Perbandingan Metode *Naïve Bayes* Dan *Decision Tree* C4.5 untuk Analisis Sentimen Produk Es Teh Indonesia di Media Sosial Twitter

Rachmat Hidayat, Abdul Barir Hakim, Risman Nugraha

Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ
abdulbarir.h@esqbs.ac.id

Diterima : 03 Januari 2024
Disetujui : 01 Februari 2024

***Abstrak*— Penelitian ini membandingkan metode Naive Bayes dan Decision Tree dalam analisis sentimen produk Es Teh Indonesia di media sosial Twitter. Analisis sentimen digunakan untuk memahami opini pengguna terhadap produk atau layanan. Meskipun kedua metode tersebut telah banyak digunakan dalam analisis sentimen, belum ada penelitian khusus yang membandingkannya untuk produk es teh Indonesia di Twitter. Data penelitian dikumpulkan melalui web scraping dari Twitter, dengan mencari tweet yang mengandung kata kunci terkait es teh Indonesia. Setiap tweet diberi label sentimen positif dan negatif berdasarkan konteks dan emosi yang terkandung di dalamnya. Hasil penelitian menunjukkan bahwa model algoritma Decision Tree C4.5 dengan memiliki tingkat akurasi yang lebih tinggi dengan tingkat akurasi sebesar 71.96% dan Naïve Bayes dengan tingkat akurasi sebesar 66.11% sehingga terbukti nilai akurasi Decision Tree C4.5 lebih baik daripada nilai akurasi Naïve Bayes dalam melakukan analisis sentimen. Kemudian hasil analisis sentimen terhadap Produk es teh indonesia menunjukkan kecenderungan pelanggan merespon negatif dengan dengan jumlah sentimen negatif 143 dan sentimen positif 28**

***Keywords* — analisis sentimen, metode Naive Bayes, Decision Tree, es teh Indonesia, Twitter**

I. PENDAHULUAN

Seorang konsumen atau pelanggan dikatakan puas jika ia senang dan memiliki kebiasaan perilaku yang kuat untuk menggunakan atau membeli kembali suatu produk atau jasa. Cara membentuk kepuasaan pelanggan tentunya dimulai dengan menyediakan produk atau jasa yang berkualitas atau bermutu tinggi, sehingga pelanggan merasa puas dengan pengalaman mengkonsumsinya. Kepuasan pelanggan berawal dari penilaian konsumen terhadap kualitas produk atau jasa yang diterimanya berdasarkan harapan yang telah terkonsep dalam pikirannya. Harapan tersebut muncul dari produk atau jasa yang telah diterima sebelumnya serta berita dari mulut ke mulut yang sampai pada pelanggan [1].

Dalam pengambilan keputusan untuk membeli suatu produk, salah satu faktor yang menjadi bahan pertimbangan adalah kualitas produk. Tentu pelanggan akan mencari kualitas produk yang terbaik untuk mereka beli. Kulitas produk ini jika dikaitkan dengan produk minuman, maka yang menjadi perhatian pelanggan adalah cita rasanya serta kemasan dalam bentuk *cup* yang menarik. Rasa yang enak serta penampilan dan tata penyajian yang unik akan membuat pelanggan lebih mudah tertarik terhadap produk yang ditawarkan. Hal itu dikaitkan dengan zaman dan trend pada saat ini yang memang sudah didukung oleh teknologi yang sesuai, sebagian besar pelanggan cenderung punya keinginan untuk mengabadikan foto makanan yang mereka beli dan memposting foto tersebut di media sosial. Jadi pelaku usaha kuliner yang kompetitif harus memperhatikan kualitas produk yang mereka jual kepada pembeli agar tetap dapat bersaing dan

bertahan di era gempuran para pengusaha saat ini [2]. Es Teh Indonesia bisnis es teh yang memiliki 300 cabang di Indonesia dengan omzet miliaran [3]. Namun, pada bulan september tahun 2022 terjadi permasalahan yang menyebabkan komentar negatif terhadap Es Teh Indonesia sumber didapatkan berdasarkan media sosial Twitter .

Berdasarkan hal ini maka menarik untuk dilakukan penelitian untuk mengetahui opini masyarakat tentang produk Es Teh Indonesia. Opini masyarakat tentang suatu produk atau jasa atau kejadian lainnya biasanya mereka tuangkan dengan menggunakan media sosial. Salah satu media sosial yang sering digunakan masyarakat Indonesia adalah Twitter. Twitter adalah sebuah *platform* yang bisa digunakan oleh banyak orang secara bersama-sama untuk menyampaikan pendapat mereka atau menyampaikan opini mereka. Dengan Twitter, banyak orang menuliskan tentang keluh kesah mereka mulai dari keluh kesah tentang kehidupan dalam keseharian mereka ataupun keluh kesah terhadap layanan yang mereka dapat atau mereka gunakan baik dari pemerintah atau dari pihak lainnya, atau pun mengenai hal-hal yang mereka anggap perlu untuk disebarkan. Salah satu isu yang sempat menjadi perbincangan warga Indonesia dalam Twitter adalah mengenai Es Teh Indonesia yang menjadi *Trending Topic* di Indonesia per tanggal 26 September 2022. Terkait dengan penelitian untuk mengetahui mengenai opini masyarakat mengenai Es Teh Indonesia, maka media sosial Twitter ini dipilih sebagai *platform* media sosial yang menjadi sumber data dalam penelitian ini. Alasan pemilihan Twitter ini adalah karena Twitter memiliki fitur-fitur yang mendukung penelitian seperti pencarian berdasarkan kata kunci, *trending topic*, tagar atau *hastag*. Pada penelitian ini, akan dilakukan pencarian dengan menggunakan kata kunci Es Teh Indonesia yang memang sempat menjadi *trending topic*.

Untuk menelaah lebih lanjut mengenai opini masyarakat mengenai Es Teh Indonesia berdasarkan opini pada media sosial Twitter, diperlukan suatu mekanisme untuk melakukan analisis sentimen yang terkandung pada kumpulan komentar dan unggahan guna mengetahui perasaan yang diberikan terhadap suatu topik atau sebuah objek di media sosial, yang pada penelitian ini adalah Es Teh Indonesia. Analisis sentimen biasa digunakan untuk menilai kesukaan atau ketidaksukaan publik terhadap suatu barang atau jasa yang mereka dapatkan atau gunakan. Sentimen yang dianalisis adalah merupakan informasi tekstual yang bersifat subjektif dan mempunyai polaritas positif dan negatif. Nilai polaritas ini dapat digunakan sebagai indikator dalam menentukan suatu kepuasan terhadap barang dan jasa.

Dengan mempertimbangkan fakta dan data yang telah diuraikan, hal ini menjadi alasan yang kuat untuk melakukan penelitian. Untuk mendukung penelitian tersebut, diperlukan metode klasifikasi yang spesifik untuk membedakan opini-opini yang ada dengan tingkat akurasi yang lebih baik. Pada penelitian ini, dipilih dua metode algoritma klasifikasi, yaitu Naïve Bayes dan Decision Tree, yang dapat digunakan untuk mengklasifikasikan tweet masyarakat terkait produk Es Teh Indonesia di media sosial Twitter.

Naïve Bayes Classifier merupakan salah satu algoritma yang mampu melakukan proses klasifikasi dengan cepat. Naïve Bayes juga merupakan salah satu algoritma yang sangat efisien dan efektif bahkan saat digunakan untuk menganalisis data berskala besar [4]. *Decision tree* adalah salah satu penerapan metode klasifikasi yang paling populer, dengan metode ini sebuah item dapat dikelompokkan dan dimodelkan dalam bentuk sebuah pohon keputusan, sehingga model yang dihasilkan dapat dipahami dengan mudah [5].

Berdasarkan penelitian yang dilakukan oleh Suciningsih dan rekan-rekannya yang berjudul *Comparation analysis of naïve bayes and decision tree C4.5 for caesarean section prediction* yang membandingkan metode Naïve Bayes dengan *Decision tree C4.5* dimana metode Naïve Bayes memiliki hasil akhir klasifikasi yang lebih tinggi daripada metode *Decision tree* yaitu 50% banding 45% [6]. Namun pada penelitian yang dilakukan oleh Romadloni dan rekan-rekannya yang berjudul perbandingan metode Naïve Bayes, KNN dan Decision Tree terhadap analisis sentimen

transportasi KRL Commuter Line, dimana penelitian ini membandingkan 3 metode yang hasilnya ternyata Decision Tree memiliki hasil akhir klasifikasi dengan tingkat akurasi yang lebih tinggi dibandingkan dengan 2 metode lainnya yaitu akurasi sampai dengan 100%, kemudian Naïve Bayes dan KNN yaitu akurasi sampai dengan 80% [7]. Berdasarkan temuan di atas yang menunjukan bahwa tidak ada *machine learning model* yang secara konsisten mengungguli yang lain. Sehingga pada penelitian ini diputuskan untuk dilakukan pengujian pada dua model klasifikasi yaitu Naïve Bayes dan *Decision Tree* C4.5.

Jadi penelitian ini berfokus untuk membandingkan nilai akurasi dari dua model algoritma klasifikasi yaitu Naïve Bayes dan Decision Tree C4.5, dalam melakukan klasifikasi *tweet* di *platform* media sosial Twitter yang berisi opini positif dan negatif tentang produk Es Teh Indonesia, pada periode September 2022 hingga Maret 2023. Melalui penelitian ini, diharapkan dapat diperoleh kesimpulan mengenai apakah opini publik cenderung lebih positif atau negatif terhadap produk Es Teh Indonesia. Selain itu, penelitian ini bertujuan untuk menentukan apakah *Decision Tree* C4.5 memiliki akurasi yang lebih baik daripada Naïve Bayes dalam melakukan analisis sentimen, sehingga dapat dijadikan rekomendasi untuk melakukan analisis sentimen.

II. METODE PENELITIAN

A. Metode yang Digunakan

Penelitian ini menggunakan metode studi perbandingan atau dikenal dengan istilah *comparative study*, yaitu studi dengan cara membandingkan dua atau lebih beberapa kondisi, beberapa kejadian, beberapa kegiatan, program dan hal-hal lainnya yang bisa dibandingkan. Penelitian ini membandingkan dua jenis metode klasifikasi yaitu klasifikasi dengan Naive Bayes dan *Decision Tree* C4.5.

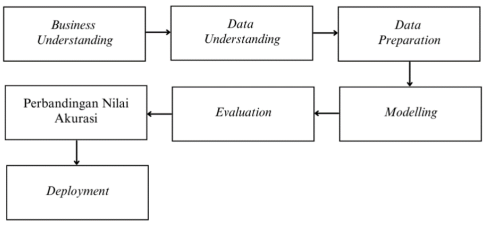
Penelitian ini menggunakan metode penelitian kualitatif deskriptif analisis, yaitu data-data yang diperoleh kemudian ditampilkan dalam bentuk skema, lalu dideskripsikan sehingga dapat memberikan kejelasan informasi yang realistis.

B. Sumber Data dan Metode Pengambilan Data

Sumber data penelitian ini adalah Data yang ditarik dari media sosial Twitter yang berkaitan dengan produk Es Teh Indonesia. Oleh karena itu akan dilajukan *Crawling Data Twitter* menggunakan *API Twitter* dan *RapidMiner* sebagai *tools*, lalu data dikumpulkan sesuai dengan fokus penelitian. Data dikumpulkan dengan latar belakang alami sebagai sumber data langsung. Jenis data dalam penelitian adalah data sekunder, yaitu data yang diperoleh dari sumber data yang diambil secara langsung dari *website Twitter* sejumlah 212 *tweet* menggunakan aplikasi *RapidMiner* dengan *keyword* pencarian Es Teh Indonesia, Rasa Es Teh Indonesia, dan tweet lainnya yang berhubungan dengan produk Es Teh Indonesia dengan memanfaatkan *search API* (*Application Programming Interface*) yang disediakan oleh *Twitter*.

C. Tahapan Penelitian

Penelitian menerapkan tahapan penelitian mengikuti framework untuk penambangan data yaitu *Cross-Industry Standard Process for Data Mining* atau disingkat CRISP-DM. CRISP-DM adalah sebuah metodologi standar industri dalam proses data mining yang dapat digunakan dalam berbagai lini bisnis atau bidang industri yang dikembangkan pada tahun 1996 oleh Daimler Chrysler, SPSS dan NCR. Sebagai sebuah metodologi, CRISP-DM membagi aktivitas-aktivitas yang perlu dilakukan menjadi tahapan-tahapan proyek. Deskripsi pekerjaan yang terkait dengan setiap fase dan hubungan antara pekerjaan ini memberikan gambaran umum tentang siklus hidup sebuah proyek penambangan data [8].

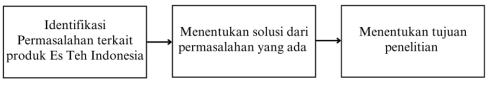


Gambar 1. Tahapan Penelitian mengikuti CRISP-DM

Dalam penelitian ini seperti yang dilihat pada gambar 1, terdapat enam tahap utama sesuai dengan tahapan dalam CRISP-DM, dan tambahan 1 tahap yang pada dasarnya adalah pendetailan dari salah satu tahap utama dari CRISP-DM. Jadi

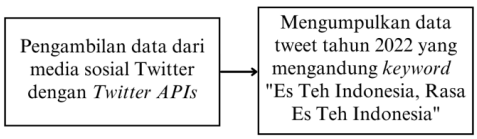
keseluruhannya ada 7 tahap yang dilakukan pada penelitian ini yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation,* Perbandingan Nilai Akurasi, dan *Deployment*. Tahap Perbandingan Nilai Akurasi adalah sebenarnya adalah bagian dari tahap Evaluasi yang dipisahkan untuk men-detailkan langkah-langkah yang dilakukan. Untuk lebih detailnya dari masing-masing tahapan adalah sebagai berikut:

1. *Business Understanding*: Pada tahap ini dilakukan pemahaman terhadap objek penelitian. Permasalahan yang diangkat adalah banyak pelanggan mengeluh terkait salah satu rasa dari produk Es Teh Indonesia yang dinilai mengandung banyak gula yang berlebih. Oleh karena itu, sebagai pemecahan masalah maka penelitian ini mencoba melakukan analisis sentimen dengan menggunakan model klasifikasi data mining untuk mengetahui lebih lanjut sebenarnya seberapa besar perbandingan sentimen positif dan sentimen negatif dari publik terhadap produk Es Teh Indonesia. Dalam penelitian ini, sebagai sumber data yang diambil dari media sosial *Twitter*, data berupa *tweet* opini masyarakat terkait tanggapan rasa dari produk Es Teh Indonesia September sampai November 2022. Motivasi pada tahap ini data tweet mengenai Es Teh Indonesia yang disajikan dalam bentuk teks akan dikelompokkan dengan algoritma klasifikasi berdasarkan kategori sentimen positif dan sentimen negatif. Pada tahap ini juga dilakukan pemahaman untuk mendapatkan metode klasifikasi yang lebih baik antara Naïve Bayes dan *Decision Tree* C4.5 ketika proses pengolahan data yang akan dilakukan dengan cara membandingkan hasil dari penerapan 2 algoritma tadi. Gambar 2 menujukkan aktivitas yang dilakukan di tahap *Business Understanding*



Gambar 2. Aktivitas Tahap Business Understanding

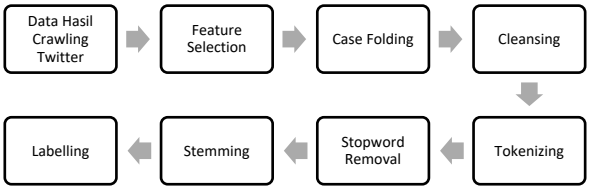
2. *Data Understanding*. Pada tahap ini tujuan yang ingin dicapai adalah memahami data yang akan digunakan sebagai bahan yang akan diteliti untuk dapat dilanjutkan ke tahap selanjutnya yaitu *Data Preparation*. Langkah-langkah yang dikerjakan pada tahap ini adalah mengumpulkan data tweet yang berisi opini masyarakat yang mengandung kata kunci “Es Teh Indonesia, rasa Es Teh Indonesia” pada September sampai November 2022 sebagai objek penelitian yang diambil dari sosial media Twitter menggunakan aplikasi *Automatic Crawling Twitter* melalui Twitter API. Gambar 2 menunjukkan aktivitas yang dilakukan pada tahap *Business Understanding*.



Gambar 2. Aktivitas Tahap *Business Understanding*

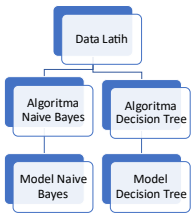
3. ***Data Preparation***: Berdasarkan data yang diperoleh dari tahap sebelumnya, dilakukan pembersihan untuk menjadi *dataset* yang siap diolah. Tahapan ini data tweet yang telah dikumpulkan akan melalui beberapa tahap pemrosesan teks yang terdiri dari *Case folding* yaitu mengubah seluruh huruf pada *dataset* menjadi huruf kecil, *Cleansing* yaitu menghilangkan karakter-karakter spesial yang tidak terbaca sistem dari tweet, *Tokenizing* yaitu mengubah kalimat menjadi potongan kata yang dipisahkan dengan koma, dimana setiap kata yang dipisahkan ini masing-masing akan menjadi *attribute dataset*, *Normalization* mengubah kalimat tidak baku/slang menjadi kalimat baku yang sesuai dengan KBBI, *Stemming* yaitu sebuah proses untuk menghilangkan awalan atau akhiran kata yang terdapat kata sambung, kata depan, kata ganti, menjadi kata dasar yang sesuai dengan KBBI, dan *stopword removal* yaitu menghilangkan kata – kata yang tidak penting seperti kata sambung dan kata ganti orang, dan kata-kata yang dianggap tidak mempunyai pengaruh terhadap kandungan sentimen positif dan negatif dari setiap *tweet*. Selanjutnya proses pe-*label*-an untuk menentukan *tweet* termasuk ke dalam kelas positif berisi pujian, saran, masukan, yang melambangkan emosi positif

seperti puas, senang dan bahagia, atau *tweet* tersebut digolongkan sebagai kelas negatif yang biasanya berisi keluhan, kalimat sindiran, kritik, dan kalimat lain yang berisi emosi negatif seperti amarah, kesal, dan kecewa Tujuan dari pemrosesan teks adalah supaya data yang didapat diolah menjadi bentuk yang sesuai untuk dilakukan proses selanjutnya yaitu *Modelling*.



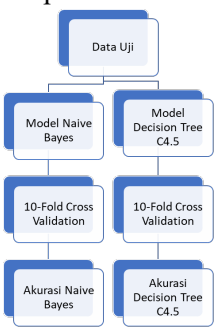
Gambar 3. Aktivitas Tahap *Data Preparation*

4. ***Modelling***: Merupakan tahap pemilihan teknik data mining dengan menentukan algoritma klasifikasi yang akan digunakan untuk membuat model. *Tool* yang digunakan adalah *RapidMiner*. Model analisis sentimen dibangun berdasarkan data latih yang telah disiapkan yang diujicobakan pada pengujian data untuk mengetahui sentimen terhadap produk Es Teh Indonesia. Data latih yang digunakan adalah data yang sudah diberikan label positif dan label negatif pada tahap sebelumnya. Hasil pengujian model yang dilakukan adalah mengklasifikasikan sentimen positif dan negatif dari Produk Es Teh Indonesia menggunakan algoritma Naive Bayes dan *Decision Tree* C4.5 yang digunakan untuk mendapatkan nilai akurasi yang lebih baik pada proses selanjutnya. Gambar 4 menunjukkan pembuatan model Naïve Bayes dan *Decision Tree* C4.5 dengan data latih yang sudah dipersiapkan yaitu yang sudah diberi label pada tahap sebelumnya.



Gambar 4. Aktivitas Tahap *Modelling*

5. ***Evaluation***: Pada tahap evaluasi ini bertujuan untuk menentukan kinerja dari model yang telah berhasil kita buat pada langkah sebelumnya. Metode untuk mengevaluasi kinerja model yang digunakan yaitu dengan menggunakan *cross validation* dengan k=10 atau disebut dengan *10-Fold Cross Validation* yang menghasilkan *accuracy*. Metode *Cross Validation* biasa digunakan untuk menghindari *overlapping* pada data testing, pada *cross validation* data uji dan data latih di-generate secara otomatis dengan melakukan melakukan pemisahan data dengan proses tertentu [9]. Hasil dari *Cross Validation* adalah nilai akurasi yang merupakan nilai rata-rata akurasi dari percobaan yang diulang berkali-kali sesuai jumlah k yang ditentukan dengan pada setiap percobaan dilakukan perubahan data uji dan data latih sedemikian rupa sehingga setiap data pasti akan mendapat peran sebagai data uji. Sehingga pada tahap ini bisa diukur nilai akurasi dari masing-masing model yaitu dengan menerapkan data uji terhadap masing-masing model yang sudah dibuat. Pada tahap ini hasil dari *10-Fold Cross Validation* dengan menggunakan bantuan *tools* Rapidminer adalah nilai *accuracy* yang mewakili kemampuan model untuk melakukan prediksi *label* dengan benar dan nilai Area Under Curve (AUC) dari penerapan data uji terhadap model. Sehingga kinerja 2 model ini bisa dibandingkan dengan membandingkan nilai akurasi yang dihasilkan. Sehingga tahap ini dilakukan perbandingan hasil akurasi Naive Bayes dengan *Decision Tree* C4.5. Aktivitas ini dapat dilihat pada Gambar 5.

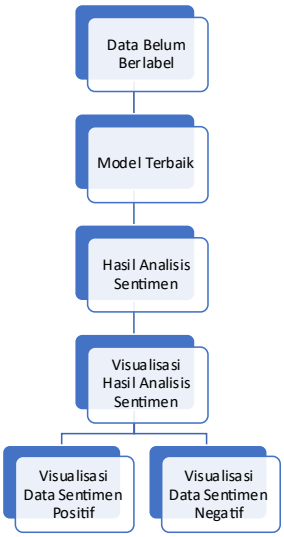


Gambar 5. Aktivitas Tahap *Evaluation*

6. Perbandingan **Nilai Akurasi**: Tahap ini adalah tahap pendetailan dari tahap Evaluation. Pada tahap ini dilakukan perbandingan nilai akurasi untuk 2 model untuk mendapatkan algoritma yang terbaik. Sehingga pada tahap selanjutnya bisa dilakukan

Analisis sentimen dengan menggunakan model dengan akurasi yang terbaik diantara 2 model.

7. ***Deployment***: Pada tahap ini, model terbaik yang sudah didapatkan setelah tahap evaluasi dan perbandingan Nilai Akurasi dipilih dan selanjutnya dilakukan proses klasifikasi keseluruhan data dengan menggunakan model ini. Setelah data diperoleh dan model dijalankan, maka akan didapatkan hasil analisis sentimen berupa klasifikasi sentimen positif dan negatif terhadap data tersebut. Kemudian dibuat kesimpulan berdasarkan hasil akurasi klasifikasi sentimen menggunakan model klasifikasi *machine learning* tersebut mana yang lebih baik yang nantinya dapat mengetahui hasil analisis sentimennya berupa jumlah sentimen positif dan jumlah sentimen negatif. Pada tahap ini dengan teknik visualisasi maka hasil analisis sentimen tersebut bisa ditampilkan dalam bentuk grafik. Selanjutnya analisis lebih lanjut bisa dilakukan dengan melihat kemunculan yang paling sering muncul untuk masing-masing kelas Sentimen Positif dan Sentimen Negatif, dengan ini berdasarkan data Sentimen Positif dan Sentimen Negatif juga bisa didapatkan kata-kata yang paling relevan untuk masing-masing kelas. Gambar 6 menunjukkan aktivitas pada tahap *deployment*.



Gambar 6. Aktivitas Tahap *Deployment*

III. HASIL DAN PEMBAHASAN

Hasil penelitian dan pembahasannya di

A. Business Understanding

Pada penelitian ini dilakukan pemahaman terhadap objek penelitian. Salah satu permasalahan yang diangkat adalah munculnya banyak komplain mengenai rasa produk Es The Indonesia yang terlalu manis, yang dari situ bisa ditarik kesimpulan bahwa terdapat kadar gula yang tinggi dalam Es Teh Indonesia. Banyak produsen es teh menggunakan jumlah gula yang berlebihan, yang dapat menyebabkan masalah kesehatan, seperti obesitas dan penyakit diabetes. Dan berdasarkan banyaknya komplain tentang rasa manis produk ini cukup mengkhawatirkan. Cuma tidak semua orang mengeluhkan hal ini dan menarik untuk lebih mendalami seberapa besar perbandingan publik yang mengeluh dan yang puas akan produk Es Teh Indonesia. Oleh karena itu perlu diteliti lebih lanjut mengenai sentimen publik terhadap produk ini dan sebagai solusi maka sebuah dibuatlah model klasifikasi algoritma yang nantinya dapat digunakan untuk melakukan klasifikasi terhadap sentimen masyarakat atau *customer* mengenai produk Es Teh Indonesia yang datanya diambil dari media sosial Twitter dari September 2022 sampai November 2022 sebagai objek penelitian ini. Motivasi pada tahap ini data *tweet* mengenai produk Es Teh Indonesia dalam bentuk teks akan dikelompokkan berdasarkan kategori sentimen positif dan sentimen negatif. Pada tahap ini juga dilakukan pemahaman untuk mendapatkan metode klasifikasi yang terbaik antara Naïve bayes dan *Decision Tree* C4.5 pada saat proses pengolahan data yang akan dilakukan dengan cara membandingkan kinerja dari kedua algoritma tadi dalam melakukan klasifikasi sentimen positif dan negatif. Indikator kinerja yang digunakan untuk perbandingan adalah akurasi yaitu kemampuan algoritma klasifikasi untuk mengklasifikasikan setiap *tweet* dengan benar sesuai dengan klasifikasinya.

B. Data Understanding

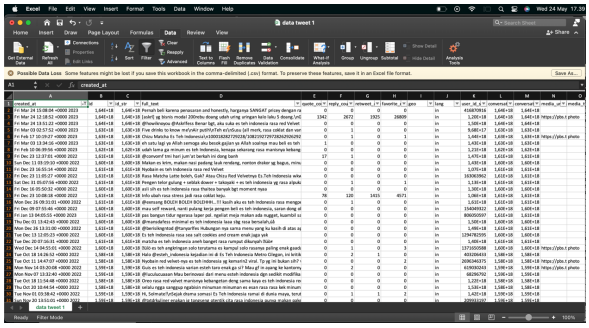
Dalam proses memahami data yang akan digunakan sebagai bahan yang akan diteliti untuk dapat dilanjutkan ke tahap selanjutnya yaitu preprocessing. Dimana Langkah-langkah yang akan dikerjakan mulai dari mengumpulkan data tweet yang berisi opini atau tanggapan pelanggan yang mengandung kata kunci “rasa es teh Indonesia” pada September 2022 sampai November 2022 sebagai objek penelitian yang

diambil dari sosial media Twitter menggunakan aplikasi *Automatic Crawling Twitter* melalui *Tweet Harvest (Twitter Crawler)* karena data yang diambil valid berdasarkan waktu dan hasil tweet dari media sosial Twitter.

C. Data Preparation

Pada penelitian ini dilakukan proses pengumpulan data, dimana data yang diambil merupakan data tweet yang berisi opini pelanggan yang mengandung kata kunci "rasa es teh Indonesia" pada September 2022 sampai Maret 2023 pada media sosial Twitter. Proses *crawling data* dijalankan secara *manual* menggunakan akses token yang didapatkan dari *Twitter API* dengan menggunakan *Tweet harvest* atau *Twitter crawler* untuk mendapatkan informasi serta data yang diinginkan terdapat beberapa langkah pengerjaan dalam melakukan *crawling data*.

Pada tahap ini *tweet harvest* sudah berhasil mengambil data dari *Twitter* terkait produk es teh Indonesia berdasarkan kata kunci rasa es teh Indonesia.



Gambar 7. Hasil Crawling

Dari hasil *crawling* ini langkah berikutnya adalah dilakukan *Feature Selection* yaitu memilih *attribute* sesuai dengan data yang dibutuhkan. Dengan bantuan *Rapidminer* dilakukan proses pemilihan *attribute* yang dibutuhkan yaitu cukup *full-text tweet*-nya saja. Hasil dari crawling data ini didapatkan 212 *tweet*. Selanjutnya dari 212 *tweet* ini dilakukan penghapusan *tweet* yang berulang atau duplikat. Setelah proses ini didapatkan 172 *tweet*. Selanjutnya pada 172 *tweet* ini dilakukan *Preprocessing* sesuai dengan gambar 3 untuk menyiapkan data sehingga bisa masuk ke tahap *modelling*.

1) Case Folding

Pada tahap *preprocessing* dilakukan proses *Case folding* dengan tujuan untuk mengubah bentuk tulisan menjadi seragam dalam huruf kecil. Karena tidak semua dokumen teks memiliki penggunaan huruf kapital yang tepat. Hal tersebut dilakukan agar kata yang memiliki huruf kecil dan besar tidak dibedakan sehingga dianggap tidak memiliki perbedaan arti atau makna. Pada penelitian ini *dataset tweet* akan diseragamkan menjadi huruf kecil. Tabel 1 menunjukkan contoh hasil *Case Folding*.

Tabel 1. Contoh Hasil *Case Folding*

Sebelum	**Sesudah**
Chizu Matcha Es Teh Indonesian Ini muanis beneran manis banget Tapi ada cheese cream yang bisa balancein rasa manisnya itu Kek gurihmanis jadi saTUUU AAAAHHH GATAU POKOKNYA ENAKK	chizu matcha es teh indonesian ini muanis beneran manis banget tapi ada cheese cream yang bisa balancein rasa manisnya itu kek gurihmanis jadi satuuu aaahhh gatau pokoknya enakk
Oreo rasa red valvet manisnya kebangetan deng sama kaya es teh indonesia red valvet	oreo rasa red valvet manisnya kebangetan deng sama kaya es teh indonesia red valvet
GES REKOMENDASI RASA ES TEH INDONESIA DONG	Ges rekomendasi rasa es teh Indonesia dong

2) Cleansing Data

Pada bagian *Cleansing* ini dilakukan dengan tujuan untuk mengubah ataupun menghapus karakter atau kata tertentu yang ada pada data *tweet*, yang dianggap tidak perlu dan tidak memberikan kontribusi makna menjadi sentimen positif dan negatif, sebagai contoh kata *Hastag Twitter* (#), alamat situs (*url*), *Retweet* (RT), simbol/tanda baca dengan menggunakan *Regular Expression* (*Regex*). Tabel 2 menunjukkan contoh hasil cleansing yang dilakukan untuk menghapus mention, url dan simbol-simbol.

Tabel 2. Contoh Hasil *Cleansing Data*

	Sebelum	**Sesudah**
Cleaning Mention	@howiknoyou @Askrlfess Benar bgt, aku suka es teh indonesia rasa red Velvet tuh enak bgtttt!!!!	Benar bgt, aku suka es teh indonesia rasa red Velvet tuh enak bgtttt!!!!
Cleaning url	Heboh polemik Es Teh Indonesia gegara cuitan soal rasa salah satu produk yang disebut terlalu manis. Kandungan gula pun menjadi pembahasan setelah polemik ini muncul. Lalu, berapa takaran gula tambahan yang sebenarnya aman dikonsumsi dalam sehari? Simak... https://t.co/NuL4FeZK2C	Heboh polemik Es Teh Indonesia gegara cuitan soal rasa salah satu produk yang disebut terlalu manis. Kandungan gula pun menjadi pembahasan setelah polemik ini muncul. Lalu, berapa takaran gula tambahan yang sebenarnya aman dikonsumsi dalam sehari? Simak...

Cleaning Simbol	Minuman sejenis Es Teh Indonesia yang kadar gulanya bisa dibilang tinggi juga ada. Mungkin malah rata-rata ya segitu gulanya. Bayangin aja pake gula, ditambah kental manis, terus dicampur susu, entah cokelat/rasa lain.	Minuman sejenis Es Teh Indonesia yang kadar gulanya bisa dibilang tinggi juga ada Mungkin malah rata-rata ya segitu gulanya Bayangin aja pake gula ditambah kental manis, terus dicampur susu entah cokelatrasa lain.

3) Stopword Removal

Pada tahap ini dilakukan proses menghapus kata-kata yang sering dipakai namun tidak memberikan kontribusi makna yang penting. Contoh *stopword* dalam bahasa Indonesia di antaranya adalah "yang", "dari", "di", "dan", "adalah", dan sebagainya. Tujuan dari tahap ini adalah supaya bisa fokus pada kata yang mempunyai informasi penting sehingga dapat mencapai akurasi yang lebih baik dan akan mengurangi ukuran *dataset* sehingga akan mengurangi waktu proses pelatihan karena jumlah data yang digunakan berkurang. Pada penelitian ini menggunakan data *stopword list* yang sudah tersedia di *Kaggle* milik Oswin Rahadyan Hartono yaitu *Indonesian stoplist*. Contoh hasil proses *Remove Stoplist* ditampilkan pada tabel 3.

Tabel 3. Contoh Hasil *Remove Stoplist*

Sebelum	**Sesudah**
udah lama ga minum es teh Indonesia, kenapa sekarang rasa manisnya kebangetan ya sampe ga berani ngabisinnya.	Udah ga minum es teh Indonesia, rasa manisnya kebangetan ya sampe ga berani ngabisinnya.

4) Tokenizing

Pada tahap dilakukan proses mengekstrak kata-kata dalam *full-text tweet* menjadi fitur-fitur dalam *dataset*. Selain itu kata yang hanya memiliki jumlah karakter sedikit (<=3 karakter) atau jumlah karakter terlalu banyak (>25 karakter) akan dihapus.

5) Labelling

Dalam penelitian ini data yang sudah terkumpul dan telah melewati tahap *preprocessing*, maka tahap selanjutnya adalah tahap pelabelan. Proses pelabelan yang dilakukan secara manual berdasarkan pengetahuan individu dalam memahami makna setiap *tweet* yang dibaca yang prosesnya akan memakan waktu yang cukup banyak. Dalam pemberian *label* pada suatu *dataset* diperlukan penambahan kolom baru dengan tujuan untuk memberi identifikasi terkait *dataset* tersebut yang diberi nama kolom *sentiment*. Di kolom inilah *dataset* akan ditandai dan diberi *label* pada setiap *tweet* berdasarkan sentimen yang ada pada *tweet* tersebut dengan cara melabeli secara *manual* menjadi dua jenis polaritas, yaitu sentimen positif dan sentimen negatif. Sentimen positif adalah opini yang memiliki unsur kebahagiaan di dalamnya yang bersifat positif, seperti menyetujui, menyemangati, memuji, bersyukur dan sebagainya. Sedangkan sentimen negatif merupakan kebalikan dari sentimen positif yaitu yang memiliki unsur yang merusak atau destruktif yang bersifat negatif, seperti hujatan, kekecewaan, ketakutan, penolakan, amarah, dan sebagainya. Contoh hasil pemberian label sentimen pada *dataset tweet* yang sudah melalui proses pembersihan data ditampilkan pada tabel pada tabel 4.

Tabel 4. Contoh Hasil *Labeling*

Teks	**Sentimen**
Benar bgt aku suka es teh indonesia rasa red Velvet tuh enak bgtttt	positif
udah lama ga minum es teh Indonesia kenapa sekarang rasa manisnya kebangetan ya sampe ga berani ngabisinnya	negatif
es teh indonesia yg rasa avocado kok bisa enakkkk bangett yaaa beneran senagih itu	positif
tapi asli pertama kali beli es teh jumat kemarin rasa chizu taro dan gila manis banget giung ke pahit jatuhnya ga enak di lidah krim kejunya ga bisa netralin manisnya tetap manis kebangetan hiks aku suka minuman manis padahal tp utk es teh maaf keknya ga lagi huhu	negatif

D. Modelling

Setelah dilakukan *dataset* selesai pada tahap *Data Preparation* berarti *dataset* sudah siap digunakan untuk membuat model *machine learning* dengan menerapkan algoritma *machine learning* sesuai dengan perencanaan penelitian. Pada tahap ini dengan menggunakan alat bantu *Rapidminer* data yang sudah berlabel manual dibuat model naïve bayes-nya dengan dikenakan algoritma Naïve Bayes dan juga dibuat kan pula

model *Decision Tree* C4.5 dengan menggunakan algoritma C4.5. Sehingga pada tahap ini dihasilkan 2 model.

E. Evaluation

Kedua Model yang dihasilkan pada tahap *Modelling* harus diukur kinerjanya dengan untuk menentukan mana dari kedua model ini yang lebih baik untuk digunakan dalam pengklasifikasian sentimen positif dan sentimen negatif. Pada tahap evaluasi dengan bantuan *Rapidminer* digunakan *10-Fold Cross Validation* terhadap masing-masing model.

Hasil penerapan *10-Fold Cross Validation* pada model Naïve Bayes mendapatkan hasil *accuracy* sebesar 66.11% dan nilai AUC 0.450 yang berarti tingkat akurasi sangat lemah.. Sedangkan pada model *Decision Tree 4.5* hasil *accuracy* dan AUC yang didapat adalah 71.96% dan 0.607. Nilai AUC ini berarti tingkat akurasi termasuk tingkat akurasi Sedang. Perbandingan nilai accuracy dan AUC beserta interpretasinya ini ditampilkan pada tabel 5.

Tabel 5. Perbandingan Kinerja Model

Model	Accuracy	AUC	Interpretasi AUC
Naïve Bayes	66.11%	0.450	Tingkat Akurasi Sangat Lemah
Decision Tree C4.5	71.96%	0.607	Tingkat Akurasi Sedang

F. Perbandingan Nilai Akurasi

Berdasarkan data *accuracy* dan Interpretasi AUC pada tabel 5, maka dari kedua model yang telah dibuat, model yang lebih baik digunakan untuk melakukan analisis sentimen pada data yang tersedia adalah *Decision Tree C4.5* dengan nilai accuracy yang tidak terpaut jauh, tetapi kalau dari interpretasi AUC-nya ada perbedaan 2 kategori antara tingkat akurasi sangat lemah dan tingkat akurasi sedang.

Hasil pengukuran akurasi ini ternyata berbeda dengan hasil penelitian oleh Suciningsih dan rekan-rekannya [6] dimana pada penelitian tersebut hasilnya yang lebih baik adalah model Naïve Bayes tetapi dengan tingkat akurasi yang lebih rendah yaitu pada 50% dengan model Naïve Bayes dan 45% pada model *Decision Tree C4.5.*

Tetapi dilain pihak hasil penelitian ini ternyata sejalan dengan penelitian yang dilakukan oleh Romadloni dan rekan-rekannya menghasilkan akurasi yang terbaik pada model *Decision Tree* [7], dengan nilai *accuracy* sebesar 100% dan metode lain sebesar 80%.

G. Deployment

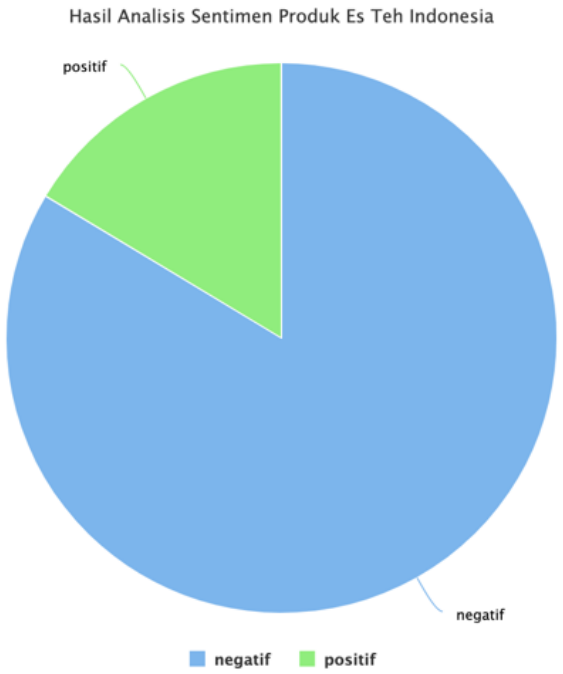
Pada tahap ini, dijalankan model yang lebih baik yaitu *Decision Tree C4.5* berdasarkan tahap evaluasi dan perbandingan nilai akurasi pada 171 tweet data Twitter yang membahas tentang produk es teh indonesia dari September 2022 sampai Maret 2023. Setelah data didapatkan, dibersihkan, dan model dengan akurasi tertinggi dijalankan, hasil klasifikasi sentimen didapatkan terhadap data tersebut. Kemudian, dibuat kesimpulan berdasarkan hasil akurasi klasifikasi sentimen positif dan negatif menggunakan model algoritma klasifikasi dengan akurasi tertinggi tersebut dapat mengetahui hasil analisis sentimennya berupa jumlah sentimen positif dan jumlah sentimen negatif dari produk es teh indonesia.

Hasil dari analisis sentimaen terhadap 171 *tweet* adalah untuk kelas sentimen negatif didapatkan 143 record dan kelas positif didapatkan 28 record. Hal ini bisa dilihat pada tabel perbandingan kelas sentimen positif dan negatif pada tabel 6. Pada tabel 6 jumlah *record* dan perbandingan persentase masing-masing kelas ditampilkan.

Tabel 6. Perbandingan Kelas Sentimen Negatif dan Positif

Index	Nominal value	Absolute count	Fraction
1	negatif	143	0.836
2	positif	28	0.164

Selanjutnya dengan teknik visualisasi data didapatkan grafik dalam bentuk *Pie Chart* yang dapat dilihat pada gambar 8. Dengan bantuan pie chart secara visual dapat dibandingkan besarnya perbedaan porsi dari sentimen positif dan negatif. Pada gambar 8 ini, sangat terlihat bahwa sentimen negatif jauh lebih besar porsinya dibandingkan sentimen positif.



Setelah dianalisis lebih lanjut untuk kelas Sentimen Negatif ternyata 3 kata yang paling relevan yang muncul di kelas negatif adalah manis, beli dan gula. Ini dapat dilihat pada tabel 7.

Tabel 7. Kata Relevan pada Kelas Negatif

Kata	Kemunculan Kata di Kelas Negatif
manis	40
beli	29
gula	22

Sedangkan pada kelas positif, 3 kata yang paling relevan yang muncul adalah enak, beli dan sih. Ini dapat dilihat pada tabel 8.

Tabel 8. Kata Relevan pada Kelas Positif

Kata	Kemunculan Kata di Kelas Positif
enak	19
beli	10
sih	9

IV. KESIMPULAN

Penelitian ini membandingkan metode Naïve Bayes dan Decision Tree dalam analisis sentimen produk Es Teh Indonesia di media sosial Twitter. Berdasarkan hasil penelitian, berikut terdapat beberapa kesimpulan :

1. Berdasarkan hasil evaluasi, model algoritma Decision Tree C4.5 menunjukkan tingkat akurasi lebih tinggi daripada model algoritma Naïve Bayes. Akurasi model Decision Tree C4.5 adalah 73,91%, sedangkan Naïve Bayes adalah 63,77%. Dengan demikian, dapat disimpulkan bahwa model algoritma klasifikasi Decision Tree memiliki performa akurasi yang lebih baik daripada Naïve Bayes dalam menganalisis sentimen pada kumpulan Tweet tentang produk es teh Indonesia.
2. Hasil analisis klasifikasi sentimen terhadap produk es teh Indonesia, terlihat bahwa pelanggan cenderung memberikan respon yang negatif. Jumlah sentimen negatif mencapai 143, sementara sentimen positif mencapai 28.

Selain itu masih banyak kekurangan dari penelitian yang telah dilakukan yang bisa diperbaiki pada penelitian selanjutnya. Sehingga berikut ini merupakan saran yang dapat diberikan pada peneliti selanjutnya yaitu :

1. Perluasan sumber data: Selain menggunakan data dari Twitter, penelitian ini dapat mempertimbangkan untuk menggabungkan data dari *platform* media sosial lainnya, seperti Instagram, Facebook, atau *platform* lain yang relevan. Hal ini akan memberikan perspektif yang lebih komprehensif tentang sentimen pengguna terhadap produk Es Teh Indonesia.
2. Penanganan bahasa yang tidak baku: Media sosial sering kali memuat teks dengan bahasa yang tidak baku, mengandung singkatan, slang, atau bahasa gaul. Oleh karena itu, dalam penelitian ini, penting untuk mempertimbangkan penggunaan teknik pemrosesan bahasa alami atau istilahnya *natural language processing* yang mampu mengatasi variasi bahasa yang tidak baku tersebut.
3. Analisis Sentimen *Real-time*. Dalam penelitian ini, data dari media sosial Twitter mungkin telah dikumpulkan pada waktu tertentu. Namun, dalam perkembangan selanjutnya, peneliti dapat mencoba untuk melakukan analisis sentimen secara *real-time*, sehingga hasilnya dapat memberikan wawasan yang lebih aktual tentang sentimen pengguna terhadap produk Es Teh Indonesia di Twitter.

4. Klasifikasi sentimen berdasarkan aspek. Sebagai pengembangan lebih lanjut, penelitian ini dapat dikembangkan lebih lanjut dengan mempertimbangkan untuk melakukan analisis sentimen berdasarkan aspek-aspek tertentu yang relevan dengan produk Es Teh Indonesia, seperti rasa, harga, kemasan, layanan pelanggan, dan lain sebagainya. Hal ini akan memberikan wawasan yang lebih rinci tentang sentimen pengguna terhadap aspek-aspek khusus dari produk tersebut.

DAFTAR PUSTAKA

[1] A. Afnina and Y. Hastuti, "Pengaruh Kualitas Produk terhadap Kepuasan Pelanggan," *Jurnal Samudra Ekonomi Dan Bisnis,* vol. 9, no. 1, p. 21–30, 2018.

[2] A. Halim, A. K. Djaelani and M. K. A. B. Suharto, "Pengaruh Kualitas Produk, Harga Dan Lokasi Terhadap Keputusan Pembelian (Studi Pada Kafe Es Teh Indonesia Tlogomas Kota Malang)," *E-JRM: Elektronik Jurnal Riset Manajemen,,* vol. 12, no. 1, 2023.

[3] S. A. Salsabilah, "Visualisasi Model Bisnis Es Teh Indonesia," *OSF Preprints.,* 2022.

[4] M. Ahmad, S. Aftab, I. Ali and N. Hameed, "Hybrid tools and techniques for sentiment analysis: A review," *Int. J. Multidiscip. Sci. Eng,* vol. 8, no. 3, pp. 29-33, 2017.

[5] J. Fadlil and W. F. Mahmudy, "Pembuatan Sistem Rekomendasi Menggunakan Decision Tree dan Clustering," *Jurnal Ilmiah Kursor,* vol. 3, no. 1, 2007.

[6] I. G. A. Suciningsih, M. A. Hidayat and R. A. Hapsari, "Comparation Analysis of Naïve Bayes and Decision Tree C4.5 for Caesarean Section Prediction," *Journal of Soft Computing Exploration,* vol. 2, no. 1, pp. 46-52, 2021.

[7] N. V. Romadloni, I. Santoso and S. Budilaksono, "Perbandingan Metode Naive Bayes, KNN dan Decision Tree terhadap Analisis Sentimen Transportasi KRL Commuter Line," *Ikraith Informatika,* vol. 3, no. 2, 2019.

[8] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi),* vol. 5, no. 6, p. 1120–1126, 2021.

[9] M. R. Fahdia, D. Riana, F. Amsury, I. Saputra and N. Ruhyana, "Komparasi Algoritma Klasifikasi untuk Orientasi Minat Mahasiswa dalam Penuntasan Studi," *JIRA: Jurnal Inovasi Dan Riset Akademik,* vol. 2, no. 7, pp. 970-1007, 2021.

[10] I. Susianti, S. S. Ningsih, M. al Haris and T. W. Utami, "Analisis Sentimen pada Twitter Terkait New Normal dengan Metode Naïve Bayes Classifier," *EDUSAINTEK,* vol. 4, 2020.