenario 80:20 memberikan hasil paling optimal dengan rata-rata akurasi mencapai 77,00% pada model Mix Distribution. Hal ini menunjukkan bahwa komposisi data latih sebesar 80% mampu memberikan asupan informasi yang memadai bagi model untuk memahami pola bahasa dalam dataset tanpa terjebak pada kondisi overfitting. Sebaliknya, pada skenario 90:10, terlihat adanya sedikit penurunan performa yang mengindikasikan bahwa jumlah data uji yang terlalu minim (10%) meningkatkan sensitivitas metrik evaluasi terhadap variasi kecil dalam data, sehingga hasil yang diperoleh menjadi kurang representatif. Akan tetapi, model Mix Distribution menunjukkan stabilitas yang superior pada metrik F1-Score dengan standar deviasi terkecil dibandingkan model lainnya, yang membuktikan bahwa metode Mix Distribution tetap tangguh dalam menjaga keseimbangan antara precision dan recall di berbagai rasio pembagian data yang dapat dilihat pada gambar 4.11.

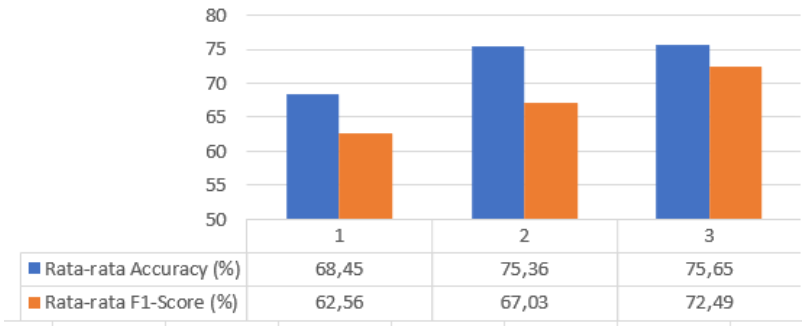


	GNB	MIX	MNB
70 / 30	68,33	75,22	74,87
80 / 20	67,44	77,00	75,97
90 / 10	69,59	74,74	75,26

Gambar 4. 11 Pengaruh Split Data

4.2.2 Pengaruh Model Distribusi

Pemilihan model klasifikasi memberikan dampak yang sangat nyata terhadap kualitas prediksi opini publik pada penelitian ini. Model Gaussian Naïve Bayes (GNB) secara konsisten memberikan performa terendah di seluruh skenario pengujian dengan rata-rata akurasi sebesar 68,45%. Rendahnya performa ini disebabkan oleh karakteristik dasar GNB yang mengasumsikan data terdistribusi secara normal (Gaussian), padahal fitur teks pada data opini Twitter cenderung bersifat diskrit dan tidak mengikuti kurva normal. Hal ini mengakibatkan GNB sering kali gagal menangkap pola probabilitas kata dengan akurat. Sebaliknya, model Multinomial Naïve Bayes (MNB) menunjukkan peningkatan performa yang cukup signifikan sebanyak 6,91% karena distribusinya memang dirancang khusus untuk menghitung frekuensi kemunculan kata dalam dokumen teks.



Gambar 4. 12 Pengaruh Model Distribusi

Puncak performa klasifikasi dicapai oleh metode usulan Mix Distribution Naïve Bayes (MIX) yang mampu menghasilkan rata-rata F1-Score tertinggi sebesar 72,49%. Keunggulan model MIX terletak pada fleksibilitasnya dalam menggabungkan kelebihan berbagai distribusi probabilitas, sehingga model dapat menyesuaikan diri dengan lebih baik terhadap variasi dan kompleksitas kata dalam

dataset opini. Jika dibandingkan dengan MNB, model MIX tidak hanya unggul dalam angka akurasi rata-rata, tetapi juga jauh lebih stabil dalam menjaga keseimbangan antara ketepatan prediksi (precision) dan jangkauan temuan data (recall). Dengan demikian, dapat disimpulkan bahwa penggunaan pendekatan distribusi campuran (mix) terbukti lebih efektif dalam menangani data opini publik dibandingkan dengan penggunaan distribusi tunggal pada model Naïve Bayes konvensional.

Penelitian analisis sentimen opini publik terhadap Program Prakerja yang dilakukan dalam skripsi ini tidak hanya berlandaskan pada pendekatan ilmiah dan teknis, tetapi juga selaras dengan nilai-nilai Islam yang menekankan kejujuran, keadilan, kehati-hatian, dan kemaslahatan dalam pengelolaan informasi. Islam memandang informasi sebagai amanah yang harus diolah secara bertanggung jawab agar tidak menimbulkan kesalahan persepsi maupun dampak negatif bagi masyarakat. Oleh karena itu, penggunaan metode analisis sentimen berbasis machine learning dalam penelitian ini dapat dipahami sebagai upaya untuk menilai opini publik secara objektif dan sistematis, sehingga hasil yang diperoleh lebih dapat dipertanggungjawabkan.

Prinsip kehati-hatian dalam menerima dan mengolah informasi ditegaskan dalam QS. Al-Hujurat ayat 6, yang berbunyi:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِنْ جَاۤءَكُمْ فَاسِقٌۢ بِنَبَاٍ فَتَبَيَّنُوْٓا اَنْ تُصِيْبُوْا قَوْمًاۢ بِجَهَالَةٍ فَتُصْبِحُوْا عَلٰى مَا فَعَلْتُمْ نٰدِمِيْنَ ۝٦

"Wahai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka telitilah kebenarannya, agar kamu tidak menimpakan musibah

kepada suatu kaum karena ketidaktahuan, yang akhirnya kamu menyesali perbuatan itu." (QS. Al-Hujurat: 6).

Ayat ini menekankan konsep tabayyun, yaitu kewajiban untuk memverifikasi informasi sebelum menarik kesimpulan. Dalam konteks penelitian ini, prinsip tabayyun diwujudkan melalui pengolahan data opini publik menggunakan tahapan pra-pemrosesan, ekstraksi fitur, dan klasifikasi berbasis algoritma yang terukur. Dengan demikian, penilaian sentimen tidak dilakukan secara subjektif, melainkan melalui analisis data yang sistematis dan berbasis bukti.

Selain kehati-hatian, Islam juga menekankan pentingnya keadilan dan objektivitas dalam menilai suatu permasalahan. Hal ini tercermin dalam QS. Al-Ma'idah ayat 8, yang menyatakan:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا كُوْنُوْا قَوّٰمِيْنَ لِلّٰهِ شُهَدَاۤءَ بِالْقِسْطِۖ وَلَا يَجْرِمَنَّكُمْ شَنَاٰنُ قَوْمٍ عَلٰٓى اَلَّا تَعْدِلُوْاۗ اِعْدِلُوْاۗ هُوَ

اَقْرَبُ لِلتَّقْوٰىۖ وَاتَّقُوا اللّٰهَۗ اِنَّ اللّٰهَ خَبِيْرٌۢ بِمَا تَعْمَلُوْنَ ۝٨

"Wahai orang-orang yang beriman, jadilah kamu penegak (kebenaran) karena Allah (dan) saksi-saksi (yang bertindak) dengan adil. Janganlah kebencianmu terhadap suatu kaum mendorong kamu untuk berlaku tidak adil. Berlakulah adil karena (adil) itu lebih dekat pada takwa. Bertakwalah kepada Allah. Sesungguhnya Allah Maha teliti terhadap apa yang kamu kerjakan." (QS. Al-Ma'idah: 8).

Ayat ini mengajarkan bahwa penilaian terhadap suatu kelompok atau fenomena harus dilakukan secara adil tanpa bias. Keterkaitan ayat ini dengan penelitian terletak pada proses klasifikasi sentimen yang dilakukan secara netral terhadap seluruh opini publik, baik yang bernada positif maupun negatif. Model

yang digunakan tidak diarahkan untuk membenarkan atau menyalahkan pihak tertentu, melainkan untuk memetakan kecenderungan opini masyarakat secara objektif berdasarkan data.

Lebih lanjut, pemanfaatan ilmu pengetahuan dan teknologi dalam Islam diarahkan untuk memberikan manfaat seluas-luasnya bagi umat manusia, sebagaimana ditegaskan dalam QS. Al-Anbiya ayat 107,

وَمَآ اَرْسَلْنٰكَ اِلَّا رَحْمَةً لِّلْعٰلَمِيْنَ ١٠٧

"Dan Kami tidak mengutus engkau (Muhammad) melainkan sebagai rahmat bagi seluruh alam." (QS. Al-Anbiya: 107).

Prinsip rahmatan lil 'alamin ini menjadi landasan bahwa ilmu dan teknologi, termasuk teknologi analisis data dan kecerdasan buatan, seharusnya digunakan untuk kemaslahatan. Dalam penelitian ini, analisis sentimen opini publik terhadap Program Prakerja diharapkan dapat menjadi masukan yang konstruktif bagi evaluasi kebijakan publik, sehingga teknologi tidak hanya berfungsi secara teknis, tetapi juga memberikan manfaat nyata bagi masyarakat.

Dengan penggunaan tiga skenario dalam penelitian ini sesuai dengan prinsip kesungguhan proses (Sa'a) untuk mencapai hasil yang objektif, di mana setiap perbedaan output pada tiap skenario merupakan representasi nyata dari variasi usaha dan parameter yang diuji.

وَاَنْ لَّيْسَ لِلْاِنْسَانِ اِلَّا مَا سَعٰىۙ ٣٩

"*bahwa manusia hanya memperoleh apa yang telah diusahakannya,*" *(QS. An-Najm: 39).*

Dalam pandangan Islam, tidak ada usaha yang sia-sia, hasil yang berbeda-beda tersebut justru menjadi nilai ilmiah yang valid karena menunjukkan bahwa penelitian ini telah melakukan eksplorasi mendalam untuk menemukan titik optimal (Ihsan) dalam metode Mix Distribution Naïve Bayes. Dengan demikian, ayat ini menegaskan bahwa keunggulan hasil pada skenario terbaik adalah buah manis dari ketelitian dalam membandingkan berbagai variabel.

Selain ayat Al-Qur'an, prinsip integrasi Islam dalam penelitian ini juga diperkuat oleh hadis Nabi Muhammad SAW. Hadis pertama berbunyi: *"Cukuplah seseorang dianggap berdusta apabila ia menceritakan setiap apa yang ia dengar"* (HR. Muslim). Hadis ini menekankan pentingnya memilah dan verifikasi informasi sebelum menyampaikan atau menyimpulkan sesuatu. Dalam penelitian ini, hadis tersebut relevan dengan tujuan analisis sentimen yang mengolah opini publik secara terstruktur dan terverifikasi, bukan berdasarkan asumsi atau opini subjektif peneliti.

Hadis kedua yang relevan adalah sabda Rasulullah SAW: "Sesungguhnya Allah mencintai seseorang yang apabila melakukan suatu pekerjaan, ia melakukannya dengan itqan (profesional dan sungguh-sungguh)" (HR. Thabrani). Hadis ini menegaskan pentingnya profesionalisme dan kesungguhan dalam setiap aktivitas, termasuk dalam kegiatan ilmiah. Penelitian ini menerapkan prinsip tersebut melalui penggunaan metodologi yang sistematis, pemilihan algoritma yang

sesuai, serta evaluasi model yang terukur dan dapat dipertanggungjawabkan secara ilmiah.

Dengan demikian, integrasi Islam dalam penelitian ini tidak sekadar bersifat normatif, tetapi menjadi kerangka nilai yang memperkuat tujuan penelitian. Nilai tabayyun, keadilan, kemaslahatan, kejujuran, dan profesionalisme menjadi landasan etis dalam penerapan teknologi analisis sentimen, sehingga hasil penelitian diharapkan tidak hanya valid secara ilmiah, tetapi juga bermanfaat dan bertanggung jawab sesuai dengan ajaran Islam.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, metode Mix Distribution Naïve Bayes (MIX) terbukti memiliki kinerja yang optimal dan stabil dalam mengklasifikasikan opini publik terhadap kebijakan Program Kartu Prakerja. Ditinjau dari nilai confusion matrix, integrasi model MIX dengan teknik RandomOverSampler (ROS) mampu menekan tingkat kesalahan prediksi (False Positive dan False Negative) secara signifikan, sehingga model dapat mengenali karakteristik data diskrit dan kontinu pada teks dengan tepat untuk menghasilkan klasifikasi sentimen positif maupun negatif yang proporsional (True Positive dan True Negative). Keunggulan ini divalidasi oleh metrik evaluasi pada skenario pembagian data terbaik (90:10), yang menghasilkan nilai rata-rata Akurasi sebesar 75,65% serta F1-Score sebesar 72,49%. Nilai F1-Score tersebut sekaligus membuktikan adanya keseimbangan yang sangat baik antara nilai Presisi (ketepatan prediksi) dan Recall (kemampuan menjaring informasi), sekaligus menegaskan bahwa penggabungan asumsi distribusi dalam metode MIX mampu mengatasi keterbatasan model tunggal seperti Gaussian Naïve Bayes (GNB) dan Multinomial Naïve Bayes (MNB) dalam klasifikasi teks media sosial.

Meskipun menunjukkan performa yang baik, model masih memiliki keterbatasan dalam mendeteksi sarkasme, ironi, dan komentar dengan sentimen campuran. Selain itu, model belum mempertimbangkan aspek temporal dan

intensitas emosi, sehingga hasil analisis belum sepenuhnya menggambarkan dinamika opini publik secara mendalam.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, beberapa saran yang dapat diberikan untuk pengembangan lebih lanjut adalah sebagai berikut:

1. Penelitian selanjutnya disarankan untuk mengeksplorasi model berbasis *deep learning*, seperti LSTM, CNN, atau transformer berbahasa Indonesia (misalnya IndoBERT), guna menangkap konteks semantik yang lebih kompleks dan meningkatkan kemampuan model dalam memahami kalimat panjang, ironi, maupun sarkasme.
2. Dataset yang digunakan dapat diperluas dengan menambahkan data dari berbagai platform media sosial lain, seperti Twitter, TikTok, atau forum daring, agar model memiliki kemampuan generalisasi yang lebih baik terhadap variasi gaya bahasa dan karakteristik teks.
3. Diperlukan pengembangan kamus slang otomatis menggunakan pendekatan *unsupervised* learning agar dapat menyesuaikan diri dengan dinamika bahasa gaul dan ekspresi baru di media sosial Indonesia yang terus berkembang.
4. Penelitian lanjutan dapat menggabungkan analisis sentimen dengan analisis emosi (emotion recognition) dan dimensi waktu (temporal analysis). Pendekatan ini dapat memberikan gambaran yang lebih mendalam tentang perubahan persepsi publik terhadap suatu topik dari waktu ke waktu.

5. Model *Mix Distribusi Naïve Bayes* yang telah dibangun dapat diimplementasikan dalam bentuk dashboard interaktif atau API analisis sentimen otomatis untuk keperluan instansi pemerintah, perusahaan, atau lembaga riset dalam melakukan pemantauan opini publik secara real-time.
6. Secara keseluruhan, penelitian ini berhasil menunjukkan bahwa model *Mix Distribusi Naive Bayes* merupakan pendekatan yang efektif, efisien, dan dapat diandalkan untuk analisis sentimen komentar publik berbahasa Indonesia. Dengan pengembangan lebih lanjut melalui integrasi teknologi NLP modern dan perluasan sumber data, sistem analisis sentimen seperti ini berpotensi menjadi alat strategis dalam memahami persepsi masyarakat digital di Indonesia.

DAFTAR PUSTAKA

Anggraini, W. P., & Utami, M. S. (2021). KLASIFIKASI SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN KARTU PEKERJA DI INDONESIA. *Faktor Exacta*, *13*(4), 255. https://doi.org/10.30998/faktorexacta.v13i4.7964

Ayogu, I. I. (2020). Exploring multinomial naïve Bayes for Yorùbá text document classification. *Nigerian Journal of Technology*, *39*(2), 528–535. https://doi.org/10.4314/njt.v39i2.23

Dewi, C., Chen, R. C., Christanto, H. J., & Cauteruccio, F. (2023). Multinomial Naïve Bayes Classifier for Sentiment Analysis of Internet Movie Database. *Vietnam Journal of Computer Science*, *10*(4), 485–498. https://doi.org/10.1142/S2196888823500100

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020). A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews. *2020 International Conference on Contemporary Computing and Applications, IC3A 2020*, 217–220. https://doi.org/10.1109/IC3A48958.2020.233300

Farisi, A. A., Sibaroni, Y., & Faraby, S. Al. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012024

Fauzan, Abd. C., & Hikmah, K. (2022). IMPLEMENTASI ALGORITMA NAIVE BAYES DALAM ANALISIS POLARISASI OPINI MASYARAKAT TERKAIT VAKSIN COVID-19. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *7*(2), 122–128. https://doi.org/10.36341/rabit.v7i2.2403

Hasibuan, E., & Heriyanto, E. A. (2022). ANALISIS SENTIMEN PADA ULASAN APLIKASI AMAZON SHOPPING DI GOOGLE PLAY STORE MENGGUNAKAN NAIVE BAYES CLASSIFIER. *JTS*, *1*(3).

Hendrastuty, N., Rahman Isnain, A., & Yanti Rahmadhani, A. (2021). *Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine*. *6*(3). http://situs.com

Jeremy Andre, S., Tresna Maulana, F., & Aryo, N. (2019). *Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor*. https://t.co/9WloaWpfD5

Novianti, E. W., & Wibowo, W. (2022). *Analisis Sentimen Pengguna Twitter terhadap Program Kartu Prakerja di Tengah Pandemi Covid-19 Menggunakan Metode Naïve Bayes Classifier*.

Prayoga, M. B., Cahyono, N., & Subektiningsih, K. (2025). PENERAPAN GRID SEARCH UNTUK OPTIMASI MODEL MACHINE LEARNING DALAM KLASIFIKASI SENTIMEN KOMENTAR YOUTUBE. In *Jurnal Mahasiswa Teknik Informatika)* (Vol. 9, Number 3).

Ramadhan, N. G., & Adhinata, F. D. (2022). Sentiment analysis on vaccine COVID-19 using word count and Gaussian Naïve Bayes. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(1), 1765–1772. https://doi.org/10.11591/ijeecs.v26.i3.pp1765-1772

Sanusi, R., Dwi Astuti, F., Indra, D., & Buryadi, Y. (2021). ANALISIS SENTIMEN PADA TWITTER TERHADAP PROGRAM KARTU PRA KERJA DENGAN RECURRENT NEURAL NETWORK. In *Jurnal Informatika dan Komputer)* (Vol. 5, Number 2).

Septianingrum, F., & Irawan, A. S. Y. (2021). Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(3), 799. https://doi.org/10.30865/mib.v5i3.2983

Suhandi, Wahyu, W., & Icin, Q. (n.d.). *DINAMIKA PERMASALAHAN KETENAGAKERJAAN DAN PENGANGGURAN DI INDONESIA*.

Winahyu, J., & Suharjo, I. (2021). Aplikasi Web Analisis Sentimen Dengan Algoritma Multinomial Naïve Bayes. *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, *10*(2).

Zulfikar, W. B., Atmadja, A. R., & Pratama, S. F. (2023). Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes. *Scientific Journal of Informatics*, *10*(1), 25–34. https://doi.org/10.15294/sji.v10i1.39952

LAMPIRAN

LAMPIRAN-LAMPIRAN

Lampiran 1 Perhitungan TF-IDF.

D1: “listrik pulsa mahal”

D2: “daftar gagal butuh listrik”

D3: “pendapat pln korupsi”

Jumlah dokumen: $N = 3$

Selanjutnya *vocabulary* diperoleh dari seluruh kata unik pada kumpulan dokumen.

Kata
listrik
pulsa
mahal
daftar
gagal
butuh
pendapat
pln
korupsi

Jumlah *vocabulary*: $|V| = 9$

Menghitung Term Frequency (TF) menunjukkan jumlah kemunculan kata dalam suatu dokumen.

Dokumen	listrik	pulsa	mahal	daftar	gagal	butuh	pendapat	pln	korupsi
D1	1	1	1	0	0	0	0	0	0
D2	1	0	0	1	1	1	0	0	0
D3	0	0	0	0	0	0	1	1	1

Selanjutnya menghitung Document Frequency (DF) adalah jumlah dokumen yang mengandung kata tertentu.

Kata	df
listrik	2
pulsa	1
mahal	1
daftar	1
gagal	1
butuh	1
pendapat	1
pln	1
korupsi	1

Lalu, menghitung IDF menggunakan rumus:

$$idf(t) = \ln\left(\frac{1+N}{1+df(t)}\right) + 1$$

Dengan: *N = 3*

IDF untuk kata dengan df = 2

$$idf = \ln\left(\frac{4}{3}\right) = +1$$

$$idf = 1.2877$$

Sehingga idf(listrik) = 1.2287

IDF untuk kata dengan df = 1

$$idf = \ln\left(\frac{4}{2}\right) = +1$$

$$idf = 1.6931$$

Sehingga:

- idf(listrik) = 1.6931
- idf(pulsa) = 1.6931
- idf(mahal) = 1.6931
- idf(daftar) = 1.6931
- idf(gagal) = 1.6931
- idf(butuh) = 1.6931
- idf(pendapat) = 1.6931
- idf(pln) = 1.6931
- idf(korupsi) = 1.6931

Dan terakhir, menghitung bobot TF-IDF dengan rumus:

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

TF-IDF Dokumen 1 (listrik pulsa mahal)

Kata	TF	IDF	TF-IDF
listrik	1	1.2877	1.2877
pulsa	1	1.6931	1.6931
mahal	1	1.6931	1.6931

TF-IDF Dokumen 2 (daftar gagal butuh listrik)

Kata	TF	IDF	TF-IDF
listrik	1	1.2877	1.2877
daftar	1	1.6931	1.6931
gagal	1	1.6931	1.6931
butuh	1	1.6931	1.6931

TF-IDF Dokumen 3 (pendapat pln korupsi)

Kata	TF	IDF	TF-IDF
pendapat	1	1.6931	1.6931
pln	1	1.6931	1.6931
korupsi	1	1.6931	1.6931

Sehingga diperoleh Matriks TF-IDF:

Dokumen	listrik	pulsa	mahal	daftar	gagal	butuh	pendapat	pln	korupsi
D1	1.2877	1.6931	1.6931	0	0	0	0	0	0
D2	1.2877	0	0	1.6931	1.6931	1.6931	0	0	0
D3	0	0	0	0	0	0	1.6931	1.6931	1.6931

Lampiran 2 Perhitungan Multinomial Naïve Bayes.

- **Data Training**
  - D1: listrik pulsa mahal → **Negatif**
  - D2: daftar gagal butuh listrik → **Negatif**
  - D3: pelayanan listrik stabil → **Positif**

Vocabulary yg terbentuk:

$$V = \{listrik, pulsa, mahal, daftar, gagal, butuh, pelayanan, stabil\}$$

Jumlah vocabulary:

$$|V| = 8$$

- **Matriks TF-IDF**

Dok	listrik	pulsa	mahal	daftar	gagal	butuh	pelayanan
D1(Neg)	1.0000	1.6931	1.6931	0	0	0	0
D2(Neg)	1.0000	0	0	1.6931	1.6931	1.6931	0
D3(Pos)	1.0000	0	0	0	0	0	1.6931

Keterangan:
  - Bobot listrik = 1.0000 karena muncul disemua dokumen
  - Kata lain memiliki bobot lebih tinggi karena lebih spesifik pada dokumen tertentu.

- **Perhitungan Prior Probability**
  - Kelas Negatif = 2

- ❖ Kelas Positif = 1
- ❖ Total Dokumen = 3

Kelas	Jumlah Dokumen	Prior
Negatif	2	$\frac{2}{3} = 0.6667$
Positif	1	$\frac{1}{3} = 0.3333$

- **Menjumlahkan Bobot TF-IDF per kelas**
  - ❖ Kelas Negatif

Kata	Total Bobot
listrik	2.0000
pulsa	1.6931
mahal	1.6931
daftar	1.6931
gagal	1.6931
butuh	1.6931
pelayanan	0
stabil	0

Total kelas negatif:

$$\sum_j N_{j,neg} = 2.0000 + 5(1.6931) = 10.4655$$

  - ❖ Kelas Positif:

Kata	Total Bobot
listrik	1.0000
pelayanan	1.6931
stabil	1.6931
daftar	0
gagal	0
butuh	0
pulsa	0
mahal	0

Total seluruh bobot kelas positif:

$$\sum_j N_{j,pos} = 1.0000 + 1.6931 + 1.6931 = 4.3862$$

- **Menghitung Likelihood**

Pada contoh ini digunakan smooting $\alpha = 1$

  - ❖ Likelihood untuk kelas negatif

$$\sum_j N_{j,neg} + \alpha|V| = 10.4655 + 8 = 18.4655$$

Contoh perhitungan kelas *listrik* kelas negatif:

$$P(listrik|neg) = \frac{2.0000 + 1}{18.4655} = \frac{3.0000}{18.4655} = 0.1625$$

Untuk kelas *mahal* kelas negatif:

$$P(mahal|neg) = \frac{1.6931 + 1}{18.4655} = \frac{2.6931}{18.4655} = 0.1459$$

Untuk kata yg tidak muncul, misalnya *pelayanan* kelas negatif:

$$P(pelayanan|neg) = \frac{0 + 1}{18.4655} = 0.0542$$

- Likelihood untuk kelas positif:

$$\sum_j N_{j,pos} + \alpha|V| = 4.3862 + 8 = 12.3862$$

Contoh perhitungan kelas *listrik* kelas positif:

$$P(listrik|pos) = \frac{1.0000 + 1}{12.3862} = \frac{2.0000}{12.3862} = 0.1615$$

Untuk kelas *stabil* kelas positif:

$$P(stabil|pos) = \frac{1.6931 + 1}{12.3862} = \frac{2.6931}{12.3862} = 0.2174$$

Untuk kata yg tidak muncul, misalnya *mahal* kelas negatif:

$$P(mahal|pos) = \frac{0 + 1}{12.3862} = 0.0807$$

- **Prediksi Dokumen Uji**

$$\alpha = "listrik\ mahal"$$

- Skor untuk kelas Negatif:

$$score_{neg} = log\ P(neg) + \alpha_{listrik} log\ P(listrik|neg) + \alpha_{mahal} log\ P(mahal|neg)$$
$$score_{neg} = log(0.6667) + 1.0000\log(0.1625) + 1.6931\log(0.1459)$$
$$score_{neg} = -0.4055 + 1.0000\ (1.8170) + 1.6931(1.9247)$$
$$score_{neg} = -0.4055 - 1.8170 - 3.2595$$
$$score_{neg} = -5.4820$$

- Skor untuk kelas Positif:

$$score_{pos} = log\ P(pos) + \alpha_{listrik} log\ P(listrik|pos) + \alpha_{mahal} log\ P(mahal|pos)$$
$$score_{pos} = log(0.3333) + 1.0000\log(0.1615) + 1.6931\log(0.0807)$$
$$score_{pos} = -1.0986 + 1.0000(-1.8231) + 1.6931(-2.5171)$$
$$score_{pos} = -1.0986 - 1.8231 - 4.2620$$
$$score_{pos} = -7.1837$$

- **Hasil Klasifikasi**

$$score_{neg} > score_{pos}$$

Maka dokumen uji “listrik mahal” diklasifikasikan sebagai **Negatif**.

Lampiran 3 Perhitungan Gaussian Naïve Bayes.

- **Data Training**
  - ❖ D1: listrik pulsa mahal → **Negatif**
  - ❖ D2: daftar gagal butuh listrik → **Negatif**
  - ❖ D3: pelayanan listrik stabil → **Positif**

Dokumen	Komponen-1	Komponen-2	Kelas
D1	1.20	0.80	Negatif
D2	1.00	1.10	Negatif
D3	0.30	2.10	Positif

- Perhitungan Prior Probability

  Jumlah dokumen:
  - ❖ Kelas Negatif = 2
  - ❖ Kelas Positif = 1
  - ❖ Total Dokumen = 3

Kelas	Jumlah Dokumen	Prior
Negatif	2	$\frac{2}{3} = 0.6667$
Positif	1	$\frac{1}{3} = 0.3333$

- Menghitung Mean dan Varian Tiap Kelas
  - ❖ Kelas Negatif (D1 dan D2)

    Mean kelas negatif komponen-1:

$$\mu_{1,neg} = \frac{1.20 + 1.00}{2} = 1.10$$

    Mean kelas negatif komponen-2:

$$\mu_{2,neg} = \frac{0.80 + 1.10}{2} = 0.95$$

    Varians kelas negatif komponen-1

$$\sigma^2_{1,neg} = \frac{(1.20 - 1.00)^2 + (1.00 - 1.10)^2}{2}$$

$$\sigma^2_{1,neg} = \frac{0.01 + 0.01}{2} = 0.01$$

    Varians kelas negatif komponen-2

$$\sigma^2_{2,neg} = \frac{(0.80 - 0.95)^2 + (1.10 - 0.95)^2}{2}$$
$$\sigma^2_{2,neg} = \frac{0.0225 + 0.0225}{2} = 0.0225$$

Fitur	Mean (μ)	Varians (σ)
Komponen-1	1.10	0.0100
Komponen-2	0.95	0.0225

❖ Kelas Positif (D3 = 0.30, 2.10)

$$\mu_{1,pos} = 0.30$$
$$\mu_{2,pos} = 2.10$$

$$\sigma^2_{1,pos} = 0.01$$
$$\sigma^2_{2,pos} = 0.01$$

Fitur	Mean (μ)	Varians (σ)
Komponen-1	0.30	0.0100
Komponen-2	2.10	0.0100

- Prediksi Dokumen Uji

Misalkan terdapat 1 dokumen uji setelah melakukan TF-IDF dan SVD menghasilkan fitur:

$$x = (1.05, 1.00)$$

Menghitung likelihood pada kelas Negatif komponen-1

$$P(x_1|neg) = \frac{1}{\sqrt{2\pi(0.01)}} \exp\left(-\frac{(1.05 - 1.10)^2}{2(0.01)}\right)$$
$$= \frac{1}{\sqrt{0.0628}} \exp\left(-\frac{0.0025}{0.02}\right)$$
$$= \frac{1}{\sqrt{0.2507}} \exp(-0.125)$$
$$= 3.989 \times 0.8825 = 3.520$$

Menghitung likelihood pada kelas Negatif komponen-2

$$P(x_2|neg) = \frac{1}{\sqrt{2\pi(0.0225)}} \exp\left(-\frac{(1.00 - 0.95)^2}{2(0.0225)}\right)$$
$$= \frac{1}{\sqrt{0.1414}} \exp\left(-\frac{0.0025}{0.045}\right)$$
$$= \frac{1}{\sqrt{0.3760}} \exp(-0.0556)$$

$$= 2.660 \times 0.9460 = 2.516$$

Menghitung likelihood pada kelas Positif komponen-1

$$P(x_1|pos) = \frac{1}{\sqrt{2\pi(0.01)}} \exp\left(-\frac{(1.05-0.30)^2}{2(0.01)}\right)$$
$$= 3.989 \times \exp\left(-\frac{0.5625}{0.02}\right)$$
$$= 3.989 \times \exp(-28.125) \approx 0$$

Menghitung likelihood pada kelas Positif komponen-2

$$P(x_2|pos) = \frac{1}{\sqrt{2\pi(0.01)}} \exp\left(-\frac{(1.00-2.10)^2}{2(0.01)}\right)$$
$$= 3.989 \times \exp\left(-\frac{1.21}{0.02}\right)$$
$$= 3.989 \times \exp(-60.25) \approx 0$$

- Menghitung Skor Log Posterior

  Skor kelas Negatif:

$$score_{neg} = \log P(neg) + \log P(x_1|neg) + \log P(x_2|neg)$$
$$= \log(0.6667) + \log(3.520) + \log(2.516)$$
$$= -0.4055 + 1.2585 + 0.9228$$
$$= 1.7758$$

  Skor kelas Positif:

  Karena likelihood pada kedua fitur sangat kecil, maka skor kelas positif menjadi jauh lebih rendah.

$$score_{pos} = \log(0.3333) + \log P(x_1|pos) + \log P(x_2|pos)$$
$$score_{neg} \ll score_{pos}$$

- Hasil Klasifikasi

  Karena:

$$score_{neg} > score_{pos}$$

  Maka dikumen diuji dengan fitur:

$$x = (1.05, 1.00)$$

  Diklasifikasikan ke dalam kelas:

Negatif

Lampiran 4 Perhitungan Mix Distribusi

- Nilai Probabilitas MNB dan GNB
  Dokumen uji:

$$x = "listrik\ mahal"$$

Model	Kelas Negatif	Kelas Positif
MNB	0.7300	0.2700
GNB	0.6800	0.3200

  Misal hasil GridSearchCV menentukan:

$$\alpha = 0.6$$

  Maka bobot GNB adalah:

$$1 - \alpha = 0.4$$

- Menghitung Log-Probability
  Log-Probability MNB

$$\log P_{MNB}(neg) = \log(0.7300) = -0.3147$$
$$\log P_{MNB}(pos) = \log(0.2700) = -1.3093$$

  Log-Probability GNB

$$\log P_{GNB}(neg) = \log(0.6800) = -0.3857$$
$$\log P_{GNB}(pos) = \log(0.3200) = -1.1394$$

- Menghitung Mix Distribusi Log-Probability
  $\log P_{mix}(y = c|x) = \alpha \log P_{MNB}(y = c|x) + (1 - \alpha) \log P_{GNB}(y = c|x)$
  dengan $\alpha = 0.6$
  Untuk kelas Negatif:

$$\log P_H(neg) = 0.6(-0.3147) + 0.4(-0.3857)$$
$$= -0.1888 + (-0.1543)$$
$$= -0.3431$$

  Untuk kelas Positif:

$$\log P_H(pos) = 0.6(-1.3093) + 0.4(-1.1394)$$
$$= -0.7856 + (-0.4558)$$
$$= -1.2414$$

- Konversi ke Nilai Eksponensial:
  Agar dapat dinormalisasi menjadi probabilitas, log-probability diubah dengan fungsi eksponensial.
  Untuk kelas Negatif

$$P'_H(neg) = e^{-0.3431} = 0.7097$$

  Untuk kelas Positif

$$P'_H(pos) = e^{-1.2414} = 0.2890$$

- Normalisasi Probabilitas

$$0.7097 + 0.2890 = 0.9987$$

Maka probabilitas akhir Mix Distribusi adalah:
Kelas Negatif:

$$P_H(neg) = \frac{0.7097}{0.9987} = 0.7106$$

Kelas Positif:

$$P_H(pos) = \frac{0.2890}{0.9987} = 0.2894$$

- Hasil Klasifikasi

$$P_H(neg) > P_H(pos)$$

Maka dokumen uji "listrik mahal" diklasifikasikan ke dalam kelas:

Negatif

Kelas	P_{MNB}	P_{GNB}	$\log P_{MNB}$	$\log P_{GNB}$	$\log P_H$	e^{logP_H}	Proba Akhir
Negatif	0.7300	0.6800	-0.3147	-0.3857	-0.3431	0.7097	0.7106
Positif	0.2700	0.3200	-1.3093	-1.1394	-1.2414	0.2890	0.2894