

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN KARTU PRAKERJA
MENGUNAKAN METODE *NAIVE BAYES***

SKRIPSI

**Oleh:
ADISA DWI WANTI
NIM. 19650037**



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023**

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN KARTU PRAKERJA
MENGUNAKAN METODE *NAIVE BAYES***

SKRIPSI

Oleh:
ADISA DWI WANTI
NIM. 19650037

**Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023**

HALAMAN PERSETUJUAN

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN KARTU PRAKERJA
MENGUNAKAN METODE *NAIVE BAYES***

SKRIPSI

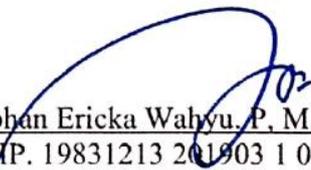
Oleh:
ADISA DWI WANTI
NIM. 19650037

Telah Diperiksa dan Disetujui untuk Diuji
Tanggal: 05 Mei 2023

Pembimbing I


Fajar Rohman Hariri, M.Kom
NIP. 19890515 201801 1 001

Pembimbing II


Johan Ericka Wahyu, P. M.Kom
NIP. 19831213 201903 1 004

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




Dr. Fatchul Kurniawan, M.MT .IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN KARTU PRAKERJA
MENGUNAKAN METODE *NAIVE BAYES***

SKRIPSI

Oleh:
ADISA DWI WANTI
NIM. 19650037

Telah Dipertahankan di Depan Dewan Penguji Skripsi
Dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 15 Juni 2023

Susunan Dewan Penguji

Ketua Penguji : Prof. Dr. Suhartono, M.Kom
NIP. 19680519 200312 1 001

Anggota Penguji I : Agung Teguh Wibowo Almais, M.T
NIDT. 19860103 20180201 1 235

Anggota Penguji II : Fajar Rohman Hariri, M.Kom
NIP. 19890515 201801 1 001

Anggota Penguji III : Johan Ericka Wahyu Prakasa, M.Kom
NIP. 19831213 201903 1 004



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



Dr. Fachrud Kurniawan, M.MT.,IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Adisa Dwi Wanti
NIM : 19650037
Fakultas/Program Studi : Sains dan Teknologi/Teknik Informatika
Judul Skripsi : Analisis Sentimen Mengenai Kebijakan Kartu
Prakerja Menggunakan Metode *Naive Bayes*

Menyatakan dengan sebenarnya bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilan daya, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka. Apabila di kemudian hari terbukti atau dapat dibuktikan skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan saya tersebut.

Malang, 15 Juni 2023
Yang membuat pernyataan,



Adisa Dwi Wanti
NIM. 19650037

HALAMAN MOTTO

“Setetes keringat orang tuaku, seribu langkahku untuk maju.”

HALAMAN PERSEMBAHAN

Skripsi ini saya persembahkan untuk Kedua Orang Tua saya,
Bapak Sampe dan Ibu Juwanik'ah, Keluarga,
Seluruh Dosen, Sahabat, Teman-Teman Seperjuangan,
Seluruh Pihak yang Terlibat, dan Diri Saya Sendiri

Terima kasih

KATA PENGANTAR

Bismillahirrahmanirrahim, Alhamdulillah segala puji syukur penulis limpahkan atas kehadiran Allah SWT yang telah melimpahkan rahmat dan karunia-Nya, sehingga peneliti mampu menyelesaikan skripsi dengan baik. Salawat serta salam senantiasa tercurahkan kepada Baginda Nabi Muhammad SAW yang telah memberikan panutan bagi seluruh umat manusia. Semoga kita termasuk dalam golongan yang mendapatkan pertolongan Nabi Muhammad SAW di hari akhir.

Penulis menyadari bahwa penyusunan skripsi ini tak luput dari dukungan dan bantuan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis menyampaikan terimakasih kepada:

1. Prof. Dr. M. Zainuddin, MA selaku rektor Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
2. Dr. Sri Hariani, M.Si selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, M.MT.,IPM selaku ketua Program Studi Teknik Informatika Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang.
4. Fajar Rohman Hariri, M.Kom selaku dosen pembimbing I yang telah bersedia meluangkan waktunya untuk membimbing dan memberikan arahan dalam penyusunan skripsi ini.
5. Johan Ericka Wahyu Prakasa, M.Kom selaku dosen pembimbing II yang telah membimbing serta memberikan kritik dan saran dalam penyusunan skripsi ini.

6. Prof. Dr. Suhartono, M.Kom selaku dosen penguji I yang telah memberikan saran dan arahan dari seminar proposal hingga sidang skripsi penulis.
7. Agung Teguh Wibowo Almais, M.T selaku dosen penguji II yang telah memberikan masukan pada seluruh proses dari seminar proposal hingga sidang skripsi penulis.
8. Seluruh dosen dan jajaran staf Program Studi Teknik Informatika yang telah memberikan ilmu serta yang secara langsung maupun tidak langsung terlibat dalam proses penyusunan skripsi ini.
9. Kedua orang tua penulis, bapak Sampe dan ibu Juwanik'ah yang tidak pernah berhenti memberikan semangat, dukungan, dan doanya, serta saudara-saudara kandung penulis Ismail Bondan Efendi dan Imam Firmansyah yang selalu menjadi penguat untuk menyelesaikan skripsi ini dengan baik dan tepat waktu.
10. Keluarga besar penulis, Pakde dan Bude, bapak Tarjo, ibu Samikati, bapak Paino, dan ibu Suwarni, sepupu-sepupu penulis, Elisa Sabahwanti dan Aris Winarta, serta keponakan-keponakan penulis Kahiyang Putri Kanaya dan Muhammad Althaf Artanabil yang selalu memberikan semangat, dukungan, dan doa-doanya sehingga penulis dapat menyelesaikan skripsi ini dengan baik dan tepat waktu.
11. Sahabat-sahabat seperjuangan Teknik Informatika angkatan 2019 "ALIEN" dan seluruh keluarga besar Teknik Informatika Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang yang telah memberikan dukungan dan bantuan selama penyusunan skripsi ini.

12. Seluruh pihak yang selalu berada di samping penulis, teruntuk Edla, Salma, Laila, Imada, Widya, Salsa, Nada, Fifi, Tutut, Alfiyatus, Moudy, Fani dan Mafa yang senantiasa memberikan dukungan dan doa yang luar biasa, semoga Allah SWT membalas segala kebaikan yang telah diberikan berkali-kali lipat, aamiin.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih jauh dari kesempurnaan, baik dari segi keilmuan dan kepenulisan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun. Semoga penulisan skripsi ini dapat bermanfaat bagi banyak pihak.

Malang, 15 Juni 2023

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN.....	iii
PERNYATAAN KEASLIAN TULISAN.....	iv
HALAMAN MOTTO	v
HALAMAN PERSEMBAHAN	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	xi
DAFTAR GAMBAR.....	xiii
DAFTAR TABEL	xiv
ABSTRAK	xv
ABSTRACT	xvi
البحث مستخلص	xvii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Pernyataan Masalah	7
1.3 Tujuan Penelitian	8
1.4 Batasan Masalah	8
1.5 Manfaat Penelitian	8
BAB II STUDI PUSTAKA	9
2.1 Penelitian Terdahulu	9
2.2 Landasan Teori.....	13
2.2.1 Media Sosial Instagram.....	13
2.2.2 Kartu Prakerja	15
2.2.3 Analisis Sentimen.....	17
2.2.4 <i>Text Preprocessing</i>	18
2.2.5 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	20
2.2.6 <i>K-Fold Cross Validation</i>	22
2.2.7 <i>Naive Bayes Classifier</i>	23
BAB III METODOLOGI PENELITIAN	26
3.1 Prosedur Penelitian	26
3.2 Pengumpulan Data	27
3.3 Pelabelan Data	28
3.4 Desain Sistem.....	29
3.5 <i>Text Preprocessing</i>	29
3.5.1 <i>Cleansing</i>	30
3.5.2 <i>Case Folding</i>	31
3.5.3 <i>Tokenizing</i>	31
3.5.4 <i>Normalization</i>	31
3.5.5 <i>Stopword Removal</i>	32
3.5.6 <i>Stemming</i>	32
3.6 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	33
3.7 <i>Naive Bayes</i>	35
3.7.1 <i>Training</i>	37

3.8 <i>K-Fold Cross Validation</i>	39
3.9 Skenario Pengujian	40
BAB IV UJI COBA DAN PEMBAHASAN	42
4.1 Data Penelitian	42
4.2 Uji Coba	44
4.2.1 Tahap Penelitian.....	44
4.3 Hasil Uji Coba.....	54
4.4 Pembahasan.....	59
BAB V KESIMPULAN DAN SARAN.....	64
5.1 Kesimpulan	64
5.2 Saran	65
DAFTAR PUSTAKA	

DAFTAR GAMBAR

Gambar 3.1	Prosedur Penelitian.....	26
Gambar 3.2	Desain Sistem.....	29
Gambar 3.3	Diagram Alir <i>Text Preprocessing</i>	30
Gambar 3.4	Diagram Alir Klasifikasi <i>Naive Bayes</i>	36
Gambar 4. 1	Grafik Perbandingan Kelas Sentimen	42
Gambar 4.2	<i>Pseudocode</i> Tahap <i>Cleansing</i>	45
Gambar 4.3	Hasil <i>Cleansing</i>	45
Gambar 4.4	<i>Pseudocode</i> Tahap <i>Case Folding</i>	46
Gambar 4.5	Hasil <i>Case Folding</i>	46
Gambar 4.6	<i>Pseudocode</i> Tahap <i>Tokenizing</i>	47
Gambar 4.7	Hasil <i>Tokenizing</i>	47
Gambar 4.8	<i>Pseudocode</i> Tahap <i>Normalization</i>	48
Gambar 4.9	Hasil <i>Normalization</i>	48
Gambar 4.10	<i>Pseudocode</i> Tahap <i>Stopword Removal</i>	48
Gambar 4.11	Hasil <i>Stopword Removal</i>	49
Gambar 4.12	<i>Pseudocode</i> Tahap <i>Stemming</i>	49
Gambar 4.13	Hasil <i>Stemming</i>	50
Gambar 4. 14	Nilai Akurasi <i>Testing</i>	55
Gambar 4.15	Hasil <i>Confusion Matrix</i>	55

DAFTAR TABEL

Tabel 2.1 Perbedaan Penelitian Terdahulu dengan Penelitian Saat ini	13
Tabel 3.1 Kriteria Pelabelan Data	28
Tabel 3.2 Data Terlabel.....	28
Tabel 3.3 <i>Preprocessing</i> Data	32
Tabel 3.4 Data Sampel	33
Tabel 3.5 Perhitungan TF.....	34
Tabel 3. 6 Hasil Perhitungan TF-IDF	35
Tabel 3.7 Contoh Kasus Data <i>Training</i>	37
Tabel 3.8 Perhitungan Probabilitas Kata pada D1	38
Tabel 3.9 Perhitungan Probabilitas Kata pada D2	38
Tabel 3.10 Perhitungan Probabilitas Kata pada D3	38
Tabel 3.11 Perhitungan Probabilitas Kata pada D4	38
Tabel 3.12 Perhitungan Probabilitas Kata pada D5	39
Tabel 3.13 Perhitungan Probabilitas Kata pada D6	39
Tabel 3.14 Konsep <i>Confusion Matrix</i>	40
Tabel 4.1 Sampel Data Komentar	43
Tabel 4.2 Contoh Data <i>Testing</i>	51
Tabel 4.3 Nilai Probabilitas Data <i>Testing</i>	54
Tabel 4.4 Perbandingan Skor <i>K-fold Cross Validation</i>	57
Tabel 4.5 Hasil Klasifikasi	58

ABSTRAK

Wanti, Adisa Dwi. 2023. **Analisis Sentimen Mengenai Kebijakan Kartu Prakerja Menggunakan Metode Naive Bayes**. Skripsi. Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Fajar Rohman Hariri, M.Kom. (II) Johan Ericka Wahyu Prakasa, M.Kom.

Kata Kunci: Analisis Sentimen, Naive Bayes, Kartu Prakerja, K-fold Cross Validation

Peluncuran kartu prakerja yang diklaim untuk meningkatkan kondisi perekonomian dengan cara mengembangkan kompetensi tenaga kerja, menuai berbagai pro dan kontra dari masyarakat. pembahasan seputar sistem seleksi, konten pelatihan, hingga besaran anggaran memunculkan beragam reaksi. pendapat dapat dibagi menjadi dua kelompok, yaitu sentimen positif dan sentimen negatif. Penelitian ini bertujuan untuk mengimplementasikan metode *Naive Bayes* dalam proses analisis sentimen terhadap kebijakan kartu prakerja. Data yang digunakan terdiri dari 2200 komentar instagram para akun @prakerja.co.id from 26 agustus hingga 9 desember 2022. Proses analisis sentimen dilakukan menggunakan metode *Naive Bayes*, karena metode ini memiliki karakteristik yang cepat dan akurasi yang tinggi ketika diaplikasikan dalam dataset yang besar dan beragam. Selain itu, dibandingkan dengan metode klasifikasi yang lain, metode *Naive Bayes* merupakan metode yang relatif sederhana dan hanya membutuhkan jumlah *training data* yang kecil. Validasi model dilakukan menggunakan teknik *K-fold Cross Validation* dengan nilai $k = 5, 10, 15,$ dan 20 pada model yang menggunakan teknik *Laplacian Correction* dan model yang tidak menggunakan teknik *Laplacian Correction*. Dari hasil analisis menunjukkan bahwa dari total 2200 komentar, sebanyak 1132 merupakan sentimen negatif dan 1068 merupakan sentimen positif. Hasil uji coba performa sistem menunjukkan nilai akurasi, presisi, *recall*, dan *f-measure* berturut-turut 90%, 91%, 88%, dan 89%. Hasil klasifikasi menunjukkan bahwa metode *Naive Bayes* memiliki performa yang baik dalam klasifikasi opini. Metode ini dapat digunakan untuk memahami analisis sentimen serupa di masa depan. Dalam kasus Kartu prakerja, temuan ini dapat menjadi masukan bagi para pemangku kepentingan untuk memperbaiki program di masa depan.

ABSTRACT

Wanti, Adisa Dwi. 2023. **Sentiment Analysis About Kartu Prakerja Policy Using Naive Bayes Method**. Undergraduate Thesis. Department of Teknik Informatika, Faculty of Sains and Technology, State Islamic University of Maulana Malik Ibrahim Malang, Supervisor: (I) Fajar Rohman Hariri, M.Kom. (II) Johan Ericka Wahyu Prakasa, M.Kom.

The launch of the pre-employment card, which is claimed to improve economic conditions by developing the competence of the workforce, has drawn various pros and cons from the community. discussions about the selection system, training content, to the amount of the budget have led to various reactions. opinions can be divided into two groups, namely positive sentiment and negative sentiment. This research aims to implement The Naive Bayes method in the sentiment analysis process of the pre-employment card policy. The data used consists of 2200 instagram comments on the @prakerja.co.id account from August 26 to December 9, 2022. The sentiment analysis process is carried out using The Naive Bayes method, because this method has fast characteristics and high accuracy when applied in large and diverse datasets. In addition, compared to other classification methods, The Naive Bayes method is relatively simple and only requires a small amount of training data. Model validation was conducted using The K-fold Cross Validation technique with $k = 5, 10, 15,$ and 20 values on models that use Laplacian Correction techniques and models that do not use Laplacian Correction techniques. The analysis shows that out of a total of 2200 comments, 1132 are negative sentiments and 1068 are positive sentiments. The system performance test results show the accuracy, precision, recall, and f-measure values of 90%, 91%, 88%, and 89%, respectively. The classification results show that The Naive Bayes method has good performance in opinion classification. This method can be used to understand similar sentiment analysis in the future. In the case of the Pre-Employment Card, these findings can be an input for stakeholders to improve the program in the future.

Keywords: Sentiment Analysis, Naive Bayes, Kartu Prakerja, K-fold Cross Validation

مستخلص البحث

واني، أديسا دوي. 2023. تحليل المشاعر نحو سياسة بطاقة العمل باستخدام طريقة البايز الساذج. رسالة بكالوريوس. قسم تكنولوجيا المعلومات، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم مالانغ الإسلامية الدولية، مشرف: (I) فجر رحمان هاريري، م. كوم. (II) جوهان إريكا واهيو براكاسا، م. كوم.

كلمات مفتاحية: تحليل المشاعر، البايز الساذج، بطاقة العمل، التقسيم المتقاطع K-fold.

إطلاق بطاقة العمل المزعومة لتحسين الحالة الاقتصادية عن طريق تطوير كفاءة القوى العاملة قد أثار مختلف الآراء المؤيدة والمعارضة من قبل المجتمع. تناول المناقشات حول نظام الاختيار، ومحتوى التدريب، وحجم الميزانية مجموعة متنوعة من الردود. يمكن تقسيم الآراء إلى مجموعتين، وهما المشاعر الإيجابية والمشاعر السلبية. يهدف هذا البحث إلى تنفيذ طريقة البايز الساذج في عملية تحليل المشاعر تجاه سياسة بطاقة العمل. تتكون البيانات المستخدمة من 2200 تعليق على حسابات إنستغرام @prakerja.go.id من 26 أغسطس إلى 9 ديسمبر 2022. يتم إجراء عملية تحليل المشاعر باستخدام طريقة البايز الساذج، لأن هذه الطريقة لها خصائص سريعة ودقة عالية عند تطبيقها على مجموعة بيانات كبيرة ومتنوعة. بالإضافة إلى ذلك، مقارنة بطرق التصنيف الأخرى، فإن طريقة البايز الساذج هي طريقة بسيطة نسبياً وتتطلب فقط كمية صغيرة من بيانات التدريب. يتم تنفيذ تحقق النموذج باستخدام تقنية الصليب المتعدد الطبقات مع قيمة 5 و 10 و 15 و 20 في النموذج الذي يستخدم تقنية تصحيح لابلاسيان والنموذج الذي لا يستخدم تقنية تصحيح لابلاسيان. تشير نتائج التحليل إلى أنه من أصل 2200 تعليق، يوجد 1132 تعليقاً سلبياً و 1068 تعليقاً إيجابياً. تظهر نتائج اختبار أداء النظام قيمة دقة ودقة واسترجاع وقياس f على التوالي 90% و 91% و 88% و 89%. تشير نتائج التصنيف إلى أن طريقة البايز الساذج لديها أداء جيد في تصنيف الآراء. يمكن استخدام هذه الطريقة لفهم تحليل المشاعر المماثلة في المستقبل. في حالة بطاقة العمل، يمكن أن تكون هذه النتائج مدخلاً لأصحاب المصلحة لتحسين البرنامج في المستقبل.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Negara berkembang sering menghadapi macam-macam problematika. Salah satu problematika itu adalah pengangguran. Begitu pula Indonesia, sebagai salah satu negara berkembang, pengangguran merupakan fenomena ketenagakerjaan yang sering terjadi di Indonesia. Pengangguran sendiri merupakan sebuah fenomena ketika tenaga kerja yang sudah memasuki usia siap kerja tidak dapat tertampung ke dalam lapangan kerja sehingga menimbulkan ketidakmampuan untuk memenuhi kebutuhan pokoknya (Pasuria & Triwahyuningtyas, 2022).

Kondisi perekonomian di Indonesia semakin memburuk ketika Covid-19 mulai merebak di Indonesia pada tahun 2020. Covid-19 tersebut membuat kondisi perekonomian dan perindustrian semakin melemah, yang berakibat pada banyaknya angka Pemutusan Hubungan Kerja (PHK) (Firnanda et al., 2021). PHK inilah yang menyebabkan angka pengangguran semakin meningkat. Sehingga, di tengah pandemi Covid-19, pemerintah Indonesia menerapkan sejumlah cara untuk memperkuat perekonomian negara. Percepatan penerbitan kartu prakerja adalah salah satu strategi yang diimplementasikan pemerintah Indonesia untuk meningkatkan perekonomian.

Kartu prakerja adalah salah satu kebijakan pemerintah khususnya yang dibuat oleh Kementerian Ketenagakerjaan yang bertujuan untuk mengembangkan kompetensi, meningkatkan produktivitas dan daya saing, serta mengembangkan

kewirausahaan yang dikemas dalam program bantuan biaya pelatihan dan insentif. Sasaran penduduk program kartu prakerja ini adalah penduduk yang sedang mencari pekerjaan, pekerja, atau buruh yang terkena dampak kehilangan pekerjaan, dan para usaha mikro yang mengalami penurunan daya beli atau bahkan kehilangan pekerjaan (Kemnaker, 2021). Pada awal peluncurannya, yaitu pada 11 April 2020, kartu prakerja ini dirancang sebagai program pelatihan untuk mengembangkan kompetensi serta meningkatkan produktivitas dan daya saing. Namun, untuk mencegah penurunan kesejahteraan masyarakat, program ini kemudian melibatkan skema bantuan sosial (bansos) tambahan. (Firnanda et al., 2021).

Dari awal peluncuran, hingga pendaftaran gelombang ke-47, program kartu prakerja ini masih menuai beberapa polemik (Handayani & Rachman, 2020). Mulai dari sistem seleksi yang kurang baik, konten pelatihan yang dianggap terlalu mahal dengan kualitas yang tidak terlalu baik, dianggap sebagai pemborosan anggaran, hingga masalah *conflict of interest* yang dialami oleh beberapa staf khusus presiden (Consuello, 2020). Meskipun menuai berbagai polemik, animo masyarakat terhadap kartu prakerja masih sangat tinggi. Hal tersebut menunjukkan bahwa sejatinya program kartu prakerja ini sangat dibutuhkan oleh masyarakat. Dengan keterbatasan ruang dan waktu, masyarakat umum kesulitan untuk menyampaikan aspirasi secara langsung mengenai kebijakan yang dibuat oleh pemerintah tersebut (Syarifuddin, 2020). Namun, dengan peningkatan internet yang begitu cepat, kini ada media sosial yang dapat dijadikan sebagai ruang publik baru untuk menyampaikan aspirasi masyarakat (Alvianda & Adikara, 2019). Salah satu platform jaringan sosial yang sedang terkenal pada era ini yaitu Instagram. Platform

ini banyak dimanfaatkan oleh pengguna. Program kartu prakerja sendiri memiliki akun instagram resmi dengan nama akun @prakerja.go.id. Di dalam akun instagram tersebut, masyarakat dapat dengan bebas untuk memberikan komentar atau opininya mengenai kebijakan kartu prakerja.

Komentar atau pendapat yang diberikan oleh masyarakat melalui platform media sosial Instagram umumnya dapat dibagi menjadi dua kategori utama, yaitu komentar positif dan komentar negatif. Komentar positif tentang kartu prakerja yang disampaikan oleh masyarakat sendiri kebanyakan mengenai kepuasan tentang berbagai jenis pelatihan yang ditawarkan, jumlah insentif yang dirasa cukup untuk mengurangi dampak pandemi, kegunaan pelatihan yang diberikan dalam meningkatkan keterampilan, hingga kesesuaian kurikulum dengan kebutuhan nyata di lapangan kerja. Sedangkan komentar negatif mengenai kartu prakerja meliputi mahalnya materi pelatihan dengan kualitas konten yang kurang, keterlambatan pencairan insentif. Buruknya sistem seleksi, server yang tidak stabil, serta kesulitan dalam melaporkan masalah yang dialami kepada *contact person* (Ardhani et al., 2021). Oleh karena itu, dengan banyaknya komentar atau pendapat yang disampaikan oleh masyarakat, dibutuhkan sebuah analisis sentimen yang dapat digunakan sebagai rujukan bagi pemerintah untuk mengkaji ulang atau bahkan memperbaiki kebijakan kartu prakerja, mulai dari sistem seleksi hingga perbaikan konten pelatihan.

Analisis sentimen merupakan suatu metode yang diimplementasikan dalam mengumpulkan pemahaman terkait aspirasi, pendapat, komentar, sikap, dan emosi individu terhadap suatu topik yang spesifik. Analisis sentimen sendiri dapat disebut

juga dengan *opinion mining*, yaitu salah satu bidang baru dalam *Natural Language Processing* (NLP) yang berisi proses pengolahan kata untuk menganalisis opini dari masyarakat (Deviyanto & Wahyudi, 2018). Tugas pokok dari analisis sentimen adalah untuk mengategorikan polaritas dari teks yang dapat berupa dokumen, kalimat, atau opini, dengan tujuan memahami apakah mereka memiliki sentimen positif, negatif, atau netral. Polaritas merupakan sebuah kecenderungan apakah teks yang berbentuk dokumen, kalimat, atau pendapat tersebut memiliki aspek positif ataupun negatif (Nugroho et al., 2015). Oleh karena itu, melalui penerapan analisis sentimen, perusahaan atau lembaga terkait dapat memperoleh wawasan tentang tanggapan masyarakat terhadap suatu produk atau layanan yang spesifik. Proses analisis sentimen perlu dilakukan, agar perusahaan atau instansi terkait dapat menjadi bahan evaluasi untuk memperbaiki produk ataupun kinerjanya.

Seluruh kebijakan-kebijakan yang ditetapkan oleh pemerintah tersebut dilakukan semata-mata untuk meningkatkan taraf perekonomian masyarakat. Karena pemerintahan yang baik adalah pemerintahan yang setiap keputusannya mampu menjadikan masyarakatnya sebagai bangsa yang makmur dan sejahtera. Hal ini dijelaskan secara khusus dalam Al-Qur'an ayat 58 Surat An-Nisa:

إِنَّ اللَّهَ يَأْمُرُكُمْ أَنْ تُؤَدُّوا الْأَمَانَاتِ إِلَىٰ أَهْلِهَا وَإِذَا حَكَمْتُمْ بَيْنَ النَّاسِ أَنْ تَحْكُمُوا بِالْعَدْلِ إِنَّ اللَّهَ نِعِمَّا يَعِظُكُمْ بِهِ إِنَّ اللَّهَ كَانَ سَمِيعًا بَصِيرًا

“*Sesungguhnya Allah menyuruh kamu menyampaikan amanat kepada yang berhak menerimanya, dan (menyuruh kamu) apabila menetapkan hukum diantara manusia supaya kamu menetapkan dengan adil. Sesungguhnya Allah memberi pengajaran yang sebaik-baiknya kepadamu. Sesungguhnya Allah adalah Maha mendengar lagi Maha melihat*” (Q.S. An-Nisa’: 58)

Berdasarkan penafsiran Kementerian Agama Republik Indonesia, ayat tersebut menginstruksikan untuk memberikan amanat kepada pihak yang berhak menerimanya. Dalam ayat ini, kata “amanat” memiliki pengertian yang sangat luas, baik itu amanat yang diberikan oleh Allah kepada umat-Nya maupun amanat yang dikeluarkan oleh individu satu sama lain maupun amanat yang ditujukan kepada diri sendiri. Selain itu, Q.S. An-Nisa ayat 58 tersebut lebih spesifik membahas mengenai sifat adil para pemimpin atau pemegang kekuasaan terhadap rakyatnya serta untuk selalu menjaga amanat yang telah dipercayakan kepada dirinya (Kemenag, 2022). Sehingga dapat disimpulkan bahwa pemerintah sebagai pemegang kekuasaan bertugas sebagai penentu keputusan, dan setiap keputusan tersebut harus ditetapkan dengan sebenar-benarnya dan harus membawa kebaikan bagi rakyatnya (Fuad, 2016).

Pada dasarnya, penentuan polaritas komentar positif dan komentar negatif dapat dilakukan secara manual. Namun, saat jumlah komentar yang dijadikan sebagai sumber opini semakin bertambah, maka waktu dan usaha yang dibutuhkan juga akan semakin bertambah. Karena itu, dalam studi ini, peneliti akan memanfaatkan metode Machine Learning untuk mengategorikan polaritas komentar. Machine Learning menyediakan berbagai metode klasifikasi untuk mengimplementasikan analisis sentimen, seperti *Naive Bayes Classifier*, *Support Vector Machine*, dan *Decision Tree*. Salah satu metode yang sering diimplementasikan untuk mengklasifikasikan dokumen adalah *Naive Bayes Classifier* (Nurhuda et al., 2016).

Metode *Naive Bayes* merupakan sebuah algoritma dengan pendekatan pengklasifikasian probabilitas yang sederhana. Cara kerjanya melibatkan perhitungan sejumlah probabilitas dengan menjumlahkan frekuensi kombinasi nilai pada dataset yang digunakan (Rezki et al., 2020). Metode *Naive Bayes* beroperasi dengan meramalkan probabilitas kejadian pada zaman yang akan datang berdasarkan pengetahuan pada masa lalu (Fauzan & Hikmah, 2022). Metode *Naive Bayes* menerapkan prinsip Teorema Bayes yang diusulkan oleh ilmuwan bernama Thomas Bayes (Apif Supriadi & Fatmasari, 2021). Metode *Naive Bayes* mempunyai sejumlah keutamaan, termasuk kecepatan dan akurasi yang tinggi ketika diterapkan pada dataset yang besar dan beragam (Fauzan & Hikmah, 2022). Tak hanya itu, dibandingkan dengan metode klasifikasi lainnya, *Naive Bayes* adalah metode yang relatif mudah dan hanya memerlukan sejumlah kecil data pelatihan (Septianingrum & Irawan, 2021).

Dalam suatu studi yang berjudul analisis sentimen terhadap pendapat masyarakat mengenai vaksin COVID-19, digunakan metode klasifikasi *Naive Bayes Classifier*. Temuan studi tersebut mengindikasikan bahwa metode *Naive Bayes* memperlihatkan performa yang sangat baik dalam mengklasifikasikan polaritas opini, dengan mencapai tingkat akurasi sebesar 93% (Yulita et al., 2021). Sedangkan, studi tentang analisis sentimen opini publik mengenai covid-19 pada twitter menggunakan metode *Naive Bayes* dan KNN. Dalam penelitian tersebut, dua metode klasifikasi digunakan untuk menganalisis sentimen opini publik terkait COVID-19. Selanjutnya, hasil pengklasifikasian dari kedua metode tersebut akan dipadankan guna mengetahui metode yang memiliki tingkat akurasi yang lebih

tinggi. Sehingga, peneliti menyimpulkan bahwa metode *Naive Bayes* mengungguli KNN dalam hal tingkat akurasi. *Naive Bayes* mencapai tingkat akurasi sebesar 63,21%, sedangkan KNN mencapai tingkat akurasi sebesar 58,10%, berdasarkan temuan dari penelitian tersebut (Syarifuddinn, 2020). Berdasarkan justifikasi yang diberikan, peneliti menerapkan pendekatan klasifikasi *Naive Bayes*.

Dengan mempertimbangkan masalah dan konteks yang ada, penelitian dilakukan untuk menganalisis sentimen terhadap kebijakan kartu prakerja menggunakan metode *Naive Bayes*. Tujuannya adalah untuk mengukur sentimen opini masyarakat Indonesia terkait kebijakan tersebut. Tujuan dari penelitian ini adalah untuk mempelajari penerapan metode *Naive Bayes* dalam analisis sentimen terhadap kebijakan kartu prakerja, serta mengevaluasi sejauh mana tingkat keakuratan klasifikasi yang dihasilkan oleh metode *Naive Bayes*.

1.2 Pernyataan Masalah

Berdasarkan pemaparan pada konteks sebelumnya, rumusan masalah dinyatakan sebagai berikut:

1. Bagaimana mengimplementasikan metode *Naive Bayes* dalam proses analisis sentimen terhadap kebijakan kartu prakerja?
2. Seberapa besar nilai *Confusion Matrix* yang didapatkan dalam hasil analisis sentimen terhadap kebijakan kartu prakerja menggunakan metode *Naive Bayes*?

1.3 Tujuan Penelitian

Penelitian ini mempunyai beberapa tujuan, diantaranya:

1. Mengimplementasikan metode *Naive Bayes* dalam proses analisis sentimen terhadap kebijakan kartu prakerja.
2. Mengukur nilai *Confusion Matrix* yang didapatkan dalam hasil analisis sentimen terhadap kebijakan kartu prakerja menggunakan metode *Naive Bayes*.

1.4 Batasan Masalah

Demi menghindari deviasi topik, penulis menentukan batasan masalah berikut untuk mempersempit cakupan penelitian.:

1. Penggunaan data adalah data primer yaitu teks komentar pada media sosial Instagram pada akun @prakerja.go.id.
2. Data komentar yang diimplementasikan pada studi ini terdiri dari 2200 *records* data komentar berbahasa Indonesia yang diambil dalam rentang waktu antara 26 Agustus 2022 hingga 9 Desember 2022.
3. Terdapat dua teks komentar yaitu positif dan negatif, untuk proses klasifikasi.

1.5 Manfaat Penelitian

Beberapa keuntungan diharapkan dari studi ini:

1. Bagi pihak Kementerian Ketenagakerjaan Republik Indonesia, diharapkan dengan studi ini dapat dijadikan referensi untuk memperbaiki dan meningkatkan kualitas dari kebijakan kartu prakerja.
2. Bagi peneliti, diharapkan studi ini dapat memberikan kontribusi untuk perluasan pengetahuan dan menjadi sumber penelitian yang akan datang.

BAB II

STUDI PUSTAKA

Untuk membantu penulis lebih memahami topik-topik yang terkait dengan perkembangan studi ini dan perbandingan studi yang akan diteliti, tinjauan literatur dalam penelitian ini akan menggambarkan studi sebelumnya yang digunakan sebagai referensi. Berikut ini penulis berhasil menyusun contoh-contoh penelitian terdahulu dengan subyek yang diminati dalam penelitian yang sekarang sedang dikembangkan.

2.1 Penelitian Terdahulu

Yulita dan kawan-kawan melakukan penelitian sebelumnya di tahun 2021, penelitian ini mengeksplorasi analisis sentimen opini publik terkait vaksin covid-19 dengan menggunakan algoritma *Naive Bayes Classifier* (Yulita et al., 2021). Analisis itu diimplementasikan pada 3780 data tweet yang berhubungan tentang vaksinasi covid-19. Penelitian tersebut memiliki tujuan untuk menetapkan apakah langkah vaksinasi covid-19 dapat diterima secara luas oleh masyarakat atau tidak. Analisis yang dilakukan oleh peneliti membagi opini masyarakat menjadi tiga label, yaitu label positif, negatif, dan netral. Data berupa *tweet* tersebut masih mengandung kata-kata tidak baku yang akan menyusahakan tahap pengkategorian. Oleh karena itu, perlu diimplementasikan sebuah tahap yang disebut *preprocessing*. Tahap *preprocessing* merupakan sebuah proses yang digunakan untuk menghapus bagian data yang tidak relevan. Kemudian, diimplementasikan proses pengkategorian dengan bantuan algoritma *Naive Bayes Classifier*. Berdasarkan

analisis tersebut, didapati hasil opini positif sebesar 60,3%, opini netral sebesar 34,4%, dan opini negatif sebesar 5,4%. Akurasi yang dihasilkan pada penelitian ini adalah sebesar 93% (Yulita et al., 2021).

Kurniyatul Ainiah dalam sebuah penelitiannya menganalisis sentimen terhadap aplikasi Pedulilindungi dengan metode *Support Vector Machine* serta menerapkan fitur *Query Expansion Ranking* (Ainiah, 2022). Metode *Support Vector Machine* dipilih karena merupakan metode klasifikasi tunggal yang dapat melakukan regresi linier pada data nonlinier dengan dimensi besar, sedangkan pemilihan *Query Expansion Ranking* digunakan untuk mengurangi fitur yang ambigu sehingga akurasi yang dihasilkan dapat ditingkatkan. Pada penelitian tersebut, 2.500 komentar dari aplikasi Pedulilindungi di Google PlayStore merupakan data yang diteliti. Hasil klasifikasi pada penelitian tersebut menghasilkan keakuratan 89,96%, nilai presisi 98,08%, nilai *recall* 68%, serta nilai *f-measure* sebesar 80,31% (Ainiah, 2022).

Penelitian yang dilakukan oleh Muhammad Syarifuddin pada tahun 2020 menjelaskan bagaimana algoritma *Naive Bayes* dan KNN digunakan untuk menganalisis sentimen Twitter terkait Covid-19 (Syarifuddinn, 2020). Tujuan dari studi ini untuk memberi gambaran kepada masyarakat umum mengenai kecenderungan opini mengenai covid-19, serta bertujuan untuk membandingkan hasil keakuratan dari dua metode yang diimplementasikan, yaitu *Naive Bayes* dan KNN. Data yang digunakan diambil dari *website* resmi twitter dengan kata kunci “covid-19” pada tahun 2020. Kemudian didapati data acak sebanyak 1098 *tweets*. Data-data tersebut kemudian akan diolah dan dilakukan beberapa tahapan yaitu

mengkategorikannya menjadi dua bagian, yaitu komentar positif dan komentar negatif, kemudian tahap data *cleansing*, *preprocessing*, hingga didapati nilai akhir. Berdasarkan hasil analisis, didapati metode *Naive Bayes* memiliki nilai akurasi yang lebih tinggi dari metode KNN. Metode *Naive Bayes* menghasilkan nilai akurasi sebesar 63,21%, sedangkan metode KNN menghasilkan nilai akurasi sebesar 58,10%, serta didapati pula opini masyarakat yang cenderung positif. Jumlah opini positif yang didapat sebesar 610 tweet dan opini negatif sebesar 488 tweet, serta ditunjang dengan hasil pengujian *precision* menggunakan metode *Naive Bayes* yaitu opini positif dan negatif masing-masing sebesar 66,40% dan 58,94% (Syarifuddin, 2020).

Dalam salah satu penelitian, Muhammad Robih Thuuluz Zaman mengkategorikan opini tentang kebijakan publik merdeka belajar dengan menggunakan pendekatan *Naive Bayes* dan pemilihan fitur *Information Gain* (Zaman, 2020). Penelitian tersebut menggunakan metode *Naive Bayes* dikarenakan metode itu adalah metode pengkategorian yang mudah dan simpel, ditambah lagi, karena properti metode juga bebas, mereka tidak terhubung satu dengan yang lain. Sedangkan seleksi fitur *Information Gain* digunakan untuk mengurangi fitur dan properti yang tidak sesuai, maka dari itu keakuratannya akan meningkat. Data penelitian yang diimplementasikan berasal dari media sosial twitter dengan penggunaan teknik *scraping* dengan tagar *#merdekabelajar* dan kata kunci “merdeka belajar” dengan jumlah total *tweet* sebanyak 3833 *tweet*. Hasil studi menyatakan nilai akurasi sebesar 81,48%, presisi sebesar 89,74%, *recall* 76,09%, serta *f-measure* 82,35% (Zaman, 2020).

Pada tahun 2022, Abd. Charis Fauzan dan Khoiril Hikmah melakukan sebuah penelitian mengenai analisis polarisasi opini masyarakat terkait vaksin covid-19 dengan mengimplementasikan algoritma *Naive Bayes* (Fauzan & Hikmah, 2022). Studi ini bertujuan untuk mengidentifikasi polarisasi opini masyarakat terkait vaksinasi covid-19. Tahapan-tahapan yang dilakukan untuk menganalisis polarisasi opini masyarakat, pengumpulan data pada media sosial twitter menggunakan *tools* rapidMiner dengan jumlah data yang didapat sebesar 721 data adalah tahap pertama yang diimplementasikan, kemudian dilakukan tahapan *preprocessing*, dan terakhir pengkategorian metode *Naive Bayes*. Studi ini menemukan polarisasi opini publik, dengan 67% tweet mengekspresikan sentimen positif, 9% suasana sentimen netral, dan 32% sentimen negatif, dengan level akurasi dan daya ingat masing-masing 88% dan 97% (Fauzan & Hikmah, 2022).

Muhammad Rizal Rizwan tahun 2022 mengkaji analisis sentimen program kartu prakerja dengan mengimplementasikan teknik *Naive Bayes Classifier* (RIZWAN, 2022). Subjek penelitian berasal dari komentar pada akun instagram @prakerja.go.id sebanyak 2000 data. Data tersebut diperoleh menggunakan sebuah *extension* pada *browser* Google Chrome yang bernama IGCommentExport. Untuk mengetahui seberapa baik algoritma *Naive Bayes* melakukan analisis sentimen, penelitian ini akan menggunakannya untuk menginvestigasi sentimen terkait program kartu prakerja di kolom komentar Instagram. Hasil studi tersebut membuktikan bahwa metode *Naive Bayes* mempunyai keakuratan 84%, poin *precision* 86,59%, dan poin *recall* 100% (RIZWAN, 2022).

Perbandingan antara penelitian-penelitian sebelumnya yang dijadikan acuan untuk penelitian ini dipaparkan pada Tabel 2.1.

Tabel 2.1 Perbandingan Penelitian Terdahulu

No	Referensi	Metode yang Digunakan	Hasil
1.	(Yulita et al., 2021)	Naive Bayes Classifier	Menghasilkan opini positif sebesar 60,3%, opini netral sebesar 34,4%, dan opini negatif sebesar 5,4% dengan tingkat akurasi sebesar 93%.
2.	(Ainiyah, 2022)	Support Vector Machine	Hasil klasifikasi pada penelitian tersebut yaitu keakuratan 89,96%, 98,08% poin presisi, 68% poin <i>recall</i> , serta nilai <i>f-measure</i> sebesar 80,31%
3.	(Syarifuddin, 2020)	Naive Bayes dan KNN	Pendekatan <i>Naive Bayes</i> ditemukan mempunyai keakuratan yang lebih besar dibandingkan metode KNN. Metode <i>Naive Bayes</i> memperoleh poin keakuratan 63,21% dan KNN memperoleh keakuratan 58,10%.
4.	(Zaman, 2020)	Naive Bayes	Hasil studi membuktikan nilai akurasi sebesar 81,48%, 89,74% presisi, 76,09% <i>recall</i> , dan 82,35% <i>f-measure</i> .
5.	(Fauzan & Hikmah, 2022)	Naive Bayes	Dengan 67% tweet dengan emosi positif, 9% dengan sentimen netral, dan 32% dengan sentimen negatif, penelitian ini menghasilkan opini publik yang terpolarisasi dan mencapai tingkat akurasi dan daya ingat masing-masing sebesar 88% dan 97%.
6.	(RIZWAN, 2022)	Naive Bayes Classifier	Metode <i>Naive Bayes</i> mempunyai keakuratan 84%, 86,59% <i>precision</i> , dan 100% <i>recall</i> .

2.1 Landasan Teori

2.2.1 Media Sosial Instagram

Instagram adalah jejaring sosial berbasis gambar yang memungkinkan pengguna untuk mempublikasikan foto dan video tanpa dipungut biaya (Putri, 2018). Aplikasi instagram sendiri dapat diakses melalui telepon seluler baik yang berbasis Android, iOS, maupun Windows Phone. Aplikasi Instagram

memungkinkan penggunanya untuk mengunggah foto maupun video serta membagikan postingan tersebut kepada pengikutnya atau kepada pengguna Instagram yang lain. Selain mengunggah foto dan video, aplikasi ini dapat digunakan untuk melihat, menyukai, dan mengomentari postingan orang lain yang berupa video atau foto.

Beberapa fitur-fitur interaktif yang ditawarkan oleh Instagram diantaranya yaitu (Putri, 2018):

a. *Search & Explore*

Salah satu fitur Instagram, bagian *search & explore*, menawarkan rekomendasi konten berdasarkan postingan populer dan akun yang diikuti pengguna.. Selain itu terdapat fitur *search* yang memungkinkan pengguna untuk mencari akun, tagar, maupun lokasi yang diinginkan oleh pengguna.

b. *Direct Message*

Selain itu, *Direct Message* juga mempermudah pengguna untuk berkomunikasi dengan pengguna lain secara pribadi. Selain pesan *text*, fitur *Direct Message* juga dapat digunakan oleh pengguna untuk mengirimkan foto dan video. Selain itu, pada fitur *Direct Message* juga memungkinkan penggunanya untuk berinteraksi langsung melalui pesan suara dan *video call*.

c. *Reels*

Salah satu fitur terbaru yang dirilis oleh Instagram adalah *Reels*. *Reels* sendiri adalah fitur yang dapat dimanfaatkan untuk merekam dan membagikan video singkat dengan maksimal durasi 15 detik.

d. *Stories*

Stories adalah fitur untuk berbagi momen serta pengalaman dalam menggunakan instagram yang dapat berupa foto maupun video. *Stories* instagram yang diunggah oleh pengguna hanya dapat bertahan selama 24 jam, kemudian akan terhapus secara otomatis. Pengguna juga dapat menambahkan teks, musik, stiker serta GIF pada foto atau video yang dibagikan pada fitur ini.

e. *IG TV*

Fitur *IG TV* adalah fitur yang diimplementasikan untuk membagikan video vertikal yang berdurasi lebih lama. Pengguna dapat mengunggah video dengan durasi hingga 10 menit, sedangkan untuk pengguna terverifikasi dan jumlah pengikut yang banyak dapat mengunggah dengan durasi hingga 60 menit.

f. *Shop*

Instagram *Shop* adalah fitur baru yang memungkinkan pengguna untuk mencari suatu produk yang diinginkan. Selain itu, pengguna juga dapat melihat deskripsi produk dan detail harga dari produk tersebut.

2.2.2 *Kartu Prakerja*

Kartu prakerja adalah ketentuan yang pemerintah ciptakan khususnya oleh Kementrian Ketenagakerjaan yang bertujuan untuk mengembangkan kompetensi, meningkatkan produktivitas dan daya saing, serta mengembangkan kewirausahaan yang dikemas dalam program bantuan biaya pelatihan dan intensif. Sasaran penduduk program kartu prakerja ini adalah penduduk yang sedang mencari pekerjaan, pekerja, atau buruh yang terkena dampak kehilangan pekerjaan, dan para usaha mikro yang mengalami penurunan daya beli atau bahkan kehilangan

pekerjaan (Kemnaker, 2021). Pada awal peluncurannya, yaitu pada 11 April 2020, kartu prakerja ini dirancang sebagai program pelatihan untuk mengembangkan kompetensi serta meningkatkan produktivitas dan persaingan. Untuk menghindari penurunan kesejahteraan masyarakat, program ini kemudian ditambahkan pada program bantuan sosial (Firnanda et al., 2021).

Dari awal peluncuran, hingga pendaftaran gelombang ke-47, program kartu prakerja ini masih menuai beberapa polemik (Handayani & Rachman, 2020). Mulai dari sistem seleksi yang kurang baik, konten pelatihan yang dianggap terlalu mahal dengan kualitas yang tidak terlalu baik, dianggap sebagai pemborosan anggaran, hingga masalah *conflict of interest* yang dialami oleh beberapa staf khusus presiden (Consuello, 2020). Meskipun menuai berbagai polemik, animo masyarakat terhadap kartu prakerja masih sangat tinggi. Hal tersebut menunjukkan bahwa sejatinya program kartu prakerja ini sangat dibutuhkan oleh masyarakat. Dengan keterbatasan ruang dan waktu, masyarakat umum kesulitan untuk menyampaikan aspirasi secara langsung mengenai kebijakan yang dibuat oleh pemerintah tersebut (Syarifuddinn, 2020). Namun, dengan perkembangan internet yang semakin pesat, kini ada media sosial yang dapat dijadikan sebagai ruang publik baru untuk menyampaikan aspirasi masyarakat (Alvianda & Adikara, 2019). Salah satu jejaring sosial yang paling luas digunakan ialah instagram. Program kartu prakerja sendiri memiliki akun instagram resmi dengan nama akun @prakerja.go.id. Di dalam akun instagram tersebut, masyarakat dapat dengan bebas untuk memberikan komentar atau opininya mengenai kebijakan kartu prakerja.

2.2.3 Analisis Sentimen

Cara yang dilakukan untuk memahami aspirasi, pendapat, komentar, sikap, dan emosi seseorang terhadap sebuah masalah disebut analisis sentimen. Analisis ini disebut juga dengan *opinion mining*, yaitu salah satu bidang baru dalam *Natural Language Processing* (NLP) yang berisi proses pengolahan kata untuk menganalisis opini dari masyarakat (Deviyanto & Wahyudi, 2018).

Tugas mendasar dari analisis sistem ialah mengkategorikan polaritas dari sebuah teks yang berbentuk dokumen, kalimat, ataupun opini. Polaritas merupakan sebuah kecenderungan apakah teks yang berbentuk dokumen, kalimat, atau pendapat tersebut memiliki aspek positif ataupun negatif (Nugroho et al., 2015). Sehingga, dengan adanya analisis sentimen, pihak terkait yang mempunyai topik tersebut dapat mengetahui pendapat dari masyarakat terkait produk atau pelayanan tertentu. Proses analisis sentimen perlu dilakukan, agar perusahaan atau instansi terkait dapat menjadi bahan evaluasi untuk memperbaiki produk ataupun kinerjanya.

Langkah awal dalam analisis sentimen adalah pengumpulan data, atau pengumpulan data yang akan diteliti. Data dapat berasal dari berbagai sumber seperti media sosial, platform e-commerce, forum diskusi, atau survei. Tahap berikutnya setelah pengumpulan data adalah pra-pemrosesan data. Ini melibatkan pembersihan dan pengolahan teks, seperti menghapus karakter khusus, tanda baca, dan kata-kata yang tidak penting, serta melakukan normalisasi teks seperti *stemming* atau lemmatisasi. Setelah data siap, fitur-fitur yang relevan perlu dibentuk dari teks. Fitur-fitur ini dapat berupa kata-kata, frasa, atau kumpulan kata-

kata yang dapat digunakan untuk menggambarkan sentimen, seperti kata positif atau negatif (Phan et al., 2021).

Setelah itu, barulah masuk pada langkah klasifikasi sentimen. Pada langkah ini, teknik klasifikasi digunakan untuk mengklasifikasikan teks menjadi sentimen positif, negatif, atau netral. Teknik klasifikasi yang umum digunakan termasuk metode *Naive Bayes*, *Support Vector Machines (SVM)*, *Decision Trees*, atau algoritma pembelajaran mesin lainnya. Setelah klasifikasi dilakukan, langkah selanjutnya adalah mengevaluasi dan memvalidasi hasil analisis sentimen. Hal ini dapat dilakukan dengan membandingkan hasil klasifikasi dengan label sentimen yang ada jika tersedia, atau dengan menggunakan matrik evaluasi seperti akurasi, presisi, recall, atau F1-score. Langkah terakhir adalah menginterpretasikan hasil analisis sentimen dan menghasilkan visualisasi yang relevan. Visualisasi dapat berupa word clouds, grafik batang, atau grafik garis yang memberikan gambaran tentang sentimen yang dominan dalam data (Phan et al., 2021).

2.2.4 Text Preprocessing

Text Preprocessing merupakan salah satu tahapan dalam *text mining* yang berfungsi untuk mengubah data yang awalnya data tidak terstruktur menjadi data terstruktur yang siap diuji. Data mentah tidak dapat langsung diolah, oleh karena itu tahap *text preprocessing* dilakukan agar data mentah tersebut dapat menjadi data yang siap dianalisis (Deviyanto & Wahyudi, 2018). Proses-proses pada tahapan *text preprocessing* yang digunakan dalam penelitian ini adalah sebagai berikut:

a. *Cleansing*

Cleansing merupakan tahapan yang dilakukan untuk menghapus komponen-komponen yang tidak penting, karakter-karakter yang tidak baku, atau komponen yang tidak berkaitan dengan isi dokumen seperti *emoticon*, angka, tanda baca, simbol, spasi berlebih, *enter*, tautan, dan tagar.

b. *Case Folding*

Pada tahap *case folding*, seluruh huruf yang ada pada teks akan diubah menjadi *lowercase* atau huruf kecil.

c. *Tokenizing*

Tokenizing merupakan tahapan dalam *text preprocessing* yang dilakukan untuk memotong kalimat menjadi kata-kata yang menjadi penyusunnya. Tahap *tokenizing* dilakukan untuk memudahkan perhitungan frekuensi menghadirkan kata pada dokumen (Kusnia & Kurniawan, 2022).

d. *Normalization*

Pada tahap *normalization*, seluruh kata yang memiliki kesalahan dalam ejaan, *noise*, ataupun kata yang tidak baku diganti menjadi baku. Kata yang mengandung *noise* sendiri merupakan kata yang mempunyai unsur-unsur bahasa daerah dan tidak sesuai dengan Kamus Besar Bahasa Indonesia (KBBI) (Septian et al., 2019).

e. *Stopword Removal*

Stopword removal atau dapat disebut juga dengan tahap *filtering* merupakan tahapan yang digunakan untuk menghilangkan kata-kata umum yang sering muncul

dan tidak memiliki makna tertentu (Styawati. et al., 2021). *Stopword removal* mempunyai tujuan menghilangkan dalam daftar token.

f. *Stemming*

Stemming merupakan tahap akhir dalam proses *text preprocessing* dimana seluruh kata dalam dokumen akan diganti menjadi kata dasar. Sehingga, semua kata imbuhan mulai dari awalan (prefix), akhiran (sufix), sisipan (infix), serta gabungan antara awalan dan akhiran (confix) akan dihilangkan.

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

Weighting word atau dalam bahasa Indonesia dikatakan pembobotan kata adalah sebuah cara yang diimplementasi untuk memberikan skor pada kemunculan suatu kata di dalam sebuah dokumen (Yulita et al., 2021). Proses pembobotan kata dilakukan untuk mengetahui kemiripan suatu kata dengan kata lain. Selain itu, proses pembobotan kata dilakukan agar data yang digunakan dapat memenuhi syarat untuk diproses dan dihitung lebih lanjut untuk membuat model.

TF-IDF ialah teknik pembobotan kata yang pengimplementasiannya sangat luas. Teknik ini ialah teknik statistik yang menilai signifikansi kata dalam dokumen tertentu dibandingkan dengan dokumen lain. TF-IDF adalah metode pengembangan dari metode-metode sebelumnya dan merupakan gabungan dari dua metode yang berbeda yaitu metode TF dan IDF. Sehingga, TF-IDF memiliki nilai perhitungan yang didapatkan dari hasil perkalian 2 nilai, yaitu nilai TF dan nilai IDF (Ulinuha, 2022). Karena TF-IDF merupakan metode gabungan dari TF dan IDF, maka metode ini memiliki dua konsep pembobotan kata, yaitu frekuensi

kehadiran sebuah kata dalam dokumen dan *inverse* frekuensi dokumen yang berisi kata tersebut.

TF (*Term Frequency*), yang merupakan singkatan dari Frekuensi Istilah, digunakan sebagai metode untuk mengatribusikan bobot pada kata-kata dalam makalah dengan mengidentifikasi kata kunci yang cocok atau memiliki kesamaan dengan kategori yang telah ditentukan sebelumnya. Faktor yang menentukan bobot sebuah istilah dalam sebuah dokumen adalah sejauh mana kata tersebut sering muncul dalam dokumen tersebut. Oleh karena itu, ketika suatu kata muncul dengan frekuensi yang semakin tinggi dalam sebuah dokumen, maka bobot kata tersebut akan meningkat secara proporsional. *Term Frequency* (TF) sendiri memiliki kekurangan yaitu ketika suatu kata muncul pada semua dokumen, maka kata tersebut akan dianggap sebagai *common term* dan nilainya sudah dianggap tidak penting. Maka dari itu, diperlukan metode *Inverse Document Frequency* (IDF) yang bertujuan untuk menghilangkan dominasi *common term*. IDF adalah teknik pembobotan kata yang diterapkan untuk mengurangi dominasi *common term* atau *term* yang banyak muncul pada dokumen (Noviyanti, 2018).

Proses penghitungan bobot kata menggunakan metode TF-IDF melibatkan perhitungan nilai TF (*Term Frequency*) atau frekuensi kemunculan kata dalam suatu dokumen, diikuti oleh perhitungan DF (*Document Frequency*) atau jumlah dokumen yang mengandung kata-kata yang sedang dihitung. Setelah itu dilakukan perhitungan nilai IDF, dan terakhir adalah perhitungan bobot kata TF-IDF. Pembobotan kata menggunakan TF-IDF sendiri dapat dihitung menggunakan persamaan sebagai berikut:

$$TF = \frac{\text{jumlah kemunculan kata dalam dokumen}}{\text{jumlah kata pada dokumen}} \quad (2.1)$$

$$IDF = \log \frac{|D|}{DF} \quad (2.2)$$

$$TF - IDF = TF \times IDF \quad (2.3)$$

Dimana,

TF-IDF : Bobot kata dalam sebuah dokumen
 TF : Frekuensi kemunculan kata dalam sebuah dokumen
 IDF : Frekuensi data kata
 D : Jumlah total dokumen
 DF : Jumlah dokumen yang mengandung kata

2.2.6 K-Fold Cross Validation

Cross Validation adalah sebuah teknik validasi model untuk menilai kinerja model pada dataset (Veronica et al., 2020). Teknik ini biasanya digunakan untuk memprediksi akurasi suatu model prediktif yang diimplementasikan praktiknya dengan cara memisahkan data menjadi data latih dan data uji (Ansori et al., 2022). *K-fold Cross Validation* atau teknik validasi silang adalah suatu teknik yang dilakukan dengan membedakan data menjadi k kategori dataset dengan ukuran serupa. *K-fold Cross Validation* sendiri digunakan untuk menghapuskan bias pada data (Tempola et al., 2018). Metode ini melibatkan pengelompokan data training dan test, diikuti dengan k iterasi testing. Sehingga setiap dataset yang ada memiliki kesempatan untuk menjadi data uji dan latih. Peneliti mengimplementasikan k sebanyak 5, 10, 15, dan 20-fold, yang berarti pengujian akan dilakukan sebanyak 5, 10, 15, dan 20 kali pada seluruh isi dokumen secara acak.

2.2.7 Naive Bayes Classifier

Metode *Naive Bayes* merupakan sebuah metode pengklasifikasian probabilistik sederhana dengan cara kerja yaitu menghitung sekumpulan probabilitas yang didapat dengan menjumlahkan frekuensi dari kombinasi nilai dan dataset yang digunakan (Rezki et al., 2020). Metode *Naive Bayes* memperkirakan peluang pada zaman yang akan datang sesuai dengan pengetahuan terdahulu atau biasa disebut dengan *conditional probabilities* (Fauzan & Hikmah, 2022). *Conditional probabilities* merupakan sebuah perhitungan probabilitas peristiwa H, ketika peristiwa lain yaitu peristiwa X telah terjadi, yang ditandai dengan $P(H | X)$ yang menggabungkan probabilitas dari H dan X (Suhartono et al., 2018). Metode *Naive Bayes* sendiri menggunakan konsep Teorema Bayes yang dikemukakan oleh seorang ilmuwan Inggris yang bernama Thomas Bayes (Supriadi & Fatmasari, 2021). Teorema Bayes sendiri dapat dilihat pada persamaan 2.4 berikut ini:

$$P(H | X) = \frac{P(X|H) P(H)}{P(X)} \quad (2.4)$$

Dimana,

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

$P(H | X)$ = Probabilitas hipotesis H Berdasarkan kondisi X

$P(H)$ = Probabilitas hipotesis H (prior probabilitas)

$P(X | H)$ = Probabilitas X berdasarkan kondisi tersebut

$P(X)$ = Probabilitas dari X

Metode *Naive Bayes* sendiri dapat dikategorikan menjadi tiga tipe, yaitu *Multinomial*, *Bernoulli*, dan *Gaussian*. *Multinomial Naive Bayes* merupakan metode *Naive Bayes* untuk pengklasifikasian teks atau dokumen. *Bernoulli Naive Bayes* hampir mirip dengan multinomial, bedanya *Bernoulli* menggunakan input prediktor dengan variabel *boolean*, sehingga hasil klasifikasinya lebih berfokus

pada hasil ‘ya’ atau ‘tidak’. Sedangkan *Gaussian* lebih terfokus untuk data dalam nilai numerik. Sehingga penelitian ini akan lebih berfokus pada *Multinomial Naive Bayes*. Persamaan 2.5 di bawah ini digunakan untuk menemukan keluaran probabilitas tertinggi di semua kategori VMap.:

$$\mathbf{VMap} = \mathbf{arg\ max} P(\mathbf{Vj} | \mathbf{a1}, \mathbf{a2}, \dots, \mathbf{an}) \quad (2.5)$$

Kemudian, dengan menerapkan persamaan Teorema Bayes, maka persamaannya menjadi seperti pada persamaan 2.6 berikut ini:

$$\mathbf{VMap} = \mathbf{arg\ max} (\mathbf{Vj\ eV}) \frac{P(\mathbf{a1}, \mathbf{a2}, \dots, \mathbf{an} | P(\mathbf{Vj}))}{P(\mathbf{a1}, \mathbf{a2}, \dots, \mathbf{an})} \quad (2.6)$$

Karena nilai $P(a_1, a_2, a_3)$ konstan untuk semua kategori (V_j), maka persamaan 2.6 tersebut akan menjadi persamaan 2.7 berikut ini:

$$\mathbf{VMap} = \mathbf{arg\ max} (\mathbf{Vj\ eV}) P(\mathbf{a1}, \mathbf{a2}, \dots, \mathbf{an} | P(\mathbf{Vj})) \quad (2.7)$$

Dan dapat disederhanakan menjadi persamaan 2.8 berikut:

$$\mathbf{VMap} = \mathbf{arg\ max} (\mathbf{Vj\ eV}) \pi^{n_{i=1}} P(\mathbf{Wk} / \mathbf{Vj}) P(\mathbf{Vj}) \quad (2.8)$$

Dimana,

V_j = Kategori j

$P(V_j)$ = Probabilitas V_j

$P(W_k | V_j)$ = Probabilitas W_k dalam kategori V_j

W_k = Bobot kata

Untuk menghitung $P(V_j)$, dapat menggunakan persamaan 2.9 berikut:

$$P(\mathbf{Vj}) = \frac{|\mathbf{doc\ j}|}{|\mathbf{contoh}|} \quad (2.9)$$

Dimana,

$|\mathbf{doc\ j}|$ = Jumlah dokumen dari setiap kategori

$|\mathbf{contoh}|$ = Jumlah dokumen semua kategori

Dan untuk menghitung $P(W_k | V_j)$ dapat menggunakan persamaan 2.10 berikut:

$$P(W_k | V_j) = \frac{nk+1}{n+|kosakata|} \quad (2.10)$$

Dimana,

nk = Jumlah frekuensi setiap kata

n = Jumlah kata dari dokumen berkategori V_j

Pada persamaan 2.10 tersebut, setiap perhitungan data akan ditambah 1 (satu). Penambahan angka 1 tersebut terjadi karena basis perhitungan metode *Naive Bayes* bertumpu pada probabilitas, sehingga sangat memungkinkan suatu atribut akan memiliki nilai probabilitas 0 (nol). Hal tersebut dapat menyebabkan seluruh perhitungan menjadi rusak atau tidak sesuai. Oleh karena itu, dibutuhkan teknik *Laplacian Correction* dalam mengatasi persoalan tersebut. *Laplacian Correction* adalah teknik yang diimplementasikan untuk menangani nilai probabilitas pada *Naive Bayes* agar hasilnya tidak 0 (nol) (Rizki et al., 2021).

Metode *Naive Bayes* mempunyai karakteristik yang cepat dan keakuratan yang besar ketika diaplikasikan dalam dataset yang besar dan beragam (Fauzan & Hikmah, 2022). Karena dataset yang digunakan di dalam penelitian ini cukup besar, maka dibutuhkan sebuah metode dengan performansi yang cepat dengan akurasi tinggi. Karena di dalam penelitian ini digunakan pembobotan menggunakan TF-IDF, maka metode *Naive Bayes* dipilih karena akurasinya akan meningkat jika dikolaborasikan dengan pembobotan tersebut (Syarifuddin, 2020)

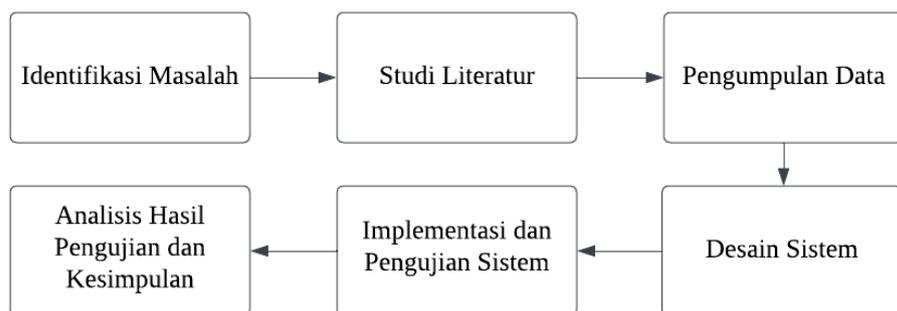
BAB III

METODOLOGI PENELITIAN

Langkah-langkah yang akan diambil selama pelaksanaan penelitian akan dijelaskan dalam bab ini. Untuk mencapai tujuan penelitian, langkah-langkah ini akan digunakan sebagai acuan dan pedoman selama pelaksanaan penelitian. Bab ini mencakup teknik penelitian, pengumpulan data, pelabelan data, arsitektur sistem, pembobotan kata, preprocessing teks, perhitungan analisis sentimen secara manual menggunakan pendekatan *Naive Bayes*, dan skenario studi.

3.1 *Prosedur Penelitian*

Rancangan aktivitas dalam suatu studi yang teratur dan sistematis untuk memenuhi target penelitian disebut sebagai prosedur penelitian. Pada prosedur penelitian ini akan dijelaskan langkah-langkah rinci dalam penelitian berawal dari pengumpulan data, pemrosesan data, hingga didapati temuan akhir. Gambar 3.1 di bawah ini menjelaskan proses yang akan diimplementasikan dalam investigasi ini.



Gambar 3.1 Prosedur Penelitian

Prosedur studi yang diterapkan pada penelitian ini diawali dengan identifikasi *problem* atau masalah. Pada tahap identifikasi masalah, dirumuskan

latar belakang masalah sebagai alasan mengapa penelitian ini dilakukan. Selain itu, pada tahap identifikasi masalah juga dirumuskan rumusan masalah dan batasan masalah. Tahap selanjutnya adalah studi pustaka, dimana dilakukan pencarian referensi pada jurnal ilmiah, tesis, dan disertasi yang sesuai dengan judul penelitian. Tahap pengumpulan data merupakan tahap pengumpulan data primer untuk penelitian. Data primer akan diolah terlebih dahulu untuk memperoleh temuan studi. Langkah berikutnya adalah perancangan sistem, dimana persyaratan untuk sistem dan tahapannya diuraikan secara menyeluruh. Tahap implementasi dan pengujian sistem dilakukan dengan mengimplementasikan desain sistem yang telah dibuat. Kemudian akan dilakukan pengujian sistem yang telah dibuat tersebut. Tahap terakhir adalah analisis hasil pengujian dan kesimpulan yang dilakukan dengan melakukan perhitungan nilai akurasi yang dihasilkan oleh sistem yang telah dibuat.

3.2 Pengumpulan Data

Menemukan informasi adalah suatu tahap yang di kenal sebagai pengumpulan data. Data yang diimplementasikan pada studi ini ialah komentar pada unggahan akun @prakerja.go.id dalam rentang waktu 26 Agustus 2022 hingga 9 Desember 2022. Pengumpulan data komentar instagram dilakukan menggunakan salah satu ekstensi pada *browser* Google Chrome yaitu *Data Miner*. Data komentar instagram pada akun @prakerja.go.id yang akan diimplementasikan pada studi ini sebanyak 2200 komentar.

3.3 Pelabelan Data

Pelabelan data adalah tahap berikutnya. Tahap ini diimplementasikan dengan maksud untuk melabeli data komentar menjadi dua kategori, yaitu positif dan negatif. Pelabelan data dilakukan secara manual oleh peneliti kemudian hasil pelabelan data tersebut akan diperiksa oleh validator data. Validasi data tersebut akan diterapkan oleh ahli bahasa yang saat ini berprofesi sebagai guru Bahasa Indonesia SMA Muhammadiyah 3 Tulangan, Sidoarjo atas nama Ibu Wahyu Putri Romadhoni. Seluruh data komentar instagram yang sudah dikumpulkan, penulis mencantumkan sepuluh data sampel dilabeli secara manual. Proses pelabelan data secara manual dilakukan dengan berdasarkan kriteria pada Tabel 3.1, sedangkan, Tabel 3.2 memaparkan hasil data komentar yang telah dilabeli.

Tabel 3.1 Kriteria Pelabelan Data

Kalimat Positif	Kalimat Negatif
Kalimat yang mengandung optimisme, unsur doa, serta mengandung kalimat motivasi, dan ajakan menuju hal yang lebih baik.	Kalimat yang mengandung prasangka buruk, mengolok, menghasut, kalimat pesimis, serta ketidakpercayaan.

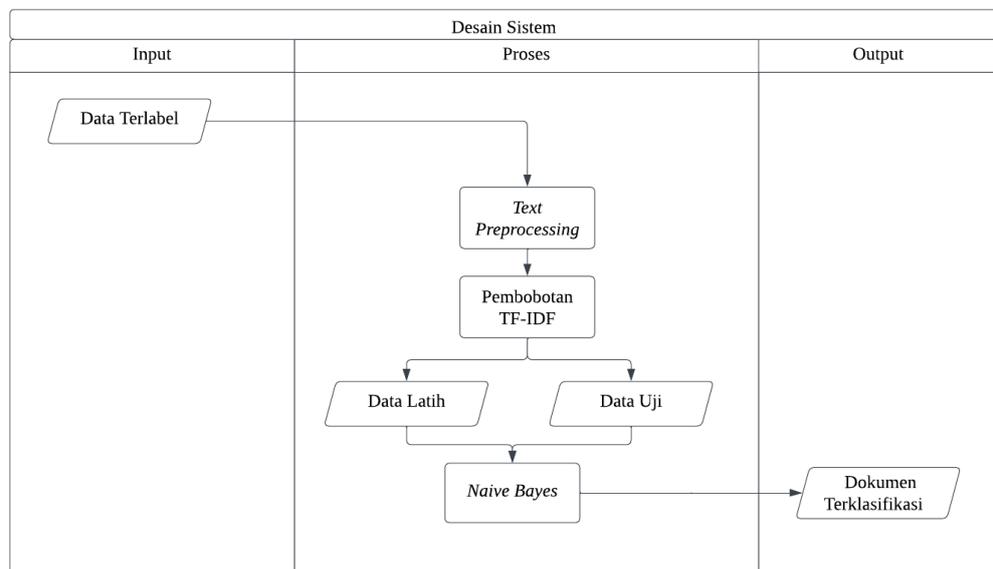
Tabel 3.2 Data Terlabel

Komentar	Pelabelan Manual
daftar mentok diveribikasi wajah ga bisa bisa	Negatif
status menerima bantuan pemerintah terus gagal, padahal mereka tidak menerima apapun 😞	Negatif
Alhamdulillah lolos,, tinggal nunggu ulasan nya keluar,, ❤️❤️	Positif
4x daftar gelombang dari tahun 2020 tapi ehh 😞 aneh gagal mulu, yg bisa ngakalin malah pda lolos	Negatif
Banyak yang lolos ya bulan skarangg	Positif
Selamat buat teman-teman yang lolos 😊	Positif
Link nya eror min	Negatif
Semangaaatttt ikut lg semoga lolos	Positif
Selamat Buat yang Lolos Kartu Prakerja Gelombang 47 Mimin Doakan yang belum lolos bisa Lolos di Gelombang 48 ya Amin 🙏	Positif
Admin kenapa lama banget utk mengupgrade link aja, mohon bantuan dan solusi @prakerja.go.id	Negatif

Dari data yang telah dilabeli secara manual dengan memenuhi kriteria pada Tabel 3.1, selanjutnya data tersebut akan dilakukan pemrosesan selanjutnya yaitu *Text Preprocessing*, pembobotan TF-IDF, pemodelan *Naive Bayes*, pengujian menggunakan *K-fold Cross Validation*, dan pengujian kinerja model menggunakan *Confusion Matrix*.

3.4 Desain Sistem

Alur sistem yang menjadi sebagai acuan bagi peneliti sebelum sistem tersebut dirancang dikenal sebagai desain sistem. Di dalam desain sistem ini mencakup tiga komponen utama yakni input, proses, dan output. Gambar 3.2 di bawah ini memaparkan desain sistem.

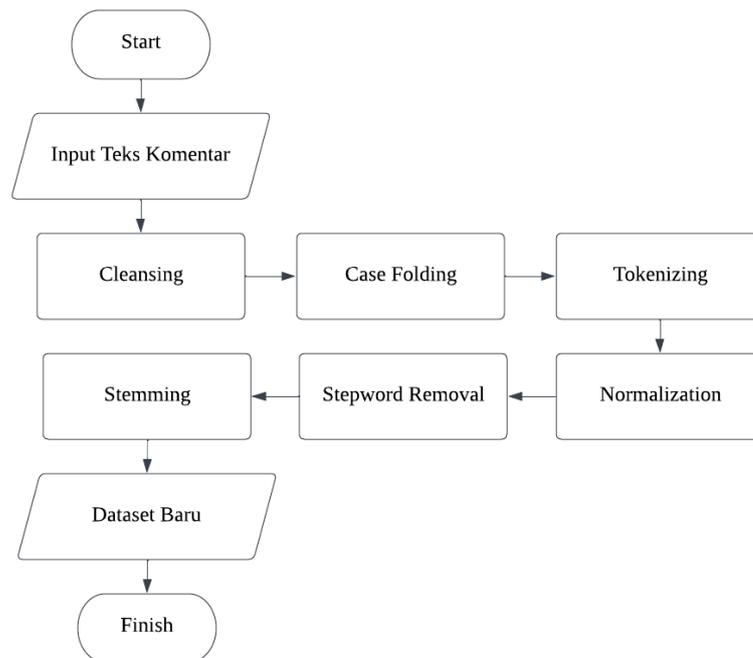


Gambar 3.2 Desain Sistem

3.5 Text Preprocessing

Text Preprocessing adalah salah satu tahapan dalam *text mining* yang berfungsi untuk mengganti data yang awalnya data tidak terstruktur menjadi data

terstruktur yang siap diuji. Data mentah tidak dapat langsung diolah, oleh karena itu tahap *text preprocessing* dilakukan agar data mentah tersebut dapat menjadi data yang siap dianalisis (Akhmad Deviyanto). Selain itu, *text preprocessing* mempengaruhi temuan pengkategorian serta nilai performa proses klasifikasi tersebut. Data yang telah dilakukan proses *preprocessing* akan menghasilkan klasifikasi data yang lebih optimal daripada data yang tidak di paparkan proses *preprocessing*. Studi ini mengimplementasikan 6 tahap *preprocessing* yaitu *cleansing*, *case folding*, *tokenizing*, *normalization*, *stopword removal*, dan *stemming*.



Gambar 3.3 Diagram Alir *Text Preprocessing*

3.5.1 *Cleansing*

Cleansing adalah langkah menghapus komponen-komponen yang tidak penting, karakter-karakter yang tidak baku, atau komponen-komponen yang tidak

ada kaitannya dengan data dokumen seperti *emoticon*, angka, tanda baca, simbol, spasi berlebih, *enter*, tautan, dan tagar. Sehingga hasil dari tahap *cleansing* berupa data bersih yang hanya berisi kata-kata yang penting saja.

3.5.2 Case Folding

Pada tahap *case folding*, seluruh huruf yang ada pada teks akan diubah menjadi *lowercase* atau huruf kecil. Sehingga semua huruf-huruf pada teks akan sama bentuknya. Proses *case folding* bertujuan untuk mempermudah proses penelusuran data. Contohnya, penelusuran data yang berisi kata “lolos” namun hasil pencarian menunjukkan tidak ada, karena kata “lolos” di indeks sebagai “LOLOS”.

3.5.3 Tokenizing

Tokenizing merupakan tahapan dalam *text preprocessing* yang dilakukan untuk memotong kalimat menjadi kata-kata yang menjadi penyusunnya. Data-data yang telah terpotong tersebut disebut dengan token. Proses *tekonizing* berfungsi untuk memudahkan tahap investigasi. Lebih lanjut lagi peneliti mengimplementasikan pembobotan kata TF-IDF, maka *tokenizing* sangat diperlukan karena TF-IDF dalam prosesnya menggunakan token.

3.5.4 Normalization

Pada tahap *normalization*, seluruh kata yang memiliki kesalahan dalam ejaan, *noise*, ataupun bahasa yang tidak baku diganti menjadi bahasa baku. Kata yang mengandung *noise* sendiri merupakan kata yang mengandung unsur-unsur bahasa daerah dan tidak sesuai dengan Kamus Besar Bahasa Indonesia (KBBI)(Septian et al., 2019).

3.5.5 Stopword Removal

Stopword removal atau dapat disebut juga dengan tahap *filtering* merupakan tahapan yang digunakan untuk menghilangkan kata-kata umum yang sering muncul dan tidak memiliki makna tertentu (Styawati. et al., 2021). Tujuan dari proses *stopword removal* adalah untuk mengurangi jumlah kata yang disimpan dalam daftar token.

3.5.6 Stemming

Stemming merupakan tahap akhir dalam proses *text preprocessing* dimana seluruh kata dalam dokumen akan diubah menjadi kata dasar. Sehingga, semua kata imbuhan mulai dari awalan (prefix), akhiran (sufix), sisipan (infix), serta gabungan antara awalan dan akhiran (confix) akan dihilangkan. Proses *stemming* dibutuhkan dalam penyesuaian variasi kata supaya tidak banyak variasi kata.

Tabel 3.3 berikut ini merupakan hasil proses *preprocessing* yang dilakukan.

Tabel 3.3 *Preprocessing* Data

<i>Preprocessing</i>	Sebelum	Sesudah
<i>Cleansing</i>	Nikmat Tuhan mana yg kau dustakan 🙌😬kejutan akhir tahun dari prakerja 🙌😊 terimakasih prakerja...#siapdarisekarang dan sampai jumpa di gelombang berikutnya tahun 2023	Nikmat Tuhan mana yg kau dustakan kejutan akhir tahun dari prakerja terimakasih prakerja siapdarisekarang dan sampai jumpa di gelombang berikutnya tahun
<i>Case Folding</i>	Nikmat Tuhan mana yg kau dustakan kejutan akhir tahun dari prakerja terimakasih prakerja siapdarisekarang dan sampai jumpa di gelombang berikutnya tahun	nikmat tuhan mana yg kau dustakan kejutan akhir tahun dari prakerja siapdarisekarang dan sampai jumpa di gelombang berikutnya tahun
<i>Tokenizing</i>	nikmat tuhan mana yg kau dustakan kejutan akhir tahun dari prakerja siapdarisekarang dan sampai jumpa di gelombang berikutnya tahun	“nikmat”, “tuhan”, “mana”, “yg”, “kau”, “dustakan”, “kejutan”, “akhir”, “tahun”, “dari”, “prakerja”

<i>Preprocessing</i>	Sebelum	Sesudah
		“siapdarisekarang”, “dan”, “sampai”, “jumpa”, “di”, “gelombang”, “berikutnya”, “tahun”
<i>Normalization</i>	“nikmat”, “tuhan”, “mana”, “yg”, “kau”, “dustakan”, “kejutan”, “akhir”, “tahun”, “dari”, “prakerja”, “siapdarisekarang”, “dan”, “sampai”, “jumpa”, “di”, “gelombang”, “berikutnya”, “tahun”	“nikmat”, “tuhan”, “mana”, “yang”, “kamu”, “dustakan”, “kejutan”, “akhir”, “tahun”, “dari”, “prakerja”, “siap”, “dari”, “sekarang”, “dan”, “sampai”, “jumpa”, “di”, “gelombang”, “berikutnya”, “tahun”
<i>Stopword Removal</i>	“nikmat”, “tuhan”, “mana”, “yang”, “kamu”, “dustakan”, “kejutan”, “akhir”, “tahun”, “dari”, “prakerja”, “siap”, “dari”, “sekarang”, “dan”, “sampai”, “jumpa”, “di”, “gelombang”, “berikutnya”, “tahun”	“nikmat”, “tuhan”, “dustakan”, “kejutan”, “akhir”, “tahun”, “prakerja”, “gelombang”, “berikutnya”, “tahun”
<i>Stemming</i>	“nikmat”, “tuhan”, “dustakan”, “kejutan”, “akhir”, “tahun”, “prakerja”, “gelombang”, “berikutnya”, “tahun”	“nikmat”, “tuhan”, “dusta”, “kejut”, “akhir”, “tahun”, “prakerja”, “gelombang”, “ikut”, “tahun”

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

Setelah dilakukan proses *text preprocessing*, maka akan dihasilkan kumpulan kata yang akan dijadikan *term* atau aturan untuk proses selanjutnya yaitu pembobotan kata. Pembobotan kata pada studi ini mengimplementasikan TF-IDF atau *Term Frequency-Inverse Document Frequency*.

Untuk menghitung bobot kata menggunakan TF-IDF, dapat dilakukan menggunakan persamaan 2.1 hingga 2.3 pada studi ini, perhitungan manual akan menerapkan 6 data sebagai sampel. Data-data tersebut disajikan pada Tabel 3.4.

Tabel 3.4 Data Sampel

Dokumen	Komentar	Pelabelan Manual
D1	gagal terus statusnya nerima bantuan pemerintah padahal ga nerima sama sekali 😞	Negatif
D2	Alhamdulillah lolos,, tinggal nunggu ulasan nya keluar,, ❤️❤️	Positif
D3	4x daftar gelombang dari tahun 2020 tapi ehh 😞 aneh gagal mulu, yg bisa ngakalin malah pda lolos	Negatif
D4	Banyak yang lolos ya bulan skarangg	Positif
D5	Selamat buat teman-teman yang lolos 😊	Positif

Dokumen	Komentar	Pelabelan Manual
D6	Link nya eror min	Negatif

Tahap pertama yang harus dilakukan dalam perhitungan TF-IDF adalah menghitung nilai TF terlebih dahulu. Nilai TF didapatkan dengan membagi jumlah kata “x” yang muncul dalam dokumen “d” dengan total kata dalam dokumen “d”. Contoh perhitungan nilai TF pada suatu kata dipaparkan pada Tabel 3.5.

Tabel 3.5 Perhitungan TF

Dokumen	Jumlah Kata	Kata	Frekuensi	TF
D1	7	gagal	1	1/7
D2	4		0	0/4
D3	8		1	1/8
D4	4		0	0/4
D5	4		0	0/4
D6	2		0	0/2

Tabel 3.5 merupakan tabel yang berisi perhitungan nilai *Term Frequency* (TF). Nilai *Term Frequency* (TF) didapatkan dengan membagi jumlah frekuensi kata terpilih dengan jumlah kata pada suatu dokumen. Sebagai contoh, pada D1 kata terpilih dengan jumlah kata pada suatu dokumen. Sebagai contoh, pada D1 terdapat sebanyak 7 kata di dalamnya dan kata “gagal” muncul sebanyak 1 kali pada D1, sehingga nilai TF untuk kata “gagal” pada D1 adalah 1/7. Nilai TF-IDF kemudian ditentukan dengan menggunakan persamaan 2.2 dan 2.3 berdasarkan nilai TF yang dihitung pada Tabel 3.5, berikut contoh perhitungan TF-IDF untuk kata atau *term* “gagal” pada D1.

$$\begin{aligned}
 IDF &= \log \frac{|D|}{DF} \\
 &= \log \frac{6}{2} \\
 &= 0,477121
 \end{aligned}$$

$$\begin{aligned}
 TF - IDF &= TF \times IDF \\
 &= 0,142857 \times 0,4771212 \\
 &= 0,068160179
 \end{aligned}$$

Perhitungan nilai TF-IDF untuk seluruh kata pada setiap dokumen sampel terdapat pada Tabel 3.6.

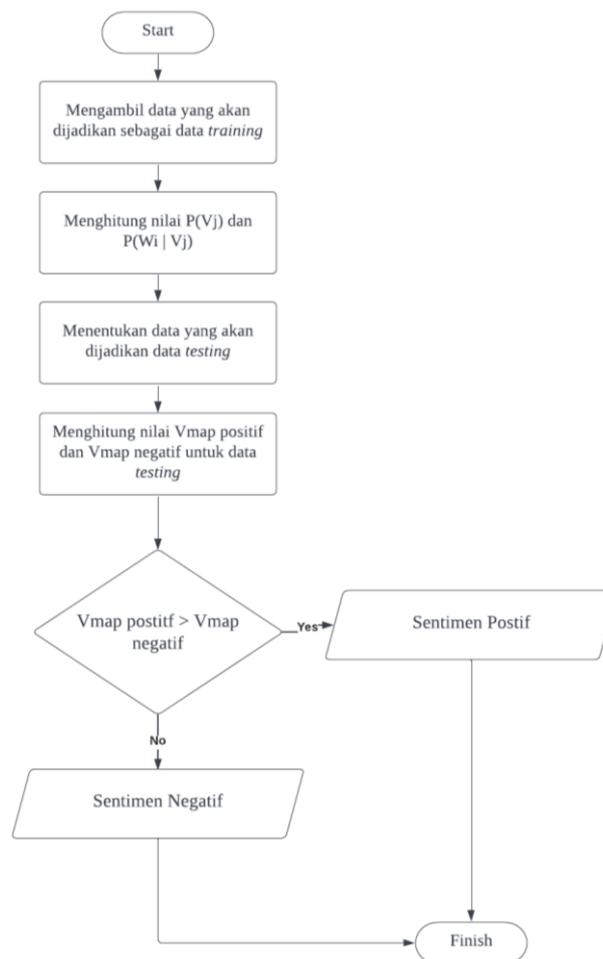
Tabel 3. 6 Hasil Perhitungan TF-IDF

Kata	Jumlah Kata					
	D1	D2	D3	D4	D5	D6
gagal	0,068160179		0,05964			
status	0,111164464					
terima	0,222328929					
bantu	0,111164464					
perintah	0,111164464					
tidak	0,111164464					
alhamdulillah		0,194538				
lolos		0,044023	0,022011	0,044023	0,044023	
tunggu		0,194538				
ulas		0,194538				
daftar			0,097269			
gelombang			0,097269			
tahun			0,097269			
tapi			0,097269			
bisa			0,097269			
akal			0,097269			
banyak				0,194538		
bulan				0,194538		
sekarang				0,194538		
selamat					0,194538	
teman					0,389076	
link						0,389076
eror						0,389076

3.7 Naive Bayes

Setelah dilakukan pembobotan kata menggunakan metode TF-IDF, langkah berikutnya menerapkan perhitungan menggunakan metode klasifikasi *Naive Bayes*. Metode *Naive Bayes* merupakan sebuah metode pengklasifikasian probabilistik

sederhana dengan cara kerja yaitu menghitung sekumpulan probabilitas yang didapat dengan menjumlahkan frekuensi dari kombinasi nilai dan dataset yang digunakan (Rezki et al, 2020). Metode *Naive Bayes* bekerja dengan memprediksi peluang di masa depan berdasarkan pengalaman di masa lalu (Fauzan & Hikmah, 2022). Tahap klasifikasi menggunakan metode *Naive Bayes* dibagi menjadi dua proses, yaitu proses *training* dan proses *testing*. Proses *training* merupakan proses pelatihan, sedangkan proses *testing* merupakan proses pengujian dengan mengacu pada probabilitas dari *dataset training*. Diagram alir untuk tahapan klasifikasi *Naive Bayes* dapat dilihat pada Gambar 3.4 berikut ini.



Gambar 3.4 Diagram Alir Klasifikasi *Naive Bayes*

3.7.1 Training

Proses *training* dilakukan dengan dokumen *training* yang telah diklasifikasi secara manual berdasarkan kelasnya dan pengimplementasian tahap *preprocessing* seperti pada Tabel 3.7.

Tabel 3.7 Contoh Kasus Data *Training*

Dokumen	Fitur	Kelas
D1	gagal status terima bantu perintah tidak terima	N
D2	daftar gelombang tahun tapi gagal bisa akal lolos	N
D3	link eror	N
D4	alhamdulillah lolos tunggu ulas	P
D5	banyak lolos bulan sekarang	P
D6	selamat teman teman lolos	P

Dari dokumen *training* tersebut, dibuat sebuah model *prior probabilities* yang mengacu pada persamaan 2.9 sebagai berikut:

$$P(V_j / positif) = \frac{|doc j|}{|contoh|} = \frac{3}{6} = \frac{1}{2}$$

$$P(V_j / negatif) = \frac{|doc j|}{|contoh|} = \frac{3}{6} = \frac{1}{2}$$

Kelas positif muncul sebanyak 3 kali dalam dokumen *training*, dan total keseluruhan dokumen adalah sebanyak 6 dokumen. Sehingga nilai $P(V_j)$ untuk kelas positif adalah $\frac{1}{2}$. Kelas negatif juga muncul sebanyak 3 kali dalam dokumen *training*, sehingga nilai $P(V_j)$ untuk kelas positif adalah $\frac{1}{2}$. Setelah itu, dibuat sebuah model *conditional probabilities* yang mengacu pada persamaan 2.10 sebagai contoh, dilakukan perhitungan *conditional probabilities* untuk kata “gagal” pada kelas positif dan negatif.

$$P(W_{gagal} / positif) = \frac{nk+1}{n+|kosakata|} = \frac{0+1}{12+23} = \frac{1}{35}$$

$$P(W_{gagal} / negatif) = \frac{nk+1}{n+|kosakata|} = \frac{2+1}{17+23} = \frac{3}{40}$$

Kemudian dilakukan perhitungan probabilitas pada setiap kata seperti pada

Tabel 3.8 hingga 3.13 berikut ini.

Tabel 3.8 Perhitungan Probabilitas Kata pada D1

Kelas	P (V _j)	P(W _i V _j)					
		gagal	status	terima	bantu	perintah	tidak
P	$\frac{1}{2}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$
N	$\frac{1}{2}$	$\frac{3}{40}$	$\frac{2}{40}$	$\frac{3}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$

Tabel 3.9 Perhitungan Probabilitas Kata pada D2

Kelas	P (V _j)	P(W _i V _j)						
		daftar	gelombang	tahun	tapi	bisa	akal	lolos
P	$\frac{1}{2}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{4}{35}$
N	$\frac{1}{2}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$	$\frac{2}{40}$

Tabel 3.10 Perhitungan Probabilitas Kata pada D3

Kelas	P (V _j)	P(W _i V _j)	
		link	eror
P	$\frac{1}{2}$	$\frac{1}{35}$	$\frac{1}{35}$
N	$\frac{1}{2}$	$\frac{2}{40}$	$\frac{2}{40}$

Tabel 3.11 Perhitungan Probabilitas Kata pada D4

Kelas	P (V _j)	P(W _i V _j)		
		alhamdulillah	tunggu	ulas
P	$\frac{1}{2}$	$\frac{2}{35}$	$\frac{2}{35}$	$\frac{2}{35}$
N	$\frac{1}{2}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$

Tabel 3.12 Perhitungan Probabilitas Kata pada D5

Kelas	P (V _j)	P(W _i V _j)		
		banyak	bulan	sekarang
P	$\frac{1}{2}$	$\frac{2}{35}$	$\frac{2}{35}$	$\frac{2}{35}$
N	$\frac{1}{2}$	$\frac{1}{40}$	$\frac{1}{40}$	$\frac{1}{40}$

Tabel 3.13 Perhitungan Probabilitas Kata pada D6

Kelas	P (V _j)	P(W _i V _j)	
		selamat	teman
P	$\frac{1}{2}$	$\frac{2}{35}$	$\frac{3}{35}$
N	$\frac{1}{2}$	$\frac{1}{40}$	$\frac{1}{40}$

Hasil perhitungan nilai probabilitas tersebut selanjutnya diimplementasikan sebagai data acuan untuk proses *testing*.

3.8 K-Fold Cross Validation

Cross Validation adalah sebuah teknik validasi model untuk menilai kinerja model pada dataset (Veronica et al., 2020). *K-fold Cross Validation* atau teknik validasi silang ialah teknik yang diimplementasikan untuk membagi data menjadi k bagian dataset dengan ukuran yang serupa. Pada penelitian ini dibuat 2 model perhitungan yang berbeda. Model pertama adalah model perhitungan menggunakan teknik *Laplacian Correction*, sedangkan model yang kedua adalah perhitungan tanpa menggunakan teknik *Laplacian Correction*. Pada penelitian ini nilai K yang digunakan adalah 5, 10, 15, dan 20-*fold Cross Validation* untuk masing-masing model, yang berarti proses uji coba model dilakukan sebanyak 5, 10, 15, dan 20 kali untuk model pertama dan 5, 10, 15, dan 20 kali juga untuk model yang kedua. Data yang akan digunakan dibagi menjadi data training atau data latih dan data testing

atau data uji. Sebagai contoh, penelitian ini akan menggunakan data sebanyak 2200 record data, jika nilai k yang digunakan adalah 5 maka data akan dibagi menjadi 5 bagian dengan masing-masing 440 data.

3.9 Skenario Pengujian

Tahap paling akhir yang dilakukan ialah tahap pengujian. Pengujian diterapkan untuk mengetahui nilai akurasi, presisi, *recall*, dan *f-measure*. Tahap pengujian atau perhitungan nilai performa ini dilakukan dengan menggunakan *Confusion Matrix*. *Confusion Matrix* merupakan sebuah tabel yang di dalamnya berisi banyak data uji yang diprediksi benar dan salah oleh model klasifikasi (Candra, 2019). Nilai *true false* dijadikan pedoman dalam menghitung nilai pefroma. Tabel 3.14 berikut ini merupakan konsep dari *Confusion Matrix*.

Tabel 3.14 Konsep *Confusion Matrix*

Prediksi Aktual	Kelas Prediksi	
	<i>Negative</i>	<i>Positive</i>
Kelas Negatif	<i>True Negative (TN)</i>	<i>False Positive (FP)</i>
Kelas Positif	<i>False Negative (FN)</i>	<i>True Positive (TP)</i>

Berdasarkan konsep *Confusion Matrix* pada Tabel 3.17 tersebut, maka dapat dilakukan perhitungan nilai performa dengan mengacu pada Persamaan 3.1, 3.2, 3.3, dan 3.4 berikut.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (3.1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (3.2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3.3)$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.4)$$

Dimana,

TP : Jumlah data yang memiliki nilai positif dan diklasifikasikan sebagai data positif.

FP : Jumlah data yang memiliki nilai negatif tetapi diklasifikasikan sebagai data positif.

FN : Jumlah data yang memiliki nilai positif tetapi diklasifikasikan sebagai data negatif.

TN : Jumlah data yang memiliki nilai negatif dan diklasifikasikan sebagai data negatif

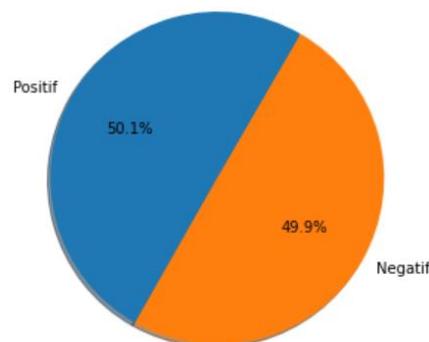
BAB IV

UJI COBA DAN PEMBAHASAN

Penyebaran dan pengujian sistem yang dikembangkan, serta hasil penilaian evaluasi sistem, dijelaskan dalam bab uji coba dan pembahasan. Selain itu, bab ini juga mencakup informasi penelitian, skenario uji coba, temuan uji coba, dan komentar yang mencakup hasil uji coba dan kesesuaiannya dengan prinsip-prinsip Islam.

4.1 Data Penelitian

Sebagaimana yang telah dipaparkan sebelumnya, data penelitian ini adalah data komentar instagram pada unggahan akun @prakerja.go.id dalam rentang waktu 26 Agustus 2022 hingga 9 Desember 2022 sebanyak 2200 komentar. Pengumpulan data komentar instagram dilakukan menggunakan salah satu ekstensi pada *browser* Google Chrome yaitu *Data Miner*. Setelah dilakukan pelabelan data dan validasi data oleh ahli bahasa, didapatkan komentar positif sebanyak 1103 komentar sedangkan komentar dengan sentimen negatif berjumlah 1097 komentar. Visualisasi perbandingan antara sentimen positif dan negatif dapat dilihat pada Gambar 4.1.



Gambar 4. 1 Grafik Perbandingan Kelas Sentimen

Pada Gambar 4.1 tersebut, terlihat bahwa data dengan sentimen positif sebesar 50,1% dan data dengan sentimen negatif sebesar 49,9% dari jumlah keseluruhan data komentar yang digunakan. Tabel 4.1 menyajikan sampel data komentar.

Tabel 4.1 Sampel Data Komentar

No	Komentar	Label
1.	Sudah 20 kali coba gak lolos2, sekarang malah gak bisa gabung gelombang lagi	Negatif
2.	Gagal lagi nggak pernah dapet gelombang, kayanya emang gak cocok ikutan ini	Negatif
3.	daftar mentok diverifikasi wajah ga bisa bisa	Negatif
4.	Cuma mau kasih saran buat admin prakerja. Di gali lagi SDM nya supaya bisa lebih baik. Banyak belajar mengenai sistem informatika yang ideal dan sesuai dengan peserta dari berbagai kalangan. Terlebih sistem otp gmail . Sangat useless dan menimbulkan masalah ke depan.	Negatif
5.	beli pelatihan di mastering skill ko lama y muncul sertifikat di dasbord nya 😞	Negatif
6.	Gangguan terus...	Negatif
7.	Udah miskin upload ktp susah dipersulit mulu dah idup gw	Negatif
8.	4x daftar gelombang dari tahun 2020 tapi ehh 😞 aneh gagal mulu, yg bisa ngakalin Negatif malah pda lolos	Negatif
9.	Gak lolos,dan udah dua kali alasannya NIK terdaftar dapat bantuan sosial 😞	Negatif
10.	Gelombang pilihan,, ga di patok kuotanya, biar akun rame di kunjungi aja, dana dikit di mainin 😊	Negatif
11.	Semangaaatttt ikut lg semoga lolos	Positif
12.	Alhamdulillah, pertama kali daftar langsung bisa loloss 😊	Positif
13.	Selamat Buat yang Lolos Kartu Prakerja Gelombang 47 Mimin Doakan yang belum lolos bisa Lolos di Gelombang 48 ya Amin 🙏	Positif
14.	Selamat buat teman-teman yang lolos 😊	Positif
15.	Halo, Selamat yaaa untuk yang lolos diGelombang 47! 🎉👏	Positif
16.	Alhamdulillah masyaAllah sekian lamanya daftar akhirnya lolos juga 😊😊😊	Positif
17.	Alhamdulillah pertama gabung lngsung lolos 😊 cuma masih blank gk tau step slanjutnya. Udah nymbung sma ewalet , tpi msih bingung pelatihan nya gmna . Prtma jadi msih bngung	Positif
18.	Alhamdulillah semua misi pelatihan sudah dilakuin sertifikat juga lengkap,ulasan juga sudah,dan sdh ada jadwal intensif nya,tinggal tunggu nya cair aja aminn 🙏🙏	Positif
19.	ALHAMDULLILAHHH 🙏🙏 Yuk coba absen dulu yang akhirnya berhasil lolos Prakerja 😊	Positif
20.	Alhamdulillah makasih prakerja	Positif

4.2 Uji Coba

Tahapan uji coba merupakan sebuah tahapan dalam penelitian yang di dalamnya berisi alur proses atau langkah-langkah untuk menguji sebuah sistem yang sudah dibuat. Tahap pengujian diimplementasikan dengan menghitung nilai *confusion matrix* dari sistem yang dibangun. Sistem tersebut sendiri dibangun menggunakan bahasa pemrograman Python pada Google Colaboratory. Selain itu, pengujian kinerja model pada dataset akan dilakukan menggunakan teknik *K-fold Cross Validation*. Pada penelitian ini dilakukan 2 skenario pengujian menggunakan *K-fold Cross Validation* yang masing-masing skenario akan digunakan 5, 10, 15, dan 20 *fold* atau nilai $k = [5, 10, 15, 20]$. Skenario pertama adalah pengujian model *Naive Bayes* menggunakan teknik *Laplacian Correction* dan skenario kedua adalah pengujian model *Naive Bayes* tanpa menggunakan teknik *Laplacian Correction*.

4.2.1 Tahap Penelitian

Tahap penelitian ini merupakan tahap dalam membangun sistem yang langkah-langkahnya telah dijelaskan pada bab sebelumnya. Sistem dibuat dengan *Google Collaboratory* dan bahasa pemrograman *Python*.

a. *Text Preprocessing*

Tahap *text preprocessing* merupakan sebuah tahapan yang bermanfaat untuk membuat data yang tidak terstruktur menjadi data terstruktur yang siap diuji. Pada penelitian ini, proses *text preprocessing* akan dilakukan seperti pada Gambar 3.3.

Tahap pertama adalah *cleansing*, tahapan *cleansing* sendiri dilakukan untuk menghapus komponen-komponen yang tidak penting, karakter-karakter yang tidak

baku, atau komponen-komponen yang tidak ada kaitannya dengan dokumen seperti *emoticon*, angka, tanda baca, simbol, spasi berlebih, *enter*, tautan, dan tagar. Dalam proses *cleansing* ini, digunakan salah satu modul pada python yaitu modul *re* atau *regex*. Proses implementasi tahap *cleansing* dapat dilihat pada Gambar 4.2. Sedangkan hasil implementasi tahap *cleansing* dapat dilihat pada Gambar 4.3.

```

Import string
Import re
FUNCTION cleansing_data (text)
    text <- remove tab, new line, and back slice
    text <- remove non ASCII
    text <- remove mention, link, hashtag
    text <- remove incomplete URL
    text <- remove number
    text <- remove punctuation
    text <- remove whitespace leading and trailing
    text <- remove multiple whitespace into single whitespace
    RETURN text
END FUNCTION
data(cleansing) <- data(komentar).apply(cleansing_data)
hasil_cleansing <- data(cleansing)
PRINT hasil_cleansing

```

Gambar 4.2 *Pseudocode* Tahap *Cleansing*

	komentar	sentimen
0	Kok sy dpt TF cm rb y bln ini	negatif
1	semoga lolos	positif
2	Gelombang selanjutnya kapan min	positif
3	Sudah kali coba gak lolos sekarang malah gak b...	negatif
4	Gelombangnyab doang buka akun lama gak di Perh...	negatif

Gambar 4.3 Hasil *Cleansing*

Tahap yang kedua adalah *case folding*, yaitu merubah semua huruf yang ada pada teks menjadi *lowercase* atau huruf kecil. Langkah ini perlu dilakukan untuk meminimalisir kesalahan saat proses pencarian kata. Tahap *case folding* sendiri menggunakan fungsi *lower* pada kelas *series string* yang ada pada *library pandas*. Proses implementasi tahap *case folding* dapat dilihat pada Gambar 4.4, sedangkan untuk hasil implementasi tahap *case folding* dapat dilihat pada Gambar 4.5.

```

Import string
Import re
data(casefolding) <- data(cleansing).lower()
hasil_casefolding <- data(casefolding)
PRINT hasil_casefolding

```

Gambar 4.4 Pseudocode Tahap Case Folding

	komentar	sentimen
0	kok sy dpt tf cm rb y bln ini	negatif
1	semoga lolos	positif
2	gelombang selanjutnya kapan min	positif
3	sudah kali coba gak lolos sekarang malah gak b...	negatif
4	gelombangnyab doang buka akun lama gak di perh...	negatif

Gambar 4.5 Hasil Case Folding

Tahap yang ketiga adalah *tokenizing*, tahap *tokenizing* merupakan sebuah tahapan memotong atau memecah kalimat menjadi kata-kata penyusunnya. Proses *tokenizing* dilakukan menggunakan fungsi *word tokenize* yang ada pada *library*

nlk. Proses implementasi tahap *tokenizing* dapat dilihat pada Gambar 4.6, sedangkan hasil implementasi *tokenizing* dapat dilihat pada Gambar 4.7.

```

Import nltk
Import word_tokenize
FUNCTION word_tokenize_wrapper(text)
    RETURN word_tokenize
END FUNCTION
data(komen_token) <- data(casefolding).apply(word_tokenize_wrapper)
PRINT(data(komen_token))

```

Gambar 4.6 Pseudocode Tahap *Tokenizing*

	komentar	sentimen
0	[kok, sy, dpt, tf, cm, rb, y, bln, ini]	negatif
1	[semoga, lolos]	positif
2	[gelombang, selanjutnya, kapan, min]	positif
3	[sudah, kali, coba, gak, lolos, sekarang, mala...]	negatif
4	[gelombangnyab, doang, buka, akun, lama, gak, ...]	negatif

Gambar 4.7 Hasil *Tokenizing*

Tahap yang keempat adalah *normalization*, yaitu tahap yang dilakukan untuk mengubah seluruh kata yang memiliki kesalahan dalam ejaan, *noise*, ataupun kata yang tidak baku menjadi kata baku. Pada tahap ini digunakan bantuan kamus perbaikan kata yang di dalamnya berisi kumpulan kata-kata yang baku. Proses implementasi tahap *normalization* dapat dilihat pada Gambar 4.8, sedangkan hasil implementasi *normalization* dapat dilihat pada Gambar 4.9.

```

Normalized_word <- read kamus-perbaikan-kata.xlsx
Set normalized_word_dict into empty dictionary
FOR index, row in normalized_word.interrows()
  IF row(0) not in normalized_word_dict
    Set normalized_word_dict(row(0) <- row(1))
  END IF
FUNCTION normalized_term(document)
  RETURN normalize_word_dict(term) if term in normalized_word_dict else
  term for term in document
END FUNCTION
data(komen_perbaikan) <- data(komen_token).apply(normalized_term)
PRINT(data(komen_perbaikan))

```

Gambar 4.8 Pseudocode Tahap Normalization

```

0      [kok, saya, dapat, transfer, cuma, ribu, iya, ...
1                                     [semoga, lolos]
2      [gelombang, selanjutnya, kapan, admin]
3      [sudah, kali, coba, tidak, lolos, sekarang, ma...
4      [gelombang, saja, buka, akun, lama, tidak, di,...
      ...
2195                                     [ntb, zonk, terus]
2196      [sekarang, sudah, bisa, kak]
2197 [kalau, lupa, surel, lupa, sandi, yang, pernah...
2198 [yang, punya, kartu, kis, bisa, lolos, tidak, ...
2199                                     [gagal, lagi]
Name: komen_perbaikan, Length: 2200, dtype: object

```

Gambar 4.9 Hasil Normalization

Tahap kelima adalah *stopword removal*, yaitu sebuah tahapan yang berfungsi untuk menghilangkan kata-kata umum yang sering muncul dan tidak memiliki makna tertentu agar jumlah kata yang tersimpan dalam daftar token tidak terlalu banyak. Proses implementasi tahap *stopword removal* dapat dilihat pada Gambar 4.10, sedangkan hasil implementasi dari *stopword removal* sendiri dapat dilihat pada Gambar 4.11 berikut ini.

```

Set list_stopwords with words that want to remove
FUNCTION stopwords_removal (words)
  RETURN word for word in words if word not in list_stopwords
END FUNCTION
data(komen_filtered) <- data(komen_perbaikan).apply(stopwords_removal)
PRINT(data(komen_filtered))

```

Gambar 4.10 Pseudocode Tahap Stopword Removal

	komentar	sentimen
0	[dapat, transfer, bulan]	negatif
1	[semoga, lolos]	positif
2	[gelombang, selanjutnya, kapan]	positif
3	[sudah, coba, tidak, lolos, sekarang, tidak, b...]	negatif
4	[gelombang, buka, akun, lama, tidak, perhatikan]	negatif

Gambar 4.11 Hasil *Stopword Removal*

Tahap terakhir dalam *text preprocessing* adalah *stemming*, tahap *stemming* sendiri merupakan sebuah tahap yang dilakukan untuk mengubah seluruh kata dalam dokumen menjadi kata dasar dengan menghapus seluruh imbuhan kata. Pada proses *stemming* ini menggunakan *library* sastrawi, yaitu *library* yang berfungsi untuk merubah kata berimbuhan menjadi kata dasar dalam Bahasa Indonesia. Proses implementasi tahap *stemming* dapat dilihat pada Gambar 4.12, sedangkan hasil implementasi *stemming* dapat dilihat pada Gambar 4.13.

```

Install sastrawi
Import stemmerfactory
Create stemmer
FUNCTION stemmed_wrapper(term)
    RETURN stemmer.stem(term)
END FUNCTION
Set term_dict into empty dictionary
FOR document in dataset(komentar)
    FOR term in document
        IF term not in term_dict
            Set term_dict(term) into whitespace
        END IF
    FOR term in term_dict
        Set term_dict(term) into stemmed_wrapper(term)
    FUNCTION get_stemmed_term (document)
        RETURN term_dict(term) for term in document
    END FUNCTION
dataset(komentar) <- dataset(komentar).apply(get_stemmed_term)
PRINT(dataset)

```

Gambar 4.12 *Pseudocode* Tahap *Stemming*

	komentar	sentimen
0	[dapat, transfer, bulan]	negatif
1	[moga, lolos]	positif
2	[gelombang, lanjut, kapan]	positif
3	[sudah, coba, tidak, lolos, sekarang, tidak, b...]	negatif
4	[gelombang, buka, akun, lama, tidak, perhati]	negatif

Gambar 4.13 Hasil *Stemming*

b. Pembobotan TF-IDF

Dari hasil kata yang telah melalui tahapan *text preprocessing*, selanjutnya akan dilakukan tahap pembobotan kata menggunakan metode TF-IDF, yaitu metode pengembangan dari metode-metode sebelumnya dan merupakan gabungan dari dua metode yang berbeda yaitu metode TF dan IDF. Sehingga, TF-IDF memiliki nilai perhitungan yang didapatkan dari hasil perkalian 2 nilai, yaitu nilai TF dan nilai IDF (Ulinuha, 2022). Pembobotan kata menggunakan TF-IDF dilakukan dengan menggunakan *library CountVectorizer* dan *TfidfTransformer*. *Library CountVectorizer* akan merubah teks menjadi matriks TF. Kemudian *library TfidfTransformer* akan menghitung skema TF-IDF dan mendapatkan matriks TF-IDF.

c. *Naive Bayes Classifier*

Proses implementasi metode *Naive Bayes* dilakukan dengan menggunakan model *MultinomialNB*. Sedangkan untuk mengimplementasikan teknik validasi model *K-Fold Cross Validation* menggunakan modul *cross_val_score*. Hal yang pertama dilakukan adalah mengundul model dengan *import MultinomialNB* dan

cross_val_score. Karena basis perhitungan metode *Naive Bayes* bertumpu pada probabilitas, sehingga sangat memungkinkan suatu atribut akan memiliki nilai probabilitas 0 (nol). Hal tersebut dapat menyebabkan seluruh perhitungan menjadi rusak atau tidak sesuai. Oleh karena itu, diperlukan teknik *Laplacian Correction* untuk menangani hal tersebut. Pada model *MultinomialNB* sendiri, telah disediakan sebuah parameter yang berfungsi sebagai *Laplacian Correction*, yaitu parameter *alpha*. Agar atribut perhitungan metode *Naive Bayes* tidak sama dengan 0 (nol), dapat dilakukan dengan menginisialisasikan parameter *alpha* sama dengan 1. Sedangkan untuk model *Naive Bayes* yang tidak menggunakan teknik *Laplacian Correction* parameter *alpha* diinisialisasikan dengan nilai 0.

Setelah itu, dilakukan perhitungan skor *K-fold Cross Validation* untuk model yang menggunakan *Laplacian Correction* dan model yang tidak menggunakan *Laplacian Correction*. Nilai K yang digunakan dalam tahap perhitungan *K-fold Cross Validation* adalah $K = [5,10,15,20]$.

Setelah itu, barulah dilakukan proses klasifikasi menggunakan metode *Multinomial Naive Bayes*. Pada Tabel 4.2 berikut ini merupakan contoh data testing yang telah melalui tahap *text preprocessing* dan pembobotan TF yang diambil dari beberapa data sampel yang disajikan pada Tabel 4.1.

Tabel 4.2 Contoh Data *Testing*

Dokumen	Fitur	Kelas
D2	gagal tidak pernah dapat gelombang tidak cocok ikut	?
D12	alhamdulillah pertama daftar lolos	?
D14	selamat teman lolos	?

Kemudian pada proses klasifikasi metode *Naive Bayes* akan dilakukan dicari perhitungan probabilitas tertinggi yang mengacu pada persamaan 2.8 dan

didasarkan pada nilai *conditional probabilities* seperti pada Tabel 3.8 hingga Tabel 3.13. Berikut ini merupakan contoh Vmap positif dan Vmap negatif yang mengacu pada persamaan 2.8 dan nilai *conditional probabilities* pada Tabel 3.8 hingga Tabel 3.13.

$$\begin{aligned}
 &P(D2 / V_{positif}) \\
 &= (P(W_{gagal} / V_{positif}) \times P(W_{tidak} / V_{positif}) \times P(W_{pernah} / V_{positif}) \times P(W_{dapat} / V_{positif}) \times \\
 &P(W_{gelombang} / V_{positif}) \times P(W_{tidak} / V_{positif}) \times P(W_{cocok} / V_{positif}) \times P(W_{ikut} / V_{positif})) \times \\
 &P(V_{positif}) \\
 &= \left(\frac{1}{35} \times \frac{1}{35} \times 1 \times 1 \times \frac{1}{35} \times \frac{1}{35} \times 1 \times 1 \right) \times \frac{1}{2} \\
 &= (0,028571 \times 0,028571 \times 1 \times 1 \times 0,028571 \times 0,028571 \times 1 \times 1) \times 0,5 \\
 &= 0,000000666349 \times 0,5 \\
 &= 0,0000003331745
 \end{aligned}$$

$$\begin{aligned}
 &P(D2 / V_{negatif}) \\
 &= (P(W_{gagal} / V_{negatif}) \times P(W_{tidak} / V_{negatif}) \times P(W_{pernah} / V_{negatif}) \times P(W_{dapat} / V_{negatif}) \\
 &\times P(W_{gelombang} / V_{negatif}) \times P(W_{tidak} / V_{negatif}) \times P(W_{cocok} / V_{negatif}) \times P(W_{ikut} / V_{negatif}) \\
 &) \times P(V_{negatif}) \\
 &= \left(\frac{3}{40} \times \frac{2}{40} \times 1 \times 1 \times \frac{2}{40} \times \frac{2}{40} \times 1 \times 1 \right) \times \frac{1}{2} \\
 &= (0,075 \times 0,05 \times 1 \times 1 \times 0,05 \times 0,05 \times 1 \times 1) \times 0,5 \\
 &= 0,000009375 \times 0,5 \\
 &= 0,0000046875
 \end{aligned}$$

$$P(D12 / V_{positif})$$

$$\begin{aligned}
&= (P(W_{alhamdulillah} / V_{positif}) \times P(W_{pertama} / V_{positif}) \times P(W_{daftar} / V_{positif}) \times P(W_{lolos} / \\
&V_{positif})) \times P(V_{positif}) \\
&= \left(\frac{2}{35} \times 1 \times \frac{1}{35} \times \frac{4}{35} \right) \times \frac{1}{2} \\
&= (0,05714 \times 1 \times 0,028571 \times 0,114285) \times 0,5 \\
&= 0,001865849 \times 0,5 \\
&= 0,0009329245
\end{aligned}$$

$$\begin{aligned}
&P(D12 / V_{negatif}) \\
&= (P(W_{alhamdulillah} / V_{negatif}) \times P(W_{pertama} / V_{negatif}) \times P(W_{daftar} / V_{negatif}) \times P(W_{lolos} / \\
&V_{negatif})) \times P(V_{negatif}) \\
&= \left(\frac{1}{40} \times 1 \times \frac{2}{40} \times \frac{2}{40} \right) \times \frac{1}{2} \\
&= (0,025 \times 0,05 \times 0,05) \times 0,5 \\
&= 0,0000625 \times 0,5 \\
&= 0,00003125
\end{aligned}$$

$$\begin{aligned}
&P(D14 / V_{positif}) \\
&= (P(W_{selamat} / V_{positif}) \times P(W_{teman} / V_{positif}) \times P(W_{lolos} / V_{positif})) \times P(V_{positif}) \\
&= \left(\frac{2}{35} \times \frac{3}{35} \times \frac{4}{35} \right) \times \frac{1}{2} \\
&= (0,05714 \times 0,085714 \times 0,114285) \times 0,5 \\
&= 0,00055973 \times 0,5 \\
&= 0,00027986
\end{aligned}$$

$$\begin{aligned}
&P(D14 / V_{negatif}) \\
&= (P(W_{selamat} / V_{negatif}) \times P(W_{teman} / V_{negatif}) \times P(W_{lolos} / V_{negatif})) \times P(V_{negatif})
\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{40} \times \frac{1}{40} \times \frac{2}{40}\right) \times \frac{1}{2} \\
&= (0,025 \times 0,025 \times 0,05) \times 0,5 \\
&= 0,00003125 \times 0,5 \\
&= 0,000015625
\end{aligned}$$

Kemudian untuk menentukan data tersebut masuk ke dalam kelas positif atau negatif, maka akan dilihat dari nilai probabilitas tertinggi dari dokumen tersebut. Pada Tabel 4.3 berikut ini merupakan nilai probabilitas setiap dokumen yang telah dihitung.

Tabel 4.3 Nilai Probabilitas Data *Testing*

Dokumen	Probabilitas	
	Positif	Negatif
D2	0,0000003331745	0,0000046875
D12	0,0009329245	0,00003125
D14	0,00027986	0,000015625

Dari hasil perhitungan probabilitas pada Tabel 3.9 dapat dianalisis pada dokumen 2 atau D2 diperoleh nilai probabilitas negatif lebih besar dari nilai probabilitas positif, sehingga dapat ditarik kesimpulan bahwa D2 termasuk ke dalam kelas sentimen negatif. Pada D12 diperoleh hasil probabilitas positif lebih besar daripada nilai probabilitas negatif, sehingga D12 termasuk ke dalam kelas sentimen positif. Sedangkan pada dokumen terakhir yaitu D14 nilai probabilitas positif lebih besar dari nilai probabilitas negatif, sehingga dapat ditarik kesimpulan bahwa D14 termasuk ke dalam kelas sentimen positif.

4.3 Hasil Uji Coba

Proses uji coba dalam penelitian ini dilakukan untuk mengevaluasi dan mengukur nilai performa sistem yang telah dibangun. Hasil dari proses uji coba

yang dilakukan adalah berupa nilai *Confusion Matrix*. Untuk mendapatkan nilai *matrix*, digunakan modul *confusion_matrix* pada *library scikit-learn*. Berikut merupakan hasil *matrix* dan nilai akurasi, presisi, *recall*, dan *f-measure*.

```

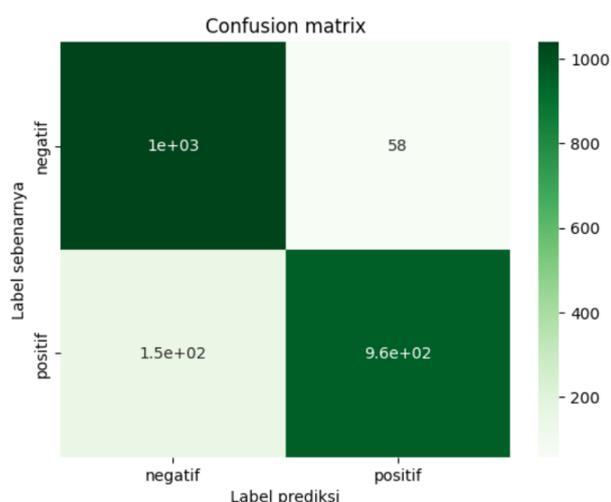
test time: 0.009s
accuracy: 0.907
      precision    recall  f1-score   support

negatif    0.88     0.95     0.91     1097
positif    0.94     0.87     0.90     1103

accuracy          0.91         0.91         0.91     2200
macro avg         0.91         0.91         0.91     2200
weighted avg      0.91         0.91         0.91     2200

```

Gambar 4.14 Nilai Akurasi *Testing*



Gambar 4.15 Hasil *Confusion Matrix*

Gambar 4.15 merupakan nilai *matrix* yang akan digunakan sebagai dasar atau acuan untuk menghitung nilai performa dari sistem yang telah dibangun. Nilai *True Positive* (TP) yang dihasilkan adalah 1500, nilai *False Negative* (FN) adalah 58, nilai *False Positive* (FP) adalah 130, dan nilai *True Negative* (TN) adalah 960.

Berikut merupakan perhitungan untuk mendapatkan nilai akurasi, presisi, *recall*, dan *f-measure*.

$$\begin{aligned}
 \text{Akurasi} &= \frac{960 + 1500}{960 + 130 + 58 + 1500} \times 100\% \\
 &= \frac{2460}{2648} \times 100\% \\
 &= 0,929 \times 100\% \\
 &= 93\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Presisi} &= \frac{960}{960 + 58} \times 100\% \\
 &= \frac{960}{1018} \times 100\% \\
 &= 0,943 \times 100\% \\
 &= 94\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Recall} &= \frac{960}{960 + 130} \times 100\% \\
 &= \frac{960}{1090} \times 100\% \\
 &= 0,88 \times 100\% \\
 &= 88\%
 \end{aligned}$$

$$\begin{aligned}
 f - \text{measure} &= 2 \times \frac{94 \times 88}{94 + 88} \times 100\% \\
 &= 2 \times \frac{8272}{182} \times 100\% \\
 &= 2 \times 45,45 \times 100\% \\
 &= 90,9 \times 100\% \\
 &= 91\%
 \end{aligned}$$

Dari hasil perhitungan, diperoleh untuk nilai akurasi adalah sebesar 93%, nilai presisi 94%, nilai *recall* 88%, dan nilai *f-measure* 91%. Langkah selanjutnya adalah mengimplementasikan model yang telah dibuat ke dalam klasifikasi sebuah kalimat. Data testing yang akan diklasifikasi berupa data sampel sebanyak 20 data komentar.

Hasil perhitungan skor *K-fold Cross Validation* untuk model yang menggunakan *Laplacian Correction* dan model yang tidak menggunakan *Laplacian Correction* dengan nilai $K = [5,10,15,20]$ dapat dilihat pada Tabel 4.4 berikut.

Tabel 4.4 Perbandingan Skor *K-fold Cross Validation*

Kondisi	K = 5	K = 10	K = 15	K = 20
Model Menggunakan <i>Laplacian Correction</i>	84,2%	85%	85,5%	85,6%
Model Tanpa Menggunakan <i>Laplacian Correction</i>	82,5%	83,7%	83,6%	84%

Berdasarkan hasil perbandingan pada Tabel 4.4, dapat dibuktikan bahwa model *Naive Bayes* yang menggunakan teknik *Laplacian Correction* menghasilkan skor *K-fold Cross Validation* yang jauh lebih baik. Karena penggunaan teknik *Laplacian Correction* berfungsi untuk menangani hasil probabilitas *Naive Bayes* agar tidak bernilai 0 (nol) yang kemudian dapat merusak hasil perhitungan. Selain itu, berdasarkan hasil pada Tabel 4.4, diketahui bahwa nilai skor *K-fold Cross Validation* terbaik didapatkan pada saat nilai $K = 20$ dengan hasil skor 85,6%.

Hasil klasifikasi model *Naive Bayes* dari 20 data dapat dilihat pada Tabel 4.5 berikut.

Tabel 4.5 Hasil Klasifikasi

No	Data Komentar	Labeling Manual	Klasifikasi Sistem
1	Sudah 20 kali coba gak lolos2, sekarang malah gak bisa gabung gelombang lagi	Negatif	Negatif
2	Gelombangnyab doang buka akun lama gak di Perhatikan	Negatif	Positif
3	daftar mentok diverifikasi wajah ga bisa bisa	Negatif	Negatif
4	Cuma mau kasih saran buat admin prakerja. Di gali lagi SDM nya supaya bisa lebih baik. Banyak belajar mengenai sistem informatika yang ideal dan sesuai dengan peserta dari berbagai kalangan. Terlebih sistem otp gmail . Sangat useless dan menimbulkan masalah ke depan.	Negatif	Negatif
5	beli pelatihan di mastering skill ko lama y muncul sertifikat di dasbord nya 😞	Negatif	Positif
6	Gangguan terus...	Negatif	Negatif
7	Udah miskin upload ktp susah dipersulit mulu dah idup gw	Negatif	Negatif
8	4x daftar gelombang dari tahun 2020 tapi ehh 😞 aneh gagal mulu, yg bisa ngakalin Negatif malah pda lolos	Negatif	Negatif
9	Gak lolos,dan udah dua kali alasannya NIK terdaftar dapat bantuan sosial 😞	Negatif	Negatif
10	Gelombang pilihan,, ga di patok kuotanya, biar akun rame di kunjungi aja, dana dikit di mainin 😊	Negatif	Negatif
11	Semangaaatttt ikut lg semoga lolos	Positif	Positif
12	Alhamdulillah, pertama kali daftar langsung bisa loloss 😊	Positif	Positif
13	Selamat Buat yang Lolos Kartu Prakerja Gelombang 47 Mimin Doakan yang belum lolos bisa Lolos di Gelombang 48 ya Amin 🙏	Positif	Positif
14	Selamat buat teman-teman yang lolos 😊	Positif	Positif
15	Halo, Selamat yaaa untuk yang lolos diGelombang 47! 🙏👏	Positif	Positif
16	Alhamdulillah masyaAllah sekian lamanya daftar akhirnya lolos juga 😊😞😞	Positif	Positif
17	Alhamdulillah pertama gabung lngsung lolos 😊 cuma masih blank gk tau step slanjutnya. Udah nymbung sma ewalet , tpi msih bingung pelatihan nya gmna . Prtma jadi msih bngung	Positif	Positif
18	Alhamdulillah semua misi pelatihan sudah dilakuin sertifikat juga lengkap,ulasan juga sudah,dan sdh ada jadwal intensif nya,tinggal tunggu nya cair aja aminn 🙏🙏	Positif	Positif
19	ALHAMDULLILAHHH 🙏🙏 Yuk coba absen dulu yang akhirnya berhasil lolos Prakerja 😊	Positif	Positif
20	Alhamdulillah makasih prakerja	Positif	Positif

4.4 Pembahasan

Temuan uji coba menyatakan sebanyak 2200 teks komentar yang telah dikumpulkan selanjutnya mengimplementasikan pelabelan secara manual oleh ahli bahasa. Data tersebut diklasifikasikan menjadi dua kelas yaitu kelas positif dan kelas negatif. Setelah dilakukan pelabelan secara manual, didapati hasil untuk sentimen positif memiliki jumlah yang lebih banyak yaitu sebanyak 1103 komentar sedangkan untuk sentimen negatif sebanyak 1097 komentar. Kemudian, data yang terlabel akan melalui tahap *text preprocessing* dan pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Setelah itu, dilakukan pemodelan menggunakan metode *Naive Bayes* dan dalam proses pemodelan tersebut dilakukan pula teknik validasi model *K-fold Cross Validation*. Pengujian menggunakan *K-fold Cross Validation* dilakukan pada dua kondisi model yang berbeda. Kondisi yang pertama dilakukan pada model *Naive Bayes* yang tidak menggunakan teknik *Laplacian Correction*. Sedangkan kondisi yang kedua dilakukan pada model *Naive Bayes* yang menggunakan teknik *Laplacian Correction*.

Berdasarkan hasil pengujian, diperoleh skor *K-fold Cross Validation* untuk model yang menggunakan teknik *Laplacian Correction* lebih tinggi dibandingkan model yang tidak menggunakan teknik *Laplacian Correction*. Hal tersebut menunjukkan bahwa teknik *Laplacian Correction* diperlukan agar hasil klasifikasi yang dilakukan oleh metode *Naive Bayes* lebih optimal. Selain itu, nilai skor *K-fold Cross Validation* terbaik didapatkan pada saat nilai $K = 20$ dengan hasil skor 85,6%.

Hasil pengukuran nilai performa menggunakan *confusion matrix* diperoleh nilai *True Positive* (TP) adalah 1500, nilai *False Negative* (FN) adalah 58, nilai *False Positive* (FP) adalah 130, dan nilai *True Negative* (TN) adalah 960. Nilai-nilai tersebut kemudian akan dijadikan sebagai acuan untuk menghitung nilai performa sistem menggunakan rumus-rumus yang telah dijelaskan pada bab sebelumnya. Berdasarkan rumus-rumus tersebut, diperoleh nilai akurasi sebesar 93%, nilai presisi sebesar 94%, nilai *recall* sebesar 88%, dan nilai *f-measure* sebesar 91%. Hasil nilai performa yang didapatkan oleh metode *Naive Bayes* pada penelitian ini dipengaruhi oleh beberapa faktor, salah satunya adalah hasil dari proses pelabelan data yang dilakukan pada tahap awal, hal tersebut terjadi karena metode *Naive Bayes* adalah metode klasifikasi pembelajaran terawasi di mana hasil prediksi dalam suatu kelas ditentukan oleh pola awal yang dibentuk dengan mengolah data pelatihan dan kemudian dimasukkan ke dalam model.

Pada penelitian ini, hasil klasifikasi/pengkategorian sistem menggunakan metode *Naive Bayes* menunjukkan bahwa komentar dengan sentimen negatif mempunyai hasil yang cukup banyak dibandingkan komentar dengan sentimen positif, yaitu untuk sentimen negatif sebanyak 1186 dan untuk sentimen positif sebanyak 1014. Dari hasil klasifikasi sistem tersebut, menunjukkan bahwa sentimen masyarakat mengenai kartu prakerja sendiri cukup berimbang antara sentimen negatif dan positif jika dilihat dari perbandingan jumlah yang sangat tipis. Dari hasil analisis sentimen terhadap kebijakan kartu prakerja tersebut, dapat digunakan oleh pemerintah yang dalam hal ini lebih spesifik kepada Kementerian Ketenagakerjaan sebagai bahan rujukan dan evaluasi untuk memperbaiki dan meningkatkan kualitas

dari kebijakan kartu prakerja. Kemudian pada uji coba klasifikasi yang dilakukan dengan menggunakan 20 data sampel menunjukkan bahwa sebanyak 18 data tersebut berhasil diklasifikasi dengan benar menggunakan model *Naive bayes*. Dari hasil klasifikasi tersebut, masih terdapat klasifikasi yang salah karena nilai performa yang dihasilkan oleh sistem tidak 100%.

Dalam data komentar yang digunakan pada penelitian ini, masyarakat masih banyak yang kurang bijak dalam memberikan komentar, seperti menggunakan kata-kata yang kasar dan kalimat yang terkesan menyudutkan beberapa pihak. Hal tersebut tidak mencerminkan identitas masyarakat Indonesia yang mayoritas penduduknya beragama Islam. Karena di dalam agama Islam sendiri, Allah telah memerintahkan umat Islam untuk menyampaikan aspirasi atau pendapat menggunakan kata-kata yang baik dan sopan. Hal ini dijelaskan sebagai berikut dalam Al-Qur'an Al-Baqarah ayat 83:

وَإِذْ أَخَذْنَا مِيثَاقَ بَنِي إِسْرَائِيلَ لَا تَعْبُدُونَ إِلَّا اللَّهَ وَبِالْوَالِدَيْنِ إِحْسَانًا وَذِي الْقُرْبَىٰ وَالْيَتَامَىٰ وَالْمَسْكِينِ وَقُولُوا لِلنَّاسِ حُسْنًا وَأَقِيمُوا الصَّلَاةَ وَآتُوا الزَّكَاةَ ثُمَّ تَوَلَّيْتُمْ إِلَّا قَلِيلًا مِّنْكُمْ وَأَنتُمْ مُّعْرِضُونَ

“Dan (ingatlah), ketika Kami mengambil janji dari Bani Israil (yaitu): Janganlah kamu menyembah selain Allah, dan berbuat kebaikanlah kepada ibu bapa, kaum kerabat, anak-anak yatim, dan orang-orang miskin, serta ucapkanlah kata-kata yang baik kepada manusia, dirikanlah shalat dan tunaikanlah zakat. Kemudian kamu tidak memenuhi janji itu, kecuali sebahagian kecil daripada kamu, dan kamu selalu berpaling.”(Q.S. Al-Baqarah: 83)

Menurut tafsir Kementerian Agama Republik Indonesia, ayat tersebut menjelaskan mengenai janji-janji yang harus dipenuhi oleh Bani Israil kepada Allah SWT. Janji-janji tersebut antara lain adalah Bani Israil dilarang menyembah sesuatu selain Allah, berbuat kebajikan kepada orang tua dan kerabat, berbuat baik kepada

anak-anak yatim dan orang-orang miskin, diperintahkan untuk mengucapkan kata-kata yang baik kepada sesama manusia, dan yang terakhir adalah perintah untuk mendirikan sholat dan menunaikan zakat. Perintah-perintah yang telah dijelaskan tersebut tentunya memiliki konteks yang lebih luas, di mana perintah tersebut tidak hanya dikhususkan kepada Bani Israil saja, melainkan kepada seluruh umat manusia di muka bumi ini. Oleh karena itu, penting bagi siapapun yang ingin menyampaikan pendapat untuk selalu menyampaikannya dalam tutur kata yang baik (Kemenag, 2022).

Pada masa modern dimana teknologi informasi semakin berkembang pesat dan media sosial kini menjadi salah satu alat komunikasi, maka menjaga ucapan atau lisan kini tidak lagi hanya dikaitkan dengan bergerakinya lisan saja, melainkan juga menjaga tulisan yang akan diketik pada media sosial. Karena melalui tulisan pada media sosial, seseorang juga dapat melakukan dosa yang biasanya dilakukan oleh lisan, seperti berbohong, mencela, dan sebagainya. Selain itu, dalam sebuah hadits yang diriwayatkan oleh Imam Bukhari Rasulullah SAW mengatakan bahwa *سلامة الإنسان في حفظ اللسان*, yang berarti keselamatan manusia sangat bergantung pada kemampuannya dalam menjaga lisan. Oleh karena itu, menuliskan komentar pada media sosial juga harus didasari dengan etika agar komunikasi yang dihasilkan tidak menyinggung, merendahkan, mencela, dan menyudutkan pihak-pihak tertentu (Alfiyah Nur Hasanah & Ikin Asikin, 2022).

Selain itu, dalam penelitian ini juga membahas mengenai kebijakan pemerintah yang berhubungan meningkatkan taraf perekonomian masyarakat. Karena pemerintahan yang baik adalah pemerintahan yang setiap keputusannya

mampu menjadikan masyarakatnya sebagai bangsa yang makmur dan sejahtera.

Hal tersebut dijelaskan di dalam Al-Qur'an surah An-Nisa ayat 58, sebagai berikut:

إِنَّ اللَّهَ يَأْمُرُكُمْ أَنْ تُؤَدُّوا الْأَمَانَاتِ إِلَىٰ أَهْلِهَا وَإِذَا حَكَمْتُمْ بَيْنَ النَّاسِ أَنْ تَحْكُمُوا بِالْعَدْلِ ۗ إِنَّ اللَّهَ نِعِمَّا يَعِظُكُمْ بِهِ ۗ إِنَّ اللَّهَ كَانَ سَمِيعًا بَصِيرًا

“Sesungguhnya Allah menyuruh kamu menyampaikan amanat kepada yang berhak menerimanya, dan (menyuruh kamu) apabila menetapkan hukum di antara manusia supaya kamu menetapkan dengan adil. Sesungguhnya Allah memberi pengajaran yang sebaik-baiknya kepadamu. Sesungguhnya Allah adalah Maha mendengar lagi Maha melihat” (Q.S. An-Nisa’: 58)

Ayat tersebut mengarahkan penyampaian amanat kepada yang berhak, sesuai dengan tafsir Kementerian Agama RI. Dalam ayat ini, kata “amanah” memiliki pengertian yang sangat luas yang mencakup kewajiban Allah kepada hamba-hamba-Nya, serta mandat seseorang kepada orang lain dan diri sendiri. Selain itu, Q.S. An-Nisa ayat 58 tersebut lebih spesifik membahas mengenai sifat adil para pemimpin atau pemegang kekuasaan terhadap rakyatnya serta untuk selalu menjaga amanat yang telah dipercayakan kepada dirinya (Kemenag, 2022). Sehingga dapat disimpulkan bahwa pemerintah sebagai pemegang kekuasaan bertugas sebagai penentu keputusan, dan setiap keputusan tersebut harus ditetapkan dengan sebenar-benarnya dan harus membawa kebaikan bagi rakyatnya (Fuad, 2016).

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Hasil studi menyatakan bahwa dengan mengimplementasikan pendekatan *Naive Bayes* untuk analisis sentimen kebijakan kartu prakerja dilakukan dengan mengumpulkan *dataset* yang diperoleh dari hasil *scraping* pada komentar instagram akun @prakerja.co.id sebanyak 2200 komentar. Kemudian *dataset* tersebut akan dilabeli secara manual menjadi dua kategori: positif dan negatif yang selanjutnya akan divalidasi oleh ahli bahasa. Setelah itu, *dataset* yang masih belum terstruktur akan diganti menjadi data yang terstruktur dalam tahap *text preprocessing*. Tahap selanjutnya adalah dilakukan pembobotan kata menggunakan metode TF-IDF. Kemudian masuk pada tahap pemodelan metode *Naive Bayes* menggunakan model *Multinomial Naive Bayes* dengan menggunakan teknik *Laplacian Correction*, yang selanjutnya penerapan validasi model dengan teknik *K-fold Cross Validation*. Tahap terakhir akan diimplementasikan pengujian kinerja sistem menggunakan teknik *Confusion Matrix*.

Dari hasil validasi model menggunakan *K-fold Cross Validation*, didapati bahwa model *Naive Bayes* yang menggunakan teknik *Laplacian Correction* memiliki temuan yang cenderung positif dari model yang tidak menggunakan teknik *Laplacian Correction*. Selain itu, dari keempat nilai K yang digunakan diperoleh hasil bahwa nilai $k = 20$ menghasilkan skor terbaik. Sedangkan dari hasil pengujian kinerja sistem, diperoleh nilai akurasi sebesar 93%, nilai presisi sebesar 94%, nilai *recall* sebesar 88%, dan nilai *f-measure* sebesar 91%. Klasifikasi yang

dilakukan pada 20 data *testing* sampel menghasilkan 18 data dikategorikan akurat, sedangkan hanya 2 data yang tidak akurat. Berdasarkan temuan tersebut, dapat dikatakan bahwa pendekatan *Naive Bayes* berhasil diterapkan, dengan tingkat akurasi sebesar 93%.

5.2 Saran

Peneliti mengetahui bahwa hasil studi ini masih membutuhkan kerja lebih untuk mendapatkan kinerja sistem yang lebih baik berdasarkan hasil pengujian yang telah terima. Akibatnya, penulis merekomendasikan tindakan berikut untuk pengembang penelitian yang akan datang.

1. Menambahkan jumlah kelas klasifikasi, misalnya kelas netral agar hasil klasifikasi lebih akurat.
2. Menambahkan jumlah *dataset* yang digunakan serta mengambil *dataset* pada media sosial lain seperti twitter maupun facebook agar data yang digunakan lebih variatif.
3. Menambahkan jumlah ahli bahasa dalam proses validasi pelabelan data.
4. Mengimplementasikan pada bahasa selain Bahasa Indonesia, misalnya Bahasa Inggris.
5. Mengimplementasikan metode pembelajaran *multi-label classification* agar komentar yang bersifat sarkasme dapat diklasifikasi dengan benar.

DAFTAR PUSTAKA

- Ainiyah, K. (2022). ANALISIS SENTIMEN TERHADAP APLIKASI PEDULILINDUNGI MENGGUNAKAN SELEKSI FITUR QUERY EXPANSION RANKING DENGAN METODE SUPPORT VECTOR MACHINE. *Material Safety Data Sheet*, 33(1), 1–12. http://www2.warwick.ac.uk/fac/sci/whri/research/mushroomresearch/mushroomquality/fungsienvironment%0Ahttps://us.vwr.com/assetsvc/asset/en_US/id/16490607/contents%0Ahttp://www.hse.gov.uk/pubns/indg373hp.pdf
- Alfiyyah Nur Hasanah, & Ikin Asikin. (2022). Nilai-Nilai Pendidikan dalam Hadits Riwayat Imam Ahmad No 11472 tentang Etika Menjaga Lisan. *Jurnal Riset Pendidikan Agama Islam*, 11472, 44–50. <https://doi.org/10.29313/jrpai.v2i1.981>
- Alvianda, F., & Adikara, P. P. (2019). Analisis Sentimen Konten Radikal Di Media Sosial Twitter Menggunakan Metode Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIHK) Universitas Brawijaya*, 3(1), 241–246.
- Ansori, Y., Fahmi, K., & Holle, H. (2022). *Perbandingan Metode Machine Learning dalam Analisis Sentimen Twitter Comparison of Machine Learning Methods in Twitter Sentiment Analysis*. 10(4), 1–6. <https://doi.org/10.26418/justin.v10i4.51784>
- Apif Supriadi, & Fatmasari. (2021). Implementasi Metode Klasifikasi Naive Bayes Pada Sistem Analisis Opini Pengguna Twitter Berbasis Web. *Jurnal Sistem Informasi*, 10(1), 46–54. <https://doi.org/10.51998/jsi.v10i1.356>
- Ardhani, B. A., Chamidah, N., & Saifudin, T. (2021). Sentiment Analysis Towards Kartu Prakerja Using Text Mining with Support Vector Machine and Radial Basis Function Kernel. *Journal of Information Systems Engineering and Business Intelligence*, 7(2), 119. <https://doi.org/10.20473/jisebi.7.2.119-128>
- Consuello, Y. (2020). Analisis Efektifitas Kartu Pra-kerja di Tengah Pandemi Covid-19. *Buletin Hukum Dan Keadilan*, 4(1), 93–100. <http://journal.uinjkt.ac.id/index.php/adalah/article/view/15479>
- Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. *JISKA (Jurnal Informatika Sunan Kalijaga)*, 3(1), 1. <https://doi.org/10.14421/jiska.2018.31-01>
- Fauzan, A. C., & Hikmah, K. (2022). *IMPLEMENTASI ALGORITMA NAIVE BAYES DALAM ANALISIS PENDAHULUAN Pada akhir tahun 2019 , terdapat wabah covid-19 di China tepatnya di Wuhan provinsi Hubei . Virus ini merupakan virus yang sangat mudah menyebar sehingga WHO (Word Health Organization) meneta*. 7(2), 122–128.

- Firnanda, A., Fitri, C., & Ardianto, F. (2021). Evaluasi Program Kartu Prakerja Ditinjau Dari Asas Kemanfaatan Dan Aksesibilitas. *Jurnal Hukum Magnum Opus*, 4(2), 170–180. <https://doi.org/10.30996/jhmo.v4i2.5243>
- Fuad, Z. (2016). *BENTUK PEMERINTAHAN IDEAL DALAM DARUSSALAM – BANDA ACEH Zahri Fuad*.
- Handayani, A. D., & Rachman, A. (2020). Dampak Pandemi COVID-19 Terhadap Motif dan Tren Penelusuran Informasi Kartu Prakerja di Indonesia. *Empower: Jurnal Pengembangan Masyarakat Islam*, 5(2), 111. <https://doi.org/10.24235/empower.v5i2.7187>
- Kusnia, U., & Kurniawan, F. (2022). Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes. *Explore IT*, 14(36), 24–28.
- Noviyanti, W. (2018). ANALISIS SENTIMEN MEDIA SOSIAL (TWITTER) TERHADAP LAYANAN PROVIDER TELEKOMUNIKASI (TELKOMSEL) MENGGUNAKAN METODE MULTINOMIAL NAIVE BAYES. 39–37, 66, עלון דגושע. https://www.fairportlibrary.org/images/files/RenovationProject/Concept_cost_estimate_accepted_031914.pdf
- Nugroho, D. G., Chrisnanto, Y. H., & Wahana, A. (2015). *Analisis Sentimen Pada Jasa Ojek Online ... (Nugroho dkk.)*. 156–161.
- Nurhuda, F., Widya Sihwi, S., & Doewes, A. (2016). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Teknologi & Informasi ITS*, 2(2), 35. <https://doi.org/10.20961/its.v2i2.630>
- Pasuria, S., & Triwahyuningtyas, N. (2022). Pengaruh Angkatan Kerja, Pendidikan, Upah Minimum, Dan Produk Domestik Bruto Terhadap Pengangguran Di Indonesia. *SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan*, 1(6), 795–808. <https://doi.org/10.54443/sibatik.v1i6.94>
- Phan, H. T., Nguyen, N. T., & Hwang, D. (2021). Sentiment Analysis for Social Media: a Survey. *Journal of Computer Science and Cybernetics*, 37(4), 403–428. <https://doi.org/10.15625/1813-9663/37/4/15892>
- Putri, A. W. (2018). *Penggunaan Instagram Stories Sebagai Personal Branding*.
- Rezki, M., Kholifah, D. N., Faisal, M., Priyono, P., & Suryadithia, R. (2020). Analisis Review Pengguna Google Meet dan Zoom Cloud Meeting Menggunakan Algoritma Naive Bayes. *Jurnal Infortech*, 2(2), 264–270. <https://doi.org/10.31294/infortech.v2i2.9286>
- Rizki, M., Arhami, M., & Huzeni, H. (2021). Perbaikan Algoritma Naive Bayes Classifier Menggunakan Teknik Laplacian Correction. *Jurnal Teknologi*, 21(1), 39. <https://doi.org/10.30811/teknologi.v21i1.2209>

- RIZWAN, M. R. (2022). Analisa Sentimen Terhadap Program Kartu Prakerja Pada Instagram Menggunakan Metode Naïve Bayes Classifier. *Tugas Akhir UIN*.
- Sari, V., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, 4(2), 1–9. <https://doi.org/10.29408/edumatic.v4i2.2202>
- Septian, J. A., Fahrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent Systems and Computation*, 43–49. <https://t.co/9WloaWpfD5>
- 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>
- Styawati., Hendrastuty, N., Isnain, A. R., & Rahmadhani, A. Y. (2021). Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine. *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, 6(3), 150–155. <http://situs.com>
- Suhartono, Kurniawan, F., & Imran, B. (2018). Identification of virtual plants using bayesian networks based on parametric L-system. *International Journal of Advances in Intelligent Informatics*, 4(1), 40–52. <https://doi.org/10.26555/ijain.v4i1.157>
- Syarifuddin, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. *INTI Nusa Mandiri*, 15(1), 23–28. <https://doi.org/10.33480/inti.v15i1.1347>
- Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(5), 577. <https://doi.org/10.25126/jtiik.201855983>
- Ulinuha, D. Z. (2022). *KLASIFIKASI KALIMAT PERBINCANGAN MASYARAKAT PADA MASA PANDEMI COVID-19 MENGGUNAKAN METODE K-NEAREST NEIGHBOR*. 1–23.
- Yulita, W., Dwi Nugroho, E., Habib Algifari, M., Studi Teknik Informatika, P., Teknologi Sumatera, I., Terusan Ryacudu, J., Huwi, W., Agung, J., & Selatan, L. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *Jdmsi*, 2(2), 1–9.
- Zaman, M. R. T. (2020). *KLASIFIKASI OPINI TERHADAP KEBIJAKAN PUBLIK MERDEKA BELAJAR PADA JEJARING SOSIAL TWITTER MENGGUNAKAN METODE NAÏVE BAYES DENGAN SELEKSI FITUR INFORMATION GAIN*.

