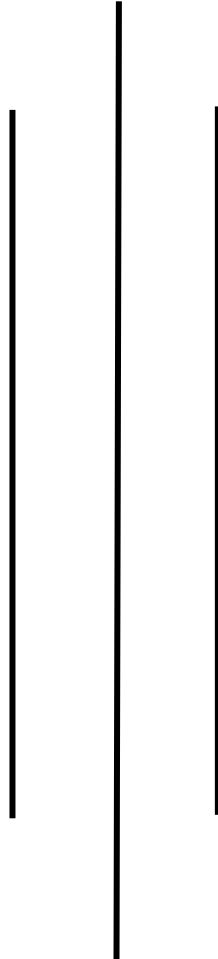


MAKALAH
ANALISIS DATA TWEET EMOTION
DAN MEMPREDIKSI 1000 LABEL KOSONG



Disusun oleh:

1. Deden Tri Aditya Kurniawan
2. Destiana Talia
3. Halidza Anya Amalia
4. Maulana Ibnu Sahban

DAFTAR ISI

A. Abstrak.....	3
BAB I	4
PENDAHULUAN.....	4
Latar Belakang.....	4
Rumusan Masalah.....	4
Tujuan dan Manfaat	5
BAB II	6
I. Landasan Teori	6
II. Metode Penelitian.....	9
B. Metode.....	10
C . Dataset.....	12
D. Langkah.....	13
III. Hasil dan Pembahasan	19
BAB III.....	21
PENUTUP	21
DAFTAR PUSTAKA.....	22

A. Abstrak

Dalam era digital yang semakin berkembang, *Twitter* telah menjadi *platform* yang penuh dengan data berharga dalam berbagai topik dan konteks. Penelitian ini membahas analisis data *tweet* menggunakan teknik *data mining* untuk menggali wawasan yang berharga dari informasi yang tersebar di *Twitter*. Makalah ini merinci langkah-langkah metodologi, termasuk pengumpulan data, pemrosesan, dan analisis statistik.

Hasil analisis data *tweet* memberikan wawasan tentang tren, sentimen, dan interaksi pengguna *Twitter* dalam topik yang diteliti. Analisis sentimen mengidentifikasi emosi seperti *love* (suka), *fear* (takut), *joy* (sukacita), *sadness* (sedih), dan *anger* (amarah) dalam *tweet*. Analisis tren memahami perkembangan topik seiring waktu, sedangkan interaksi pengguna seperti *retweet* dan *like* mengukur tingkat perhatian terhadap suatu topik.

Makalah ini juga mengeksplorasi aplikasi hasil analisis, seperti penggunaan dalam riset pasar untuk memahami preferensi pelanggan dan reaksi terhadap produk atau layanan. Analisis sentimen juga berguna dalam pemantauan opini publik, seperti dalam pemilihan umum atau kejadian berita penting.

Temuan dan rekomendasi penelitian ini memberikan panduan bagi peneliti dan praktisi yang tertarik dalam menggali informasi dari data *tweet* dalam era digital yang dinamis. Integrasi analisis data *tweet* dalam strategi bisnis atau riset dapat memanfaatkan potensi besar yang ditawarkan oleh media sosial untuk memahami tren dan pandangan masyarakat dengan lebih baik. Dengan demikian, penelitian ini membuka peluang untuk memanfaatkan kekayaan data yang ada di *Twitter* agar mendapatkan wawasan yang berharga dalam berbagai konteks.

BAB I

PENDAHULUAN

Latar Belakang

Dalam era digital yang terus berkembang, media sosial, termasuk *Twitter*, telah menjadi sumber data yang kaya akan informasi dari berbagai aspek kehidupan manusia. Jutaan *tweet* diposting setiap hari di *Twitter*, mencakup berbagai topik, termasuk berita, tren, opini, dan emosi pengguna. Analisis data *tweet* menjadi semakin penting karena potensi informasi yang dapat diekstrak dari *platform* ini.

Namun, analisis data *tweet* tidak selalu sederhana. *Tweet* seringkali singkat, penuh dengan bahasa gaul, akronim, dan slang (bahasa yang tidak resmi atau tidak formal), sehingga pemahaman konten dan sentimen yang terkandung dalam *tweet* menjadi tugas yang menantang. Oleh karena itu, permasalahan utama yang muncul adalah bagaimana kita dapat mengembangkan metode atau model yang efektif untuk memprediksi konten dan *emoticon tweet* dengan akurat.

Rumusan Masalah

Dalam konteks ini, rumusan masalah utama penelitian adalah: Bagaimana kita dapat mengembangkan model atau algoritma yang dapat memprediksi konten dan sentimen *tweet* dengan akurat? Bagaimana kita dapat mengatasi berbagai hambatan yang sering muncul dalam analisis data *tweet*, seperti bahasa yang tidak formal, variasi ejaan, dan perubahan tren bahasa di media sosial?

Penelitian ini bertujuan untuk mencari solusi yang dapat meningkatkan pemahaman dan analisis *tweet* dengan lebih baik. Hal ini

melibatkan pengembangan alat atau model yang dapat mengidentifikasi dan mengklasifikasikan konten dan sentimen *tweet*, yang dapat digunakan dalam berbagai konteks, seperti pemantauan opini publik, analisis tren, atau bahkan untuk membantu bisnis dalam memahami respon pelanggan.

Dengan menjawab pertanyaan-pertanyaan ini, penelitian ini diharapkan akan memberikan sumbangan berharga dalam pemahaman dan analisis data *tweet* di era digital yang terus berubah. Dengan kemajuan dalam pemahaman dan prediksi data *tweet*, kita dapat mengambil manfaat maksimal dari potensi data yang ada di *Twitter* untuk berbagai keperluan analisis dan pengambilan keputusan.

Tujuan dan Manfaat

Penelitian ini bertujuan untuk mengembangkan model atau algoritma yang mampu memprediksi konten dan sentimen *tweet* dengan tingkat akurasi yang tinggi. Seiring dengan perkembangan pesat era digital, *Twitter* telah menjadi sumber data yang sangat berharga, tetapi juga penuh dengan tantangan seperti bahasa yang tidak formal, variasi ejaan, dan perubahan tren bahasa. Dalam rangka mengatasi hambatan tersebut, penelitian ini berusaha meningkatkan pemahaman dan analisis data *tweet* dengan merinci langkah-langkah metodologi, termasuk pengumpulan data, pemrosesan, dan analisis statistik. Selain itu, penelitian ini juga membahas aplikasi hasil analisis seperti penggunaan dalam riset pasar, pemantauan opini publik, dan analisis tren untuk membantu bisnis memahami preferensi pelanggan, reaksi terhadap produk, serta mendukung pengambilan keputusan yang lebih baik.

Manfaat dari penelitian ini tidak hanya terbatas pada konteks akademis, melainkan juga berpotensi memberikan dampak positif dalam berbagai aspek kehidupan dan bisnis. Hasil analisis data *tweet* dapat

memberikan wawasan yang berharga kepada peneliti, praktisi, dan bisnis dalam memahami dinamika media sosial dan respons pengguna dalam era digital yang terus berkembang. Dengan integrasi analisis data *tweet* ke dalam strategi bisnis atau riset, kita dapat memanfaatkan potensi besar yang ditawarkan oleh *platform* media sosial seperti *Twitter* untuk memahami tren, pandangan masyarakat, serta meningkatkan kualitas pengambilan keputusan.

BAB II

PEMBAHASAN

I.Landasan Teori

a. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. Data mining, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Ridwan et al.,2013). Dalam Data Mining ada tiga karakteristik data (Leslie, Spits, Lumban, Trisetyarso, & Abdurachman, 2017), yaitu:

1. *Supervised*, adalah variabel atau data yang berlabel.
2. *Semi Supervised*, adalah variabel atau data yang beberapa berlabel dan beberapa tidak berlabel.
3. *Unsupervised*, adalah variabel yang tidak berlabel.

b. Analisis Sentimen

Analisis sentimen sendiri atau juga biasa disebut dengan opinion mining adalah salah satu 16 bagian dari *text mining*. Bidang ini melakukan studi mengenai opini orang-orang, sentimen, evaluasi, tingkah laku dan emosi terhadap suatu entitas seperti produk, layanan, organisasi, individu, permasalahan, topik, acara dan atribut-atributnya. Analisis sentimen sangatlah berguna untuk menganalisis komentar-komentar di *Twitter* tadi untuk kemudian diterjemahkan menjadi sesuatu yang lebih bermakna, salah satunya dalam bentuk rating. Dalam dunia bisnis rating menjadi sangat penting karena merupakan salah satu indikator kesuksesan. Di sisi lain, rating masih menjadi komoditas monopoli beberapa perusahaan seperti Nielsen, sehingga objektivitasnya menjadi kurang. Celah inilah yang kemudian dimanfaatkan penulis untuk mencoba mengaplikasikan analisis sentimen pada *Twitter* untuk membuat sistem rating berdasar komentar. (Monarizqa, Nugroho, & Hantono, 2014).

c. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas yang tidak diketahui dari suatu objek. Dalam klasifikasi terdapat dua proses, yaitu proses *training* dan proses *testing* (Bertalya, 2009). Pada proses *training* menggunakan *training set* yang telah diketahui label-labelnya untuk membangun model. Kemudian 17 proses *testing* untuk menguji keakuratan model yang telah dibangun saat proses *training*.

d. Twitter

Twitter merupakan sebuah media sosial yang memberikan layanan dari *microblogging* untuk memberikan fasilitas terhadap pengguna dalam mengirim dan membaca pesan dalam berupa *tweets*. *Microblogging* merupakan sebuah layanan berbasis web dimana penggunanya dapat menulis status dalam berupa teks, mengunggah gambar atau video secara *online* dan *real time*. *Tweet* merupakan sebuah teks tulisan yang memiliki batasan mencapai 140 karakter. Pengguna yang menuliskan status kedalam *tweets* dapat dilihat secara publik, namun juga dapat mengirim pesan melalui daftar *followers* mereka saja (*direct message*).

e. Pre-Processing

Preprocessing adalah salah satu langkah terpenting dari *Data Mining*. *Preprocessing* dilakukan untuk mendapatkan data yang akurat. Dalam *preprocessing* teks, ada banyak langkah seperti *Case Folding*, *Data Cleansing*, menghapus *stopwords*, *stemming* (Sharma, Agrawal, Lalit, & Garg, 2017).

1. *Case Folding* adalah proses dimana mengubah semua karakter pada teks menjadi huruf kecil dan menghilangkan angka atau bentuk tanda baca sehingga data yang didapat hanya mengandung karakter huruf a sampai z.
2. *Data Cleaning* adalah proses membersihkan *tweet* dari kata yang tidak diperlukan atau untuk mengurangi *noise*.
3. Penghapusan *Stopwords* adalah proses menghilangkan kata-kata yang sering muncul tapi tidak memiliki makna dalam klasifikasi.

4. *Stemming* adalah proses menyederhanakan kata yang berisi imbuhan kembali ke kata dasarnya.

II. Metode Penelitian

A. Software

RAPID MINER: DQLAB

RapidMiner sebelumnya bernama YALE (*Yet Another Learning Environment*), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit dari University of Dortmund*. *RapidMiner* merupakan salah satu *tools* yang dipakai dalam data *mining*. *RapidMiner* memiliki kurang lebih 500 operator data *mining*, termasuk operator untuk *input*, *output*, data *preprocessing* dan visualisasi.

RapidMiner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data *mining* yang dapat diintegrasikan pada produknya sendiri. *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap data *mining*, *text mining* dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik.

VISUAL STUDIO CODE(PYTHON): NIAGAHOST blog

Visual Studio Code adalah sebuah kode editor gratis yang bisa dijalankan di perangkat desktop berbasis Windows, Linux, dan MacOS. Kode editor ini dikembangkan oleh salah satu raksasa teknologi dunia, Microsoft. Visual Code adalah *software* editor yang powerfull, tapi tetap ringan ketika digunakan. Ia bisa dipakai untuk membuat dan mengedit *source code* berbagai bahasa pemrograman salah satunya adalah Python.

MICROSOFT POWER BI: IDMETAFORA

Microsoft Power BI adalah perangkat lunak intelijen bisnis yang dikembangkan oleh Microsoft yang memungkinkan Anda untuk memproses data Anda secara lebih rinci dan menyajikannya dengan cara yang jelas dan interaktif. Aplikasi ini memungkinkan Anda untuk memvisualisasikan data yang dimasukkan atau terhubung dari sistem pihak ketiga. Anda dapat dengan mudah mengontrol dan memantau data Anda. Microsoft Power BI memiliki dua *platform*: desktop, yang dapat diinstal pada komputer atau laptop Anda, dan perangkat seluler, yang dapat diinstal pada ponsel Anda berdasarkan teknologi *cloud*. Aplikasi Power BI menggunakan sistem operasi seperti Windows, iOS, dan Android.

B. Metode

Jurnal "*An empirical study of the naive Bayes classifier*" (Rish, 2001) melakukan studi empiris yang meyakinkan kinerja algoritma Naive Bayes di berbagai domain, termasuk pemrosesan teks. Mereka menyimpulkan bahwa *Naive Bayes* adalah salah satu metode yang paling cepat dan memiliki akurasi yang cukup baik dalam klasifikasi teks.

Jika melihat dari sumber-sumber yang telah disebutkan, algoritma *Naive Bayes* secara umum direkomendasikan untuk pemrosesan teks. Berikut adalah beberapa argumen yang mendukung rekomendasi tersebut:

1. Buku "*Introduction to Information Retrieval*" (Manning, Raghavan, & Schütze, 2008) memberikan pemahaman dasar tentang algoritma Naive Bayes dan mencantumkan contoh implementasi serta evaluasi kinerja. Dalam buku ini, *Naive Bayes*

diperkenalkan sebagai salah satu teknik klasifikasi teks yang cukup sederhana namun efektif.

2. Jurnal "*An empirical study of the naive Bayes classifier*" (Rish, 2001) melakukan studi empiris yang meyakini kinerja algoritma *Naive Bayes* di berbagai domain, termasuk pemrosesan teks. Mereka menyimpulkan bahwa *Naive Bayes* adalah salah satu metode yang paling cepat dan memiliki akurasi yang cukup baik dalam klasifikasi teks.
3. Jurnal "*Naive Bayes Text Classification Algorithm for Opinion Mining*" (Rajkumar & Ganesan, 2015) fokus pada penggunaan *Naive Bayes* untuk analisis sentimen pada teks. Mereka mengimplementasikan algoritma ini dan memberikan contoh kasus untuk memahami konsep secara praktis.
4. Jurnal "*The Optimality of Naive Bayes*" (Zhang, 2004) menyelidiki keunggulan dan kelemahan *Naive Bayes* dalam pemrosesan teks. Mereka membahas asumsi yang dibuat oleh algoritma ini, serta mempelajari optimasi dan modifikasi yang dapat meningkatkan kinerjanya.

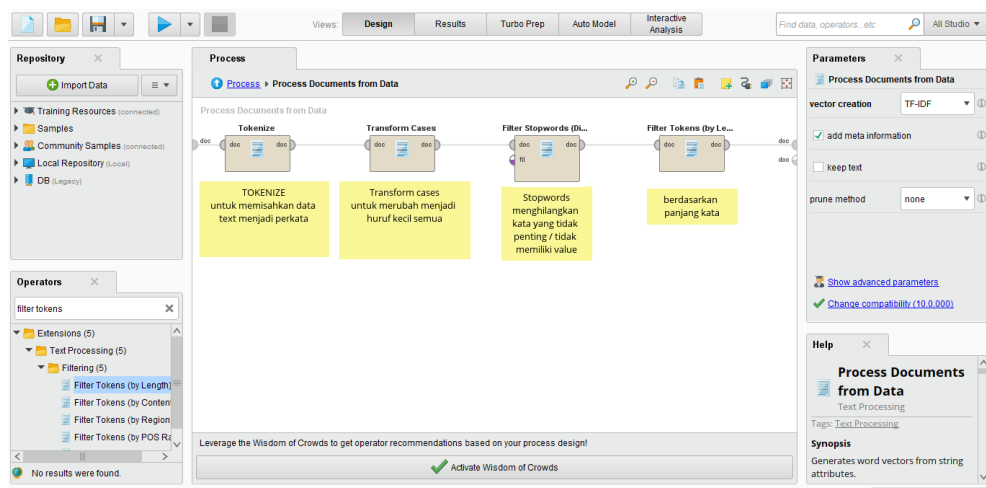
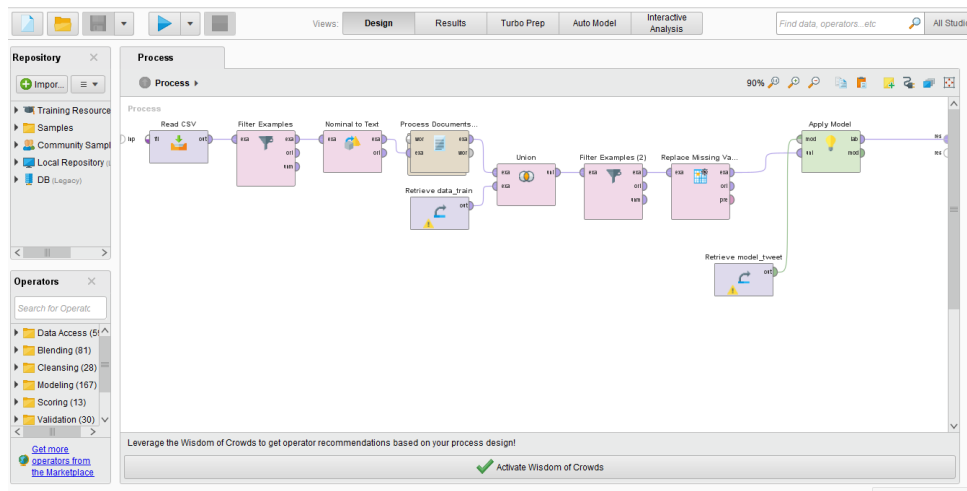
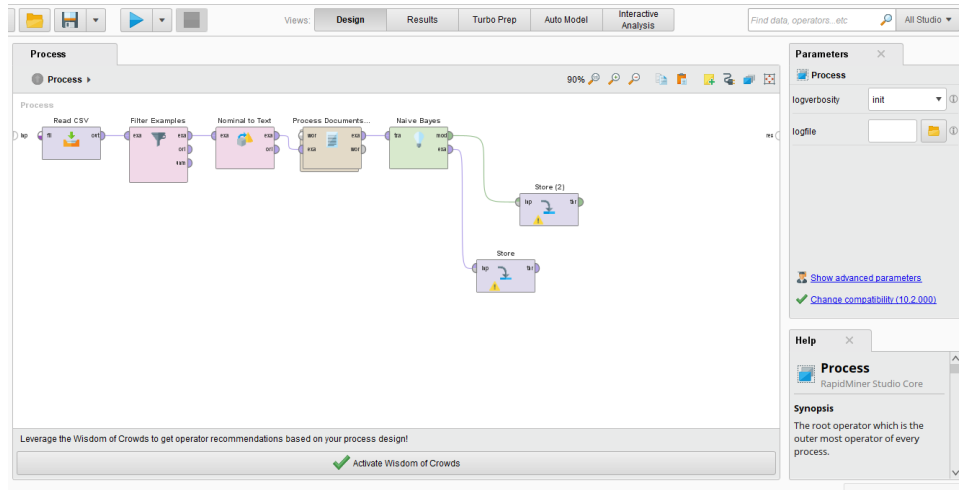
Secara keseluruhan, algoritma *Naive Bayes* direkomendasikan untuk pemrosesan teks karena kesederhanaannya, kecepatannya, dan kinerjanya yang cukup baik dalam klasifikasi. Namun, tetap penting untuk mempertimbangkan karakteristik dan kompleksitas dataset serta tujuan spesifik yang ingin dicapai dalam proyek pemrosesan teks.

C . Dataset

Dataset yang di teliti merupakan Dataset tweet merupakan kumpulan data yang berasal dari platform media sosial, terutama Twitter, dan terdiri dari sejumlah besar tweet. Data ini mencakup tweet-tweet dari berbagai pengguna yang mencakup beragam topik, opini, serta interaksi antarpengguna. Dataset tweet biasanya diperoleh melalui API Twitter atau layanan pihak ketiga yang menyediakan akses ke data tweet dan disimpan dalam format seperti CSV atau JSON. Metadata seperti waktu posting, username pengguna, jumlah retweet, dan jumlah suka sering dilampirkan untuk menganalisis interaksi pengguna dan tren seiring waktu. Isi sebenarnya dari tweet adalah elemen kunci yang mencakup teks pendek yang mengandung topik tertentu, komentar, opini, atau bahkan tautan ke sumber eksternal. Dataset ini digunakan untuk berbagai tujuan analisis, termasuk analisis sentimen, pemahaman opini publik, identifikasi trend, atau pemantauan peristiwa penting. Namun, dataset tweet memiliki tantangan unik, seperti bahasa gaul dan singkatan yang sering digunakan dalam tweet, serta keterbatasan panjang karakter yang dapat mempengaruhi pemrosesan data. Meskipun demikian, dataset tweet tetap menjadi sumber data yang berharga untuk memahami perilaku dan pandangan pengguna di era media sosial yang dinamis.

D. Langkah

➤ RapidMiner



1. *Read CSV*

Read CSV adalah proses membaca dan menguraikan data yang disimpan dalam format teks yang menggunakan tanda koma, pada proses ini dapat memungkinkan anda untuk mengkonversi data menjadi struktur yang dapat diolah.

2. *Filter Example*

Filter Example kami gunakan untuk menghilangkan baris yang memiliki *missing label*.

3. *Nominal To Text*

Nominal To Text berfungsi untuk mengubah jenis atribut nominal yang dipilih menjadi teks. Operator ini juga memetakan semua nilai dari atribut ini ke nilai *string* yang sesuai.

4. *Process Documents From Data*

Process Documents From Data adalah Langkah awal yang dilakukan saat ingin menggali dan menganalisis informasi dari text.

- ***Tokenize***

Tokenize digunakan untuk memisahkan data teks menjadi perkata.

- ***Transform Cases (Lower Case)***

Transform Cases (Lower case) berfungsi untuk merubah semua huruf dalam teks menjadi huruf kecil.

- ***Filter Stopwords (Dictionary)***

Filter Stopwords (Dictionary) berfungsi untuk menghilangkan kata penghubung dari teks di dalam data.

- ***Filter tokens (By Length)***

Filter tokens (By Length) berfungsi untuk menghapus

token dalam teks berdasarkan panjangnya. Token adalah unit kecil yang menggambarkan kata atau frasa dalam teks.

5. *Naive Bayes Operators*

Naive Bayes Operators berfungsi untuk membangun model klasifikasi sentimen yang digunakan untuk memprediksi sentimen dari teks atau ulasan.

6. *Store*

Store berfungsi untuk menyimpan hasil data yang telah diolah sehingga dapat digunakan untuk proses selanjutnya .

7. *Union*

Union berfungsi untuk menggabungkan dua atau lebih dataset menjadi satu dataset, pada tahap ini kita harus melakukan proses *Filter Example* dengan menghilangkan baris yang memiliki label, sehingga hanya tersisa 1000 kolom yang memiliki *missing label*, setelah itu kita lakukan proses *Nominal To Text* dan *Process Documents From Data* setelah itu kita munculkan data latih yang sudah kita simpan dalam store untuk digabungkan dengan data yang sudah di filter atau data yang berisi 1000 *missing label* dan jalankan .

8. *Filter Example (2)*

Setelah dijalankan maka akan ada kolom kata yang tidak memiliki *value* dan harus kita lakukan *Filter Example (2)* yang berfungsi untuk menghilangkan kata yang tidak memiliki *value* tersebut.

9. *Replace Missing Values*

Replace Missing Values kita gunakan untuk mengganti kolom yang kosong supaya bernilai 0 dengan cara mengganti menu *default* menjadi *zero*.

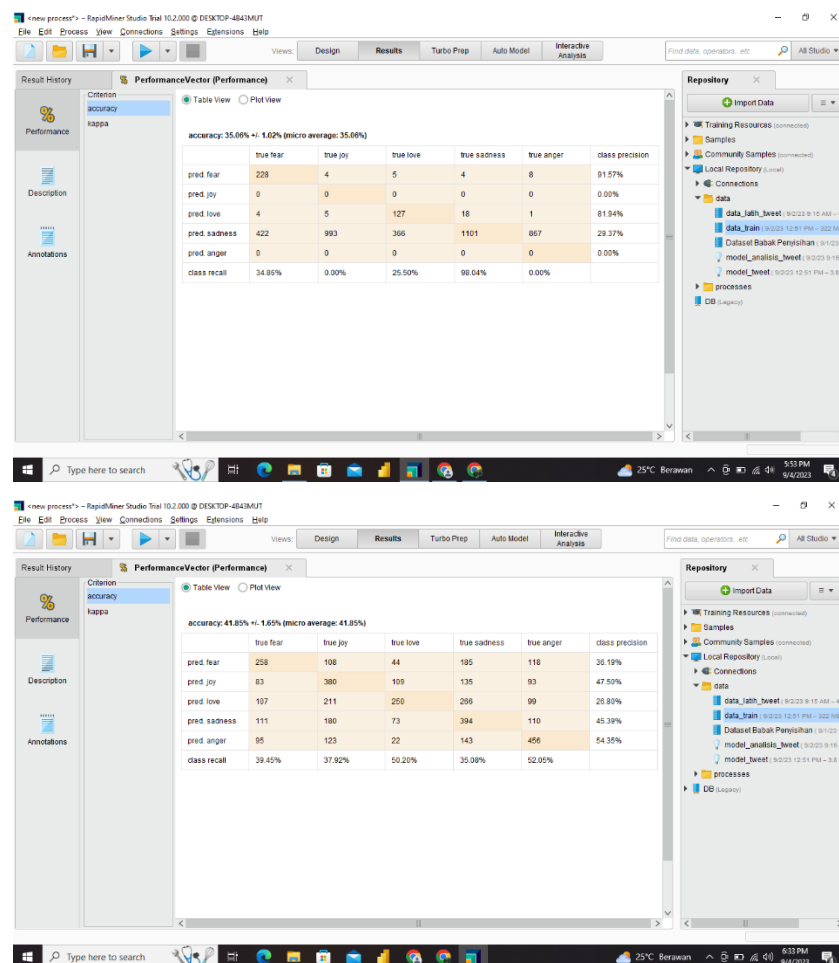
10. Apply Model

Apply Model berfungsi untuk menerapkan model yang telah dibangun atau dilatih, fungsi utamanya adalah melakukan prediksi atau klasifikasi menggunakan model yang telah ada. Kita hanya perlu munculkan model yang telah kita simpan dalam *store*.

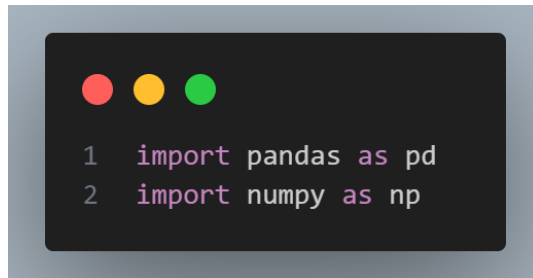
11. Hasil

Hasil yang dikeluarkan dapat kita download dengan menggunakan *Operators Write CSV* dan setelah dijalankan maka secara otomatis akan tersimpan di komputer.

12. Perbandingan Data Latih Dengan *cross validation operators* Menggunakan *Naive Bayes* dan *Decision Tree*



➤ Visual Studio Code

A screenshot of a Visual Studio Code editor window with a dark theme. At the top, there are three colored window control buttons: red, yellow, and green. Below them, the code editor shows two lines of Python code:

```
1 import pandas as pd
2 import numpy as np
```

1. *Import Library*

Import Library untuk memasukkan modul atau pustaka.

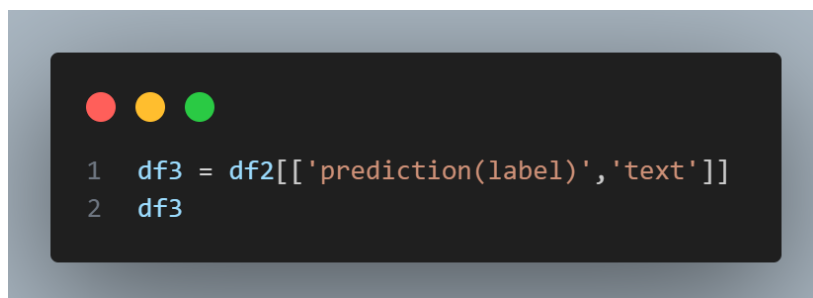
Modul adalah kumpulan fungsi, kelas, dan variabel yang dapat digunakan untuk menjalankan tugas tertentu.

A screenshot of a Visual Studio Code editor window with a dark theme. At the top, there are three colored window control buttons: red, yellow, and green. Below them, the code editor shows three lines of Python code:

```
1 df1 = pd.read_csv('file_no_missing.csv')
2 df2 = pd.read_csv('final_data.csv')
3
```

2. *Load Data*

Load Data adalah proses untuk memunculkan data yang akan kita olah. data pertama yang kita gunakan adalah file “no missing” yaitu data *tweet* yang tidak memiliki *missing value*, lalu data kedua adalah hasil prediksi yang kita lakukan di RapidMiner.

A screenshot of a Visual Studio Code editor window with a dark theme. At the top, there are three colored window control buttons: red, yellow, and green. Below them, the code editor shows two lines of Python code:

```
1 df3 = df2[['prediction(label)', 'text']]
2 df3
```

3. *Ganti Nama Kolom*

Ganti nama kolom untuk menyamakan dengan file aslinya. Kita buat variabel baru untuk data prediksi yang hanya mengambil kolom prediksi dan teks.

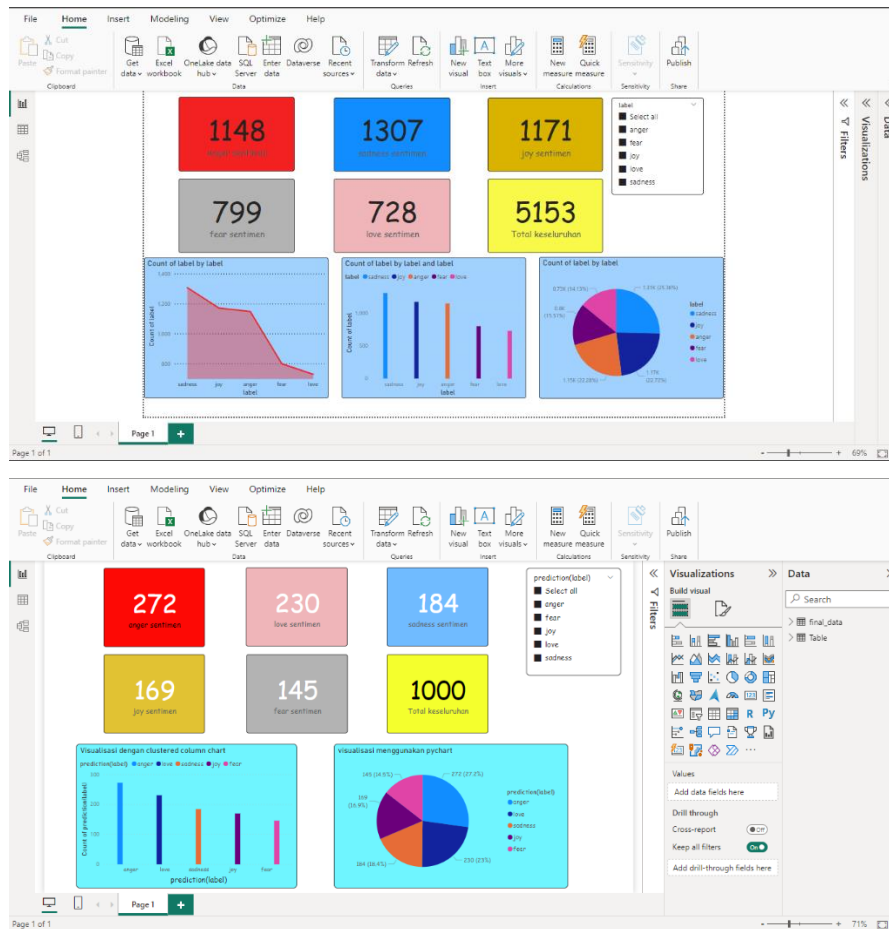
```
1 frames=[df1,df3]
2 df_asli=pd.concat(frames).reset_index(drop=True)
3 df_asli
```

```
1 # df_asli.to_csv('data_asli.csv', index=False)
2 # files.download('data_asli.csv')
```

5. Menyatukan Kolom

Setelah kedua data memiliki nama kolom yang sama maka proses penggabungan dapat dilakukan dengan menggunakan perintah *concat*, setelah berhasil maka dataset yang telah digabungkan dapat di download.

➤ POWER BI

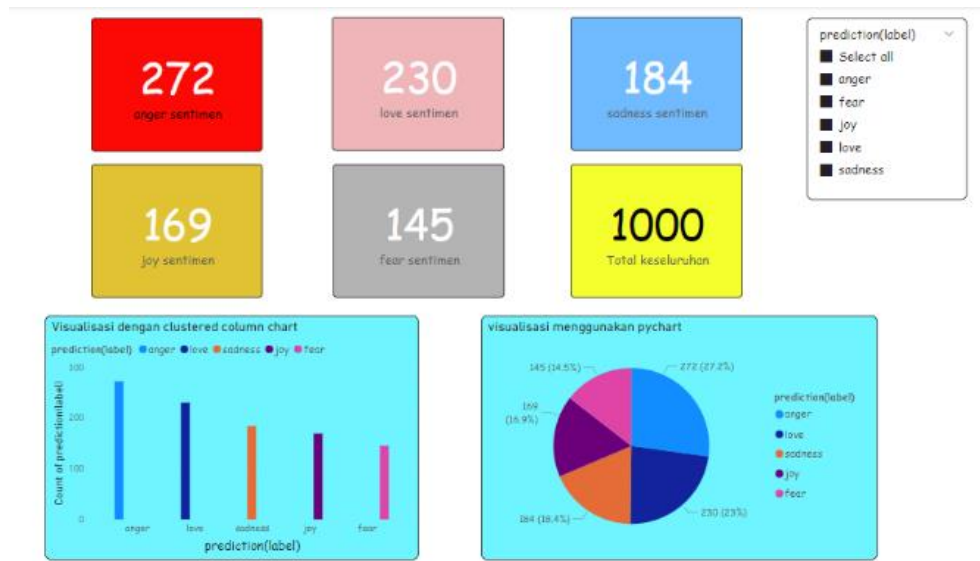


III. Hasil dan Pembahasan

Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen, kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif atau negatif. Sentiment analysis juga dapat menyatakan perasaan emosional sedih, gembira, atau marah

Seperti halnya dalam data mining, aplikasi text mining pada suatu studi kasus, harus dilakukan sesuai prosedur analisis. Langkah awal sebelum suatu data teks dianalisis menggunakan metode-metode dalam text mining adalah melakukan preprocessing teks. Database hasil ekstrak diolah melalui fase Pre Proses menggunakan aplikasi Rapidminer untuk melakukan tindakan seperti

Case Folding, Tokenizing, Cleaning, Stopword dan Stemming. Hasil dari analisis dapat di lihat pada gambar berikut :



Gambar Hasil Analisis Sentimen Tweet

Berdasarkan hasil analisis sentimen tweet menunjukkan bahwa sentimen anger merupakan yang terbanyak yaitu 272 (27,2%) sentimen, diikuti oleh sentiment love sebanyak 230 (23%) sentiment, sadness 184 (18,4%) swntimen, joy 169 (16,9%) sentiment, dan fear sebanyak 145 (14,5%)

Dari data tersebut dapat kami simpulkan bahwa kebanyakan tweet yang ada bernilai anger atau marah

BAB III

PENUTUP

I. Kesimpulan

Dapat disimpulkan bahwa penelitian ini dapat mengatasi kompleksitas analisis data tweet dalam era digital yang terus berkembang. Hal ini dilakukan melalui pengembangan model atau algoritma yang efektif, dengan fokus pada penggunaan algoritma *Naive Bayes*. *Twitter*, sebagai sumber data yang kaya dengan informasi dari berbagai aspek kehidupan manusia, memerlukan pendekatan analisis sentimen yang cermat.

Algoritma *Naive Bayes*, dengan sifat kesederhanaan dan kinerja yang cukup baik dalam klasifikasi teks, menjadi salah satu pilihan yang relevan dalam pengembangan model analisis sentimen *tweet*. Landasan teori dalam penelitian ini mencakup konsep-konsep dasar dalam pemrosesan teks dan analisis sentimen, dengan *Naive Bayes* menjadi salah satu algoritma yang telah terbukti efektif dalam pemrosesan teks.

Mengembangkan model atau algoritma untuk menghasilkan prediksi yang akurat dapat dilakukan dengan memperhatikan langkah penting seperti pembersihan data, pemilihan metode pemrosesan teks seperti TF-IDF(*Term Frequency-Inverse Document Frequency*), pembagian data yang objektif, pemilihan model yang mendukung seperti *Naive Bayes*, pelatihan model, dan yang paling penting adalah evaluasi model untuk mengukur sejauh mana model dapat memprediksi dengan akurat.

Dengan demikian, penelitian ini diharapkan dapat memberikan solusi untuk meningkatkan pemahaman dan analisis data *tweet* dengan menggunakan algoritma *Naive Bayes*. Hasil penelitian ini diharapkan dapat memberikan manfaat dalam berbagai konteks, termasuk pemantauan opini publik, analisis sentimen merek,

pemahaman tren topik, dan aplikasi bisnis lainnya yang melibatkan penggunaan data *tweet* berbasis sentimen.

DAFTAR PUSTAKA

- [1] Raghavan, Prabhakar. (2008). *Introduction to Information Retrieval*. Amerika Serikat: Cambridge University Press.
- [2] I.Rish. (2001). *An empirical study of the naive Bayes classifier, IJCAI 2001 Work. Empir. methods Artif. Intell., vol. 3, no. 22, 2001.*
- [3] Rajkumar, M. (2015). *Naive Bayes Text Classification Algorithm for Opinion Mining*. 113(18), 12-17.
- [4] Zhang, H. (2004). *The Optimality of Naive Bayes*. Amerika Serikat: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (*FLAIRS-2004*) (pp. 562-567).