Mini Project - Pemrosesan Bahasa Alami

Nama Kelompok :

Anggota :

202210370311288 – Divani Salsabila

202210370311281 - Muhammad Saifuddin Tamam

202210370311413 - Wahyu Lukytaningtyas

Berikut ini merupakan update template laporan Mini Project kuliah Pemrosesan Bahasa Alami.

Nilai Total: 120 poin

Tahap 0 (poin: 25): Business Objective

Klasifikasi sentimen merupakan teknik dalam Natural Language Processing (NLP) yang digunakan untuk memahami dan mengklasifikasikan opini dari teks. Dalam proyek ini, kami melakukan klasifikasi sentimen dalam review buku dengan empat kategori utama: Positif, Netral, Negatif atau Others. Dataset yang digunakan dalam proyek ini diambil dari Kaggle, yang berisi kumpulan ulasan buku dari berbagai pengguna. Tujuan utama dari proyek ini adalah:

- Menganalisis sentimen dalam ulasan buku guna memahami persepsi pembaca terhadap suatu karya.
- Memberikan wawasan bagi penulis dan penerbit mengenai respons audiens terhadap buku mereka.
- Mengembangkan model klasifikasi sentimen berbasis NLP dengan akurasi yang optimal.

Tahap 1 (poin: 25): Original Data

Urgensi Topik/Kasus yang Dipilih:

Proyek ini penting karena ulasan buku seringkali bersifat subjektif dan menggunakan ragam bahasa yang variatif, mulai dari bahasa formal hingga bahasa santai, serta mengandung unsur metafora, sarkasme, dan emosi yang kompleks. Hal ini menyulitkan proses klasifikasi sentimen secara manual, terlebih ketika opini dalam satu ulasan bercampur antara positif dan negatif. Dengan menggunakan teknik Natural Language Processing (NLP), klasifikasi sentimen dapat dilakukan secara otomatis berdasarkan ciri linguistik, sintaksis, dan semantik dari teks ulasan. Melalui model yang dirancang dengan baik, sistem ini dapat mengklasifikasikan review buku dengan akurasi tinggi dan efisien tanpa perlu intervensi manusia, sehingga sangat bermanfaat dalam memahami persepsi publik terhadap suatu karya secara menyeluruh dan objektif.

Data yang Digunakan:

- Deskripsi Singkat:
  - Jenis Analisis: Klasifikasi (Supervised Learning).
  - Metode Pemodelan: Natural Language Processing (NLP) + Machine Learning.
- Atribut Data:
  - user → Nama pengguna yang memberikan review.
  - review_text → Ulasan buku dalam bentuk teks.
  - genre → Kategori buku.
  - sentiment (Label/Kelas) →
    - Positif → Review yang mengandung opini positif tentang buku.
    - Netral → Review yang bersifat deskriptif tanpa opini yang jelas.
    - Negatif → Review yang mengandung opini negatif tentang buku.
    - Others → Review yang mengandung teks yang tidak termasuk review.
  - Fitur Linguistik (Ekstraksi NLP) →
    - Panjang kalimat (jumlah kata & karakter).
    - Penggunaan kata-kata positif atau negatif.
    - Struktur kalimat (kompleks atau sederhana).
    - Word Embeddings (TF-IDF, Word2Vec, BERT).
- Data NLP Task yang Digunakan:
  - Klasifikasi → Memetakan review ke dalam kelas Positif, Netral, atau Negatif.
  - Text Preprocessing → Tokenization, stopword removal, stemming/lemmatization, dll.
  - Feature Extraction → TF-IDF, Word2Vec, BERT embeddings.
  - Model Machine Learning →
    - Traditional ML (SVM, Naïve Bayes, Random Forest).
    - Deep Learning (LSTM, Transformer, BERT).
- Sumber Data:
  - Book Review Dataset (https://www.kaggle.com/datasets/)
- Contoh Data :

genre	reviewer_name	review	Label
Romance, Contemporary, Sports, New Adult	*TANYA*	Itâ€™s me, all me. I was over this book after chapter 5, but I refuse to DNF. I couldnâ€™t wait for this book to be over.	Negatif

Young Adult, Contemporary, Lgbt, Fiction	Kat O'Keeffe	I LOVED THIS. beautiful and gripping and tragic and inspiring. I've enjoyed all of Adam's books, but this one is definitely my favorite of his (so far!) I actually listened to the audiobook for this and it was EXCELLENT. highly recommended if you're into audiobooks! if you wanna know more of my spoilery thoughts and feelings, check out the recording of the booksplosion liveshow - https://www.youtube.com/watch?v=ViS0Z...	Positif
Mystery, Fiction, Thriller, Mystery Thriller, Contemporary	Eryn	I received a copy off of NetGalley in exchange for an honest review. This does not persuade my true opinion of the novel. 2.25 So, this was quite ... strange. Very, very, strange - if I might say. At first, the story showed signs of promise, even if it seemed like all of the characters were border-line perverts (this is an exaggeration, but yes, there were just too many scenes focusing on sex - at least to me - plus, these characters are kids, so come on). Gradually as the novel progressed, it slowly began getting worse and worse - up to the point where I didn't want to pick it up at all. That brief moment of a "oh, interesting, it's a fresh perspective" from the beginning, slowly drifted away as I tired of the shallow characters and plot. Overall, I won't say more because I don't want to ruin anything for those who will read this. Regardless, I'm upset to give this such a low rating - however, because I did not enjoy myself and often found myself cringing as I was reading, I would say the rating is accurate. *Thank you to NetGalley, for giving me the opportunity to review this.* (less)	Negatif

Romance, M M Romance, Contemporary	Sandra	MINE!!! This was a fun romp. I enjoyed reading it, even if the characters are a bit of a stereotype, what with the washboard abs, the massive dicks (except for the bad guy; his dick is reportedly on the smaller side), the super-toned bods, and the fact that nearly everyone in this mostly fictional condo building just north of downtown Atlanta is gay and most of them fuck indiscriminately but safely. Wheeeee, they're young, they're hung, and they're having a good time. There is lots and lots of steamy sexy times in this book too, and it would be very easy to dismiss it as just another stereotypical erotic romance BUT the two main characters actually do struggle to make sense of their lives and themselves. They do have feelings, they feel hurt and pain and grief, joy and happiness and heartbreak, and while they don't always make the best choices, they do try to get it right. What begins as a fake relationship with hot sex turns into much more, even if there are roadblocks and fears. If you can look past the first person present tense and superficial descriptions of the hot bods all around, there is actually a sweet romance within. I enjoyed it. YMMV. (less)	Positif
Romance, Paranormal Romance, Fantasy, Paranormal, Vampires,	Alex is The Romance Fox	I struggled to finish this one.......there were a few times that I almost gave up...and the only reason I turned the last page was that I kept thinking it would get better. There was something missing between the two mc's....I never felt a chemistry or connection there. I love the IAD Series but I am not sure	Netral

M M Romance		about The Dacians. However, I do love KC so I am not giving up on the Immortals now!!	
Fantasy, Young Adult, Did Not Finish, Magic, Fiction	Montzalee Wittmann	The Waking Land by Callie Bates is a book I was allowed to read from NetGalley and I am so glad indeed! This book is so rich in fantasy, world building, character depth, plot, and twists that I was totally enthralled in its wonders. Elanna has earth magic, more than anyone knows, in a time when magic is forbidden except in the far north where the land shifts and protects the people, the old ways, and the land itself. Elanna is forced/kidnapped from her family as a young child and is held by the King to make her father be submissive to the King. She is raised by the King and told so many lies about her birth land that she believes them. Then, the King is murdered and she is blamed. The daughter of the King is now Queen and has always hated her. Elanna meets a man that also knows magic and knows about hers but he wants her to go to her real father and she believes all the lies. It is so action packed, so many twists, so much magic mixed in there, so much emotion...I am not doing it justice here and only touching on the tip of the iceberg. The Queen's men are after her, the witch hunters are after her...She has the power of the land, earth, and things of the earth. She will need all of these if she can get the land to wake up. It is so exciting to see how and when and with who ....so excellent!!! (less)	Positif
Science Fiction, Fiction, Fantasy, Novella	Kaitlin	I have to say I'm really sad I don't love this series as much as some people do. I remember reading the first one in the series a while back and liking it, but still not getting the hype a lot of my friends were	Negatif

		giving it. Going back into Binti's world, I still don't get the hype. What I do like about this series is that it's showing a female character who I think is pretty great and stands out from her Himba people in positive ways. She is trying to change things and show that you don't have to just settle, you can go for your dreams, and I think that's a great message. Alongside the human character of Binti we have the alien characters in the book who show something different, like the un-gendered Meduse, and I think that's really nice to see in more books and is definitely a step in the right direction for Sci Fi as a whole. Unfortunately, for me, plot and characters are really key to my enjoyment of a book and this storyline just didn't capture my attention the way I wanted. I think Binti is doing good things, but I just find her a bit of a boring character to read about even so, and I find her really hard to connect to which is a shame. I really do wish the story worked more for me, but I just can't help but think that in 160+ pages we actually don't do a lot, and I felt that I didn't really feel any intense emotion at all through the book. Clearly, Okorafor's writing is modern and challenging at times, but for me it just isn't working out (I've previously also read Lagoon by her). I would give this a 2*s overall, it's okay, but I can't see why there's all the love this series gets, but maybe that's just me :) (less)	
Did Not Finish	Carole Sherriff		others

Nonfiction, Autobiography, Memoir, War, Military Fiction, Biography, Feminism, Biography Memoir	Jeanette	This is MJ's story from her own voice. She's over the top enthusiastic and an adrenaline junkie. But she also cries buckets at times. Intense highs and ferocious lows. But I hear the response here- "hey, that isn't fair". Yet it is true and because I do know she is quite a hero, that also makes her a natural for the work. Not only for her war flights and Taliban attack wounds events, but also with her honest and continuing attitude. Which encourages presently with petitioning legally within all the current organizational levels to get women into combat. Officially. I'm not sure I agree with her on specific points at all, but I do know she knows what she likes and what paths she traipsed to get her Medvac pilot career. And that discrimination against women in the military (especially upon the personal sexual assault reality) continues to exist. The writing had lacks and enough foul language that she could be a sailor. After dialogue response of expletive after expletive! It all sounds so dumb. But that's how they talk? Well, I for one, found that the terrible incident of being knocked out of air space and being lost with nearly no fuel remaining because of the Texas A&M game glitch while she was just learning and didn't even know the navigation equipment or maps yet at ALL- that was just as brave as the others for the 3 tours in Afghanistan. Many would not have continued. Often I wonder when I hear about celebs or politicos calling off for "spaces" when highways, tollways, airports or spaces adjacent get "shut down" what fall outs actually do occur for the rest of us. MANY! And not all of them are safe nor	Netral

		easy alternatives. MJ is certainly herself. And she has at a very young age, all the scars and redone joints to prove it. It's part of the process for those who love the adrenaline rush mode of living and push the physical. Never old bones intact. The photos were 4 star and her story is worth the read. Hard row to hoe and I hope her future includes some measure of contentment. (less)	
Romance, Contemporary, Erotica, Menage, Romance, Adult Fiction, Erotica	Isabel Love	Unconventional is now available on AUDIO and on sale for the first time ever! You need Charlie and Quinn in your ears (but grab your earbuds first)!! ðŸ“˜ Amazonâžœ mybook.to/UnconventionalKindle ðŸŽ§ Audible âžœ http://bit.ly/UnconventionalAudible ðŸŽ§ Tantor âžœ http://bit.ly/UnconventionalTantor ðŸŽ§ iTunes âžœ http://bit.ly/UnconventionaliTunes ðŸŽ§ Audiobooks.com âžœ http://bit.ly/UnconventionalAudiobook... ðŸŽ§ Google Play âžœ http://bit.ly/UnconventionalGooglePlay ðŸŽ§ RB Digital âžœ http://bit.ly/UnconventionalRBDigital	others

Tabel 1. Sample ulasan buku yang sudah dilabeli secara manual

Tahap 2 (poin: 10): Target Data (Optional)

Target data dalam klasifikasi ini adalah Sentimen Review Buku, yaitu apakah suatu ulasan bersifat positif, netral, atau negatif. Sentimen ini mencerminkan persepsi pembaca terhadap buku berdasarkan ulasan teks yang mereka berikan.
Untuk menentukan label sentimen, kita dapat menggunakan pendekatan berikut:

- "**Positif**" jika review menunjukkan kepuasan atau apresiasi terhadap buku.

- "**Netral**" jika review tidak terlalu memihak, bersifat deskriptif tanpa emosi yang kuat.
- "**Negatif**" jika review berisi kritik, kekecewaan, atau ketidakpuasan terhadap buku.
- "**Others**" Teks yang tidak termasuk review atau tidak memberikan penilaian terhadap isi buku, seperti spam, komentar tidak relevan, pertanyaan, atau sangat singkat/ambigu.

● **Atribut yang Digunakan**

Beberapa atribut dalam dataset memiliki potensi mempengaruhi analisis sentimen. Atribut yang digunakan meliputi:

Atribut	Alasan Dipilih
Review	Teks ulasan dari pengguna yang menjadi sumber utama analisis sentimen. Akan diproses dengan Natural Language Processing (NLP) untuk mengidentifikasi sentimen.
Genre	Bisa membantu dalam analisis tren sentimen berdasarkan kategori buku tertentu.
Reviewer Name	Bisa digunakan untuk melihat apakah ada pola sentimen berdasarkan reviewer tertentu.

Tabel 2.1. Atribut yang digunakan dalam klasifikasi

● **Atribut yang Tidak Digunakan**

Beberapa atribut tidak memiliki pengaruh langsung terhadap analisis sentimen atau bersifat metadata yang tidak relevan untuk model klasifikasi. Atribut yang diabaikan meliputi:

Atribut	Alasan Tidak Digunakan

book_title, Book_series, book_series_url	Hanya informasi identifikasi buku, tidak berpengaruh pada sentimen.
book_image, book_image_url	Data visual yang tidak relevan untuk NLP.
book_author, author_url	Tidak memiliki dampak langsung terhadap sentimen review.
book_rating,	Beberapa rating yang diberikan tidak sesuai dengan isi review
reviewer_image, reviewer_image_url	Tidak berkontribusi terhadap analisis teks.
ID	Hanya sebagai penomoran, tidak relevan untuk prediksi.

Tabel 2.2. Atribut yang tidak digunakan dalam klasifikasi

Dengan memilih atribut yang relevan, kita dapat meningkatkan efektivitas model dalam mengklasifikasikan sentimen review buku dengan lebih akurat.

Tahap 3-4 (poin: 25): Data Pre-processing & Transformation

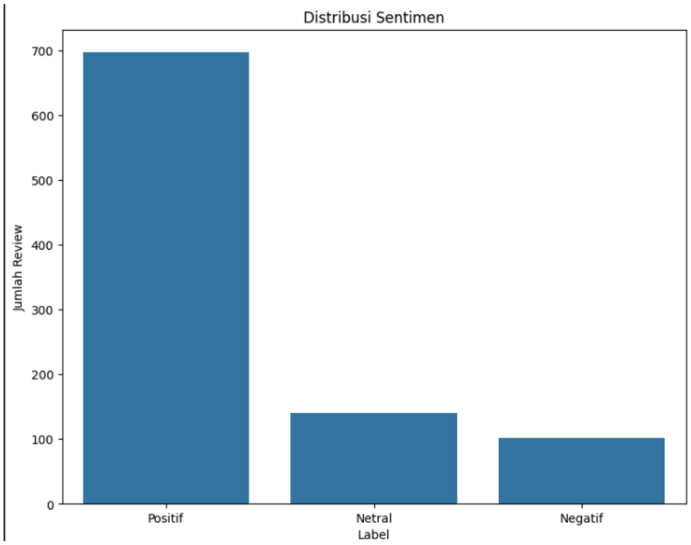
- **Exploratory Data Analysis**.
    1. Distribusi Sentimen
    Kami mengklasifikasikan setiap ulasan ke dalam tiga kategori sentimen: positif, netral, dan negatif menggunakan pelabelan secara manual.

        - Jumlah Review Positif: 697
        - Jumlah Review Netral: 140
        - Jumlah Review Negatif: 102
        - Jumlah Review Others : 16

        Kesimpulan: Distribusi sentimen terlihat tidak seimbang), di mana sentimen positif mendominasi.

2. Panjang Teks Review
   Kami menghitung rata-rata jumlah kata dan karakter dari setiap review berdasarkan kategori sentimen.

   - **Rata-rata jumlah kata**:
     1. Positif: 331 kata
     2. Netral: 264 kata
     3. Negatif: 310 kata
     4. Others : 40
   - **Rata-rata jumlah karakter**:
     1. Positif: 1868 karakter
     2. Netral: 1462 karakter
     3. Negatif: 1762 karakter



Gambar 3.1 Distribusi Sentimen Review

3. Top Words
   Kami membandingkan kata-kata yang paling sering muncul dalam review **positif** dan **negatif**.

   1. **Top Words Review Positif**: Banyak mengandung kata-kata seperti *love*, *amazing*, *beautiful*, dll.
   2. **Top Words Review Negatif**: Sering muncul kata seperti *boring*, *disappointed*, *bad*, dll.

   Kesimpulan: Kata-kata positif menunjukkan kepuasan emosional, sedangkan kata negatif cenderung mengungkapkan kekecewaan terhadap plot atau karakter.

4. Genre vs Sentimen

Kami mengeksplorasi hubungan antara genre dan distribusi sentimen review.

- Genre dengan review positif terbanyak: [Fiction, Romance, Contemporary]
- Genre dengan review negatif terbanyak: [Fiction, Romance, Contemporary]

Kesimpulan: Genre (**Fiction, Romance, Contemporary**) adalah yang paling banyak menerima review negatif, tapi juga paling banyak review positif. Jadi jumlahnya tinggi karena genre ini paling umum.

- **Data Pre-processing & Transformation**

- Data Pre-processing

Beberapa teknik yang bisa digunakan yaitu :

- Data Cleaning (Pembersihan Data) :
    - Menghapus nilai yang hilang (missing values).
    - Menghapus karakter atau simbol yang tidak diperlukan dalam teks review.
- Data Transformation (Transformasi Data)
    - Mengubah teks menjadi huruf kecil.
    - Menghapus tanda baca, angka, dan karakter khusus.
    - Menghapus stopwords (kata umum yang tidak memiliki makna penting).
    - Melakukan stemming (mengubah kata ke bentuk dasar).
- Data Reduction (Opsional)
    - Karena dataset terlalu besar, Kita mengambil 1000 data dengan menggunakan random sampling
- Feature Selection (Pemilihan Fitur)
    - Memastikan fitur yang digunakan adalah yang relevan untuk analisis sentimen.

Tahap 5 (poin: 25): Data Mining

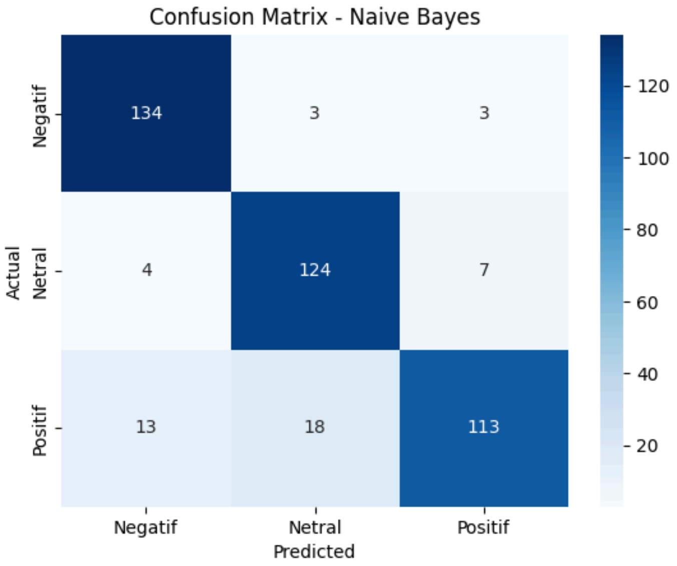
Algoritma NLP yang Digunakan

Untuk klasifikasi sentimen review buku menjadi **Positif**, **Negatif**, **Netral** dan **Others**, dilakukan preprocessing teks menggunakan:

- **TF-IDF Vectorization**: Untuk mengubah teks review menjadi representasi numerik berbasis frekuensi kata

Skenario Eksperimen

- **Dataset**: Diambil dari file selected_reviews - Label , terdiri atas kolom review dan label.
- **Split Data**: 80% untuk pelatihan dan 20% untuk pengujian.
- **Evaluasi Performa**: Digunakan metrik *classification report* dan *confusion matrix*.
- **Model Klasifikasi**:
    - Naive Bayes

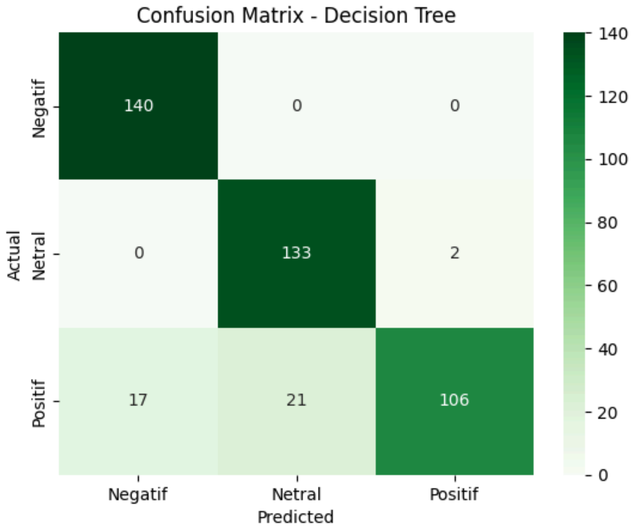


Gambar 5.1 Visualisasi Hasil Naive Bayes

Label	Precision	recall	f1-score	Support
Negatif	0.89	0.96	0.92	140
Netral	0.86	0.92	0.89	135
Positif	0.92	0.78	0.85	144

Tabel 5.1 Hasil classification report Naive Bayes

○ Decision Tree



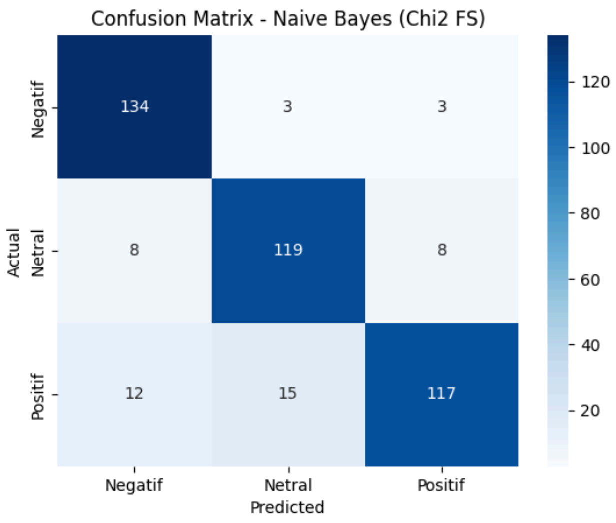
○

Gambar 5.2 Visualisasi Hasil Decision Tree

Label	Precision	recall	f1-score	Support
Negatif	0.89	1.00	0.94	140
Netral	0.86	0.99	0.92	135
Positif	0.98	0.74	0.84	1144

Tabel 5.2 Hasil classification report decision tree

○ Naive Bayes menggunakan Feature selection Chi2



Gambar 5.3 Visualisasi Hasil Naive Bayes Chi2

Label	Precision	recall	f1-score	Support
Negatif	0.87	0.96	0.91	140
Netral	0.87	0.88	0.88	135
Positif	0.91	0.81	0.86	144

Tabel 5.3 Hasil classification report Logistic Regression

Model	Akurasi	Precision	Recall	F1-Score
Naive Bayes	88.5%	0.89	0.89	0.88
Decision Tree	90.4%	0.91	0.91	0.90
Naive Bayes Chi2	88.3%	0.88	0.88	0.88

Gambar 5.4 Rangkuman hasil classification report metode klasifikasi

Tahap 6 (poin: 20): Knowledge Interpretation

- Pola-pola *useful* yang telah ditemukan.

Tahap 7 (poin: 15): Reporting

- Academic Poster.
- Report.
- Notebook (Google Colab.)
- Dashboard