ISSN: 2962-3545
Prosiding SAINTEK: Sains dan Teknologi Vol.2 No.1 Tahun 2023
Call for papers dan Seminar Nasional Sains dan Teknologi Ke-2 2023
Fakultas Teknik, Universitas Pelita Bangsa, Februari 2023

Pentingnya Algoritma Naïve Bayes Sebagai Pengklasifikasi Data

The importance of the Naïve Bayes Algorithm as a data classifier

Rhendy Diki Nugraha
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
rendydikinugraha@gmail.com

Abstract

Data is one of the most important things, currently a lot of data has been created, so a method is needed to organize or classify all this data, the goal is to help humans in solving problems. This method is the Naive Bayes Algorithm. The research method used is to collect several interrelated journals, and contains the use of the Naive Bayes Algorithm method. Of course with the results already in each journal. By conducting research on fake news data, disease data and other data. Also compare with other algorithmic classification methods that are still related to data classification. The result is to prove that the Naive Bayes algorithm method is the most accurate data classification method. Proven by the results of a study of some examples of data classification and comparison with other classification methods. The benefit is that we get accurate data results and the data can be used as a reference for the future.

***Keywords**: Naïve Bayes, Hoax, Data, Classification, Algorithm*

Abstrak

Data merupakan salah satu hal yang sangat penting, saat ini telah banyak sekali data yang dibuat, sehingga diperlukan suatu metode untuk menyusun atau mengklasifikasikan semua data tersebut, tujuannya adalah untuk membantu manusia dalam menyelesaikan masalah. Metode ini adalah Algoritma Naïve Bayes. Metode penelitian yang digunakan adalah dengan mengumpulkan beberapa jurnal yang saling terkait, dan memuat penggunaan metode Algoritma Naïve Bayes. Tentunya dengan hasil yang sudah ada di jurnal masing-masing. Dengan melakukan penelitian terhadap data berita hoaks, data penyakit dan data lainnya. Juga membandingkan dengan metode klasifikasi algoritma lainnya yang masih berkaitan dengan klasifikasi data. Hasilnya adalah membuktikan bahwa metode algoritma Naive Bayes adalah metode klasifikasi data yang paling akurat. Dibuktikan dengan hasil kajian terhadap beberapa contoh klasifikasi data dan perbandingan dengan metode klasifikasi lainnya. Manfaatnya, kita akan mendapatkan hasil data yang akurat dan data tersebut dapat digunakan sebagai acuan untuk masa yang akan datang.

Kata kunci: Naïve Bayes, Hoaks, Data, Klasifikasi, Algoritma

Pendahuluan

Perangkat komunikasi menjadi lebih mudah saat ini karena peningkatan dalam beberapa teknologi baru seiring dengan peningkatan yang dilakukan pada sistem yang tersedia untuk komputasi dan protokol di Internet [1]. Persebaran arus informasi melalui internet juga, saat ini sangatlah mudah dan cepat dimana waktu serta jarak tidak menjadi sebuah penghalang. Jumlah pengguna internet di Asia dimana pada Maret 2021 Indonesia telah mencapai 212,35 juta jiwa pengguna internet. Menggunakan data tersebut, Indonesia menempati peringkat ketiga di antara negara-negara dengan pengguna internet terbanyak di Asia. Sehingga dengan meningkatnya pengguna internet, masyarakat dapat mengkonsumsi tiap informasi yang tersebar dengan cepat [2].

Untuk berbagi informasi, media sosial telah menjadi platform yang populer. User memiliki akses mudah kesana dengan biaya minimal. Mendeteksi berita palsu adalah tugas penting yang memberikan privasi pengguna dan meningkatkan kepercayaan. Untuk mempermudah mendeteksi berita palsu dibutuhkanlah sebuah aplikasi atau metode, metode ini akan membantu kita mendeteksi dan melacak berita palsu dari media sosial [3].

Aplikasi atau metode yang dapat membantu mendeteksi tersebut adalah algoritma Naïve Bayes. Naïve Bayes merupakan metode klasifikasi yang mengacu pada teorema bayes yang pertama kali dikemukakan oleh Thomas Bayes seorang ilmuwan yang berasal dari inggris. Metode ini menggunakan metode klasifikasi statistik. Ciri utama dari metode ini adalah asumsi akan independensi dari masing- masing kondisi [4]. Keuntungan dari teorema ini adalah kita dapat membangunnya dengan sangat mudah dan akan bekerja untuk kumpulan data yang besar, dan kelemahan dari teorema ini adalah menganggap semua variabel bergantung [3].

Metode Penelitian

Dalam penarikan kesimpulan untuk membuktikan bahwa metode klasifikasi Naïve Bayes sangat penting. Saya telah berhasil mengumpulkan sepuluh buah jurnal (5 jurnal Indonesia terindeks SINTA dan 5 jurnal Internasional terindeks SCOPUS), yang tentunya saling berhubungan, dan berisi tentang penggunaan metode algoritma Naïve Bayes. Dengan semua hasil yang sudah dipaparkan dalam setiap jurnalnya, yaitu dengan melakukan penelitian kepada data berita hoax, data penyakit, dan data lainnya yang menggunakan metode Naive Bayes. Juga, membandingkan dengan metode klasifikasi algoritma lainnya yang masih berhubungan dengan klasifikasi data.

Hasil dan Pembahasan

Untuk hasil dan pembahasan disini, Saya sudah mengambil isi dari kesepuluh jurnal yang sudah dikumpulkan. Hasilnya didapatkanlah beberapa bahasan, mengenai penggunaan metode klasifikasi Naive Bayes ini. Pembahasan yang sudah didapatkan diantaranya adalah Naive Bayes dalam penanganan Covid – 19, Naive Bayes dalam penyintasan berita hoax, beberapa data yang pernah diklasifikasikan oleh Naive Bayes, dan yang terakhir perbandingan dengan algoritma klasifikasi data lainnya. Berikut ini adalah hasil dari pembahasan yang sudah Saya kumpulkan :

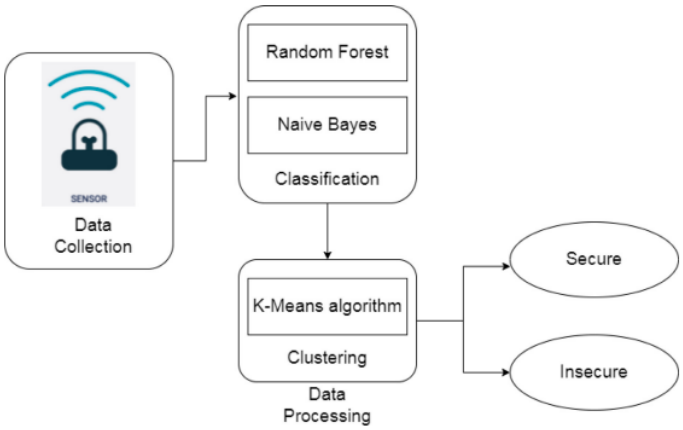
1. Naïve Bayes dalam penanganan Covid-19

Coronavirus disease 2019 atau disingkat Covid-19, adalah suatu penyakit menular yang ditimbulkan oleh sejenis coronavirus baru yaitu severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Covid-19 diidentifikasi pada awal Januari 2020 sebagai penyebab epidemi pneumonia yang menyerang kota Wuhan, provinsi Hubei, dan menyebar dengan cepat ke seluruh Cina. Setelah menginfeksi dan menyebabkan kematian ribuan orang di Cina, virus kemudian menyebar mencapai Italia dan negara dibenua Eropa lainnya serta Amerika Serikat dengan jumlah kasus baru yang dikonfirmasi meningkat setiap harinya [5].

Perekonomian dunia juga, saat ini dipengaruhi oleh penyebaran penyakit Covid-19 dan social distancing perlu dijaga. Ekonomi dunia dan gaya hidup serta teknologi juga terpengaruh. Sebuah metode baru untuk memantau penyakit menggunakan algoritma pembelajaran mesin seperti Random Forest dan Naïve Bayes diusulkan [1].

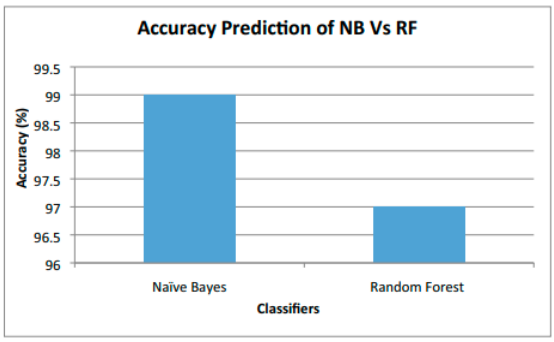
Penggunaan sistem pakar naïve bayes akan membantu dalam mendiagnosis virus Covid-19 pada tahap yang jauh lebih awal. Manfaatnya akan didapat baik oleh dokter maupun pasien karena para dokter akan dapat menghemat waktu dalam mendiagnosis pasien dan lebih fokus dalam merawat pasien Covid-19 dan pasien akan dapat mendiagnosis dirinya sendiri dan mendapatkan penanganan yang sesuai secepatnya[5]. Tujuan yang ingin dicapai adalah dikembangkannya sebuah model prediksi terkait pertumbuhan kasus Covid-19 menggunakan algoritma Naïve Bayes sebagai salah satu algoritma machine learning yang dapat diukur tingkat akurasi hasil prediksinya. Selain itu, melalui penelitian ini dilakukan analisa hasil peramalan terhadap data statistik pertumbuhan kasus Covid-19 [6].

Bagian-bagian dari sistem yang diusulkan dalam penggunaan Naïve Bayes : [1]



Gambar 2 Sistem yang diusulkan

(i)Pengumpulan Data: Data
dikumpulkan menggunakan sensor IoT dan data ini berisi detail tentang jarak antar orang beserta gambarnya. Penggunaan masker oleh orang, suhu tubuh, usia dan nama orang tertentu. Data dapat dikumpulkan dengan bantuan media sosial dan juga dari beberapa domain [1].
(ii)Pemrosesan Data: Dalam modul pemrosesan data, dilakukan klasifikasi beserta proses pengelompokannya. Algoritma Random Forest (RF) dan Naive Bayes (NB) untuk prediksi hasil. Analisis kinerja dilakukan untuk mengevaluasi kerja sistem. Untuk interpretasi data, algoritma pembelajaran mesin ini diperlukan. Masukan dianalisis untuk menghasilkan keluaran. Klasifikasi dilakukan pada awalnya ketika data diklasifikasikan sebagai "Secure" dan "Insecure". Pengukuran parameter dilakukan dengan bantuan perangkat IoT dan untuk deteksi digunakan RF dan NB. Untuk perbandingan input dan data yang dilatih, teknik K-means clustering digunakan dimana kondisi data tampak serupa [1].
(iii)Algoritma Pembelajaran Mesin: RF dan NB adalah dua algoritma yang digunakan untuk tujuan prediksi. Ekspresi matematika digunakan untuk tujuan prediksi dalam algoritma RF sementara NB menggunakan aturan klasifikasi untuk mengklasifikasikan data secara otomatis [1].



Gambar 3 Tingkat akurasi prediksi dari kedua algoritma

(iv)Hasil: Setelah data dikumpulkan, dikirim untuk diproses diikuti oleh algoritma pembelajaran mesin. Akhirnya mendapatkannya data dari model menunjukkan keamanan yang terkait [1].

2. Naïve Bayes dalam penyintasan berita hoax

Hoax adalah informasi atau berita yang mengandung hal-hal yang belum teridentifikasi atau bukan fakta yang sebenarnya terjadi. Demi mendapatkan keuntungan dan mencapai tujuan pribadi, hoax seringkali sengaja dibuat dan dibagikan sehingga dapat menyebar lebih cepat [2].

Hal pertama yang harus dilakukan adalah pengumpulan data. Pengumpulan data dilakukan dengan cara crawler secara manual (copy-paste) dari portal berita online berbahasa Indonesia seperti Tribunnews, Detik, dan CNN untuk berita netral dan dari website turnbackhoax dan kominfo pada fitur publikasi laporan info hoaks untuk berita hoax. Berita yang digunakan pada penelitian ini berfokus pada isu mengenai kesehatan. Total berita yang diambil sebanyak 287 data yang muncul selama periode Agustus 2019 – Mei 2022. Kemudian dilakukan label encoder secara manual untuk memberikan label yang diklasifikasi kedalam hoax atau valid. Dua value tersebut akan diubah setiap nilai dalam kolomnya menjadi angka yang berurutan, dimana untuk hoax akan dilabeli sebagai 0 dan valid sebagai 1 yang nantinya akan menjadi acuan proses klasifikasi berita hoax untuk diujikan. Setelah pengumpulan data, tahap yang kedua yaitu melakukan *preprocessing* untuk membantu menghapus informasi yang tidak relevan pada data dan dapat mengurangi ukuran data mentahnya. Tahapan yang akan dilakukan diantaranya proses *lower casing* (merubah huruf besar menjadi huruf kecil), *remove punctuation* (menghapus tanda baca), tokenization (membagi kalimat menjadi baris huruf kecil), *stemming* (menghilangkan imbuhan yang ada pada tiap kata) dengan menggunakan library python yaitu Sastrawi yang cocok digunakan untuk menghilangkan imbuhan berbahasa Indonesia, dan yang terakhir adalah *stopwords removal* (menyeleksi dan menghilangkan kata yang memiliki kemunculan tinggi pada data, seperti kata sambung atau kata yang tidak memilki arti) [2].

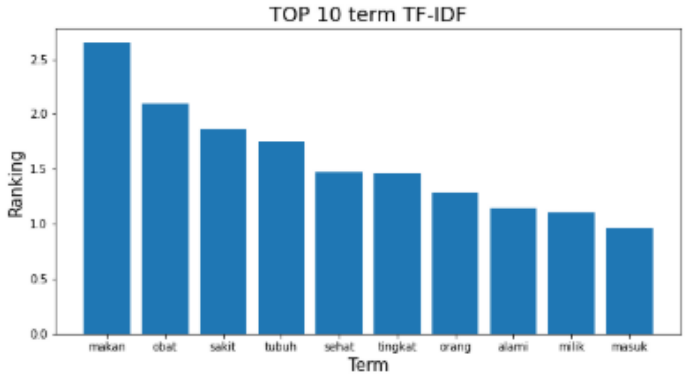
Tabel 1 Hasil Proses Preprocessing

Text Berita	Media covid... Perhatikan bahwa pH virus corona bervariasi dari 5,5 hingga 8,5... karena itu, yang harus kita lakukan untuk menghilangkan virus adalah mengkonsumsi lebih banyak makanan dan minuman alkali diatas tingkat keasaman virus...
Lowering Case	media covid... perhatikan bahwa ph virus corona bervariasi dari 5,5 hingga 8,5... karena itu, yang harus kita lakukan untuk menghilangkan virus adalah mengkonsumsi lebih banyak makanan dan minuman alkali diatas tingkat keasaman virus...
Remove Punctuation	media covid perhatikan bahwa ph virus corona bervariasi dari 5,5 hingga 8,5 karena itu yang harus kita lakukan untuk menghilangkan virus adalah mengkonsumsi lebih banyak makanan dan minuman alkali diatas tingkat keasaman virus
Tokenization	'media', 'covid', 'perhatikan', 'bahwa', 'ph', 'virus', 'corona', 'bervariasi', 'dari', 'hingga', 'karena', 'itu', 'yang', 'harus', 'kita', 'lakukan', 'untuk', 'menghilangkan', 'virus', 'adalah', 'mengkonsumsi', 'lebih', 'banyak', 'makanan', 'dan', 'minuman', 'alkali', 'diatas', 'tingkat', 'keasaman', 'virus'
Normalization	'media', 'covid', 'perhatikan', 'bahwa', 'ph', 'virus', 'corona', 'bervariasi', 'dari', 'hingga', 'karena', 'itu', 'yang', 'harus', 'kita', 'lakukan', 'untuk', 'menghilangkan', 'virus', 'adalah', 'mengkonsumsi', 'lebih', 'banyak', 'makanan', 'dan', 'minuman', 'alkali', 'diatas', 'tingkat', 'keasaman', 'virus'
Stemming	'media', 'covid', 'perhati', 'bahwa', 'ph', 'virus', 'corona', 'variasi', 'dari', 'hingga', 'karena', 'itu', 'yang', 'harus', 'kita', 'laku', 'untuk', 'hilang', 'virus', 'adalah', 'konsumsi', 'lebih', 'banyak', 'makan', 'dan', 'minum', 'alkali', 'diatas', 'tingkat', 'asam', 'virus'
Stopwords Removal	'media', 'covid', 'perhati', 'bahwa', 'ph', 'virus', 'corona', 'variasi', 'laku', 'hilang', 'virus', 'konsumsi', 'makan', 'minum', 'alkali', 'tingkat', 'asam', 'virus'

Dataset pada penelitian ini dibagi menjadi 2 set, *train-set* dan *test-set* untuk dimasukkan kedalam algoritma pengklasifikasian dengan perbandingan *train-set* 80% dan *test-set* 20% dari keseluruhan dataset. Kemudian data *training* dilakukan dan menghasilkan pembelajaran yang nantinya akan digunakan sebagai acuan pada proses testing dengan algoritma *Naïve Bayes Classifier* yang bekerja untuk memastikan apakah data yang diuji telah tepat terhadap data testing. Dari *modeling* tersebut didapatkan hasil kinerja yang tampilkan dalam bentuk visualisasi *confusion matrix* dan classification report yang menunjukkan *precision, recall, f1-score, support* dan *accuracy* [2].

Setelah melakukan preprocessing berupa *lower case, remove punctuation, tokenization, normalization, stemming,* dan *stopwords removal*, proses yang dilakukan selanjutnya yaitu TF- IDF (*Term Frequency – Inverse Document Frequency*). TF-IDF dapat digunakan untuk mengetahui frekuensi dari istilah tertentu yang relatif terhadap sebuah kata dalam kumpulan dokumen dan melihat seberapa umum atau tidak umum sebuah kata yang ada diantara *corpus* (sekumpulan teks yang terstruktur). Pada proses TF-IDF ini menggunakan urutan token berupa *unigram* dalam implementasinya sehingga jumlah token dari TF-IDF hanya satu kata saja. Kemudian mengubah perhitungan series TF-IDF menjadi berbentuk Sparse Matrix dengan matrix size nya yaitu 10

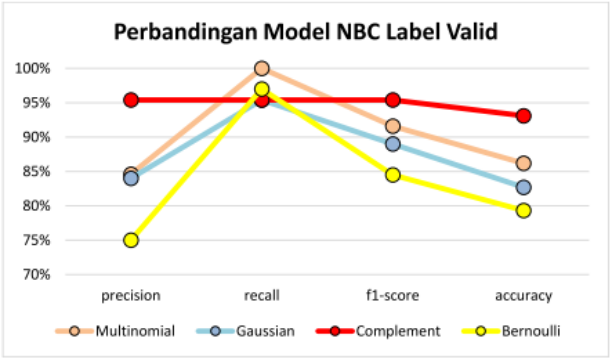
yang artinya top 10 term dengan TF- IDF terbesar. Dengan menggunakan library matplotlib yang tersedia di python untuk memvisualisasikan 10 besar term TF-IDF seperti dalam gambar [2].



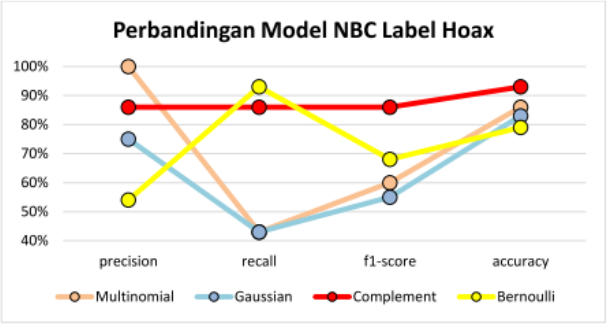
Gambar 4 Grafik Top 10 Term TF-IDF Terbesar

Model yang ada pada Naïve Bayes Classifier digunakan untuk membandingkan hasil dari masing-masing algoritma yang digunakan dengan dibantu menggunakan *library* Scikit Learn yaitu sklearn.naive_bayes untuk melakukan *import* Naive Bayes Classifier (MultinomialNB, GaussianNB, ComplementNB, dan BernoulliNB) pada python dan parameternya yang berupa default [2].

Pada pemodelan ini menggunakan fitur independen X yaitu text dan fitur dependen Y yaitu label. Berikut merupakan perbandingan model Naïve Bayes Classifier dari penelitian ini:



Gambar 5 Grafik perbandingan model NBC label Valid



Gambar 6 Grafik perbandingan model NBC label Hoax

Dalam gambar 3 dan gambar 4 menampilkan grafik tingkat akurasi dari masing-masing model NBC berlabel valid dan hoax menggunakan data testing, menghasilkan akurasi terendah sebesar 79.3% pada model Bernoulli Naïve Bayes dan akurasi tertinggi sebesar 93.1% pada model Complement Naïve Bayes. Naïve Bayes memiliki tingkat precision, recall dan f1-score konstan yang artinya model tersebut mampu mengklasifikasikan berita valid dan hoax dengan baik karena model tersebut sangat cocok untuk bekerja

pada dataset yang tidak seimbang, dimana data pada penelitian ini yang digunakan merupakan dataset imbalance dan karena alasan tersebut, model NBC lainnya memiliki tingkat akurasi yang lebih rendah [2].

3. Peran Naïve Bayes dalam klasifikasi data lainnya

Selain dari yang sudah disebutkan diatas, masih ada hal lain yang bisa dilakukan oleh metode klasifikasi Naive Bayes ini. Salah satunya adalah promosi atau pengiklanan, promosi
adalah salah satu cara paling efektif untuk mempromosikan bisnis, dan kebanyakan orang menyukai promosi. Biasanya pelaku bisnis ini mengumumkan promosinya dengan mengunggah gambar ke media sosial seperti Instagram. Namun, seringkali gambar promosi ini
terkubur di lautan gambar non promosi lainnya. Akan lebih praktis jika komputer dapat digunakan untuk mencari gambar yang berisi penawaran promosi secara otomatis. Oleh karena itu kita membutuhkan sebuah sistem yang dapat mengetahui apakah suatu gambar mengandung informasi tentang penawaran promosi atau tidak secara otomatis tanpa campur tangan manusia, caranya adalah dengan menggunakan Optical Character Recognition (OCR) dan Algoritma Naïve Bayes sebagai classifier.Model Naïve Bayes mencapai presisi 94,31%, penarikan kembali 94,33%, presisi 94,11%, dan skor F1 rata-rata 0,93. Berdasarkan hasil tersebut, dapat disimpulkan bahwa Optical Character Recognition (OCR) dan Algoritma Naïve Bayes cukup cocok untuk masalah ini [7].

Selain dari hal promosi diatas, Naïve Bayes juga sudah digunakan dalam menganalisis sentimen media sosial twitter dengan kasus kampanye Anti-LGBT di Indonesia. Orang menggunakan media sosial sebagai alat untuk mengekspresikan pikiran, minat, dan pendapat mereka tentang berbagai hal. Ribuan kiriman terjadi setiap hari di setiap media sosial. Setiap orang dapat mengutarakan pendapatnya melalui media sosial dengan bebas. Pendapat ini berisi sentimen positif, negatif, dan netral terhadap suatu topik. Studi kasus yang diambil oleh peneliti adalah kampanye Anti-LGBT di Indonesia. Kasus tersebut diambil karena kampanye Anti-LGBT ramai diperbincangkan oleh masyarakat Indonesia di media sosial Twitter. Jika ingin mengetahui kecenderungan komentar masyarakat terhadap kampanye Anti-LGBT di Indonesia apakah positif, negatif atau netral, kemudian dilakukan analisis sentimen. Naive Bayes digunakan dalam analisis sentimen ini, karena memiliki nilai yang tinggi, dan keunggulan akurasi dalam mengklasifikasikan analisis sentimen. Tahapan dalam melakukan analisis sentimen pada penelitian ini adalah preprocessing data, pengolahan data, klasifikasi, dan evaluasi. Analisis sentimen yang diperoleh dalam penelitian ini menunjukkan bahwa pengguna Twitter di Indonesia memberikan komentar yang lebih netral. Pada penelitian ini diperoleh akurasi sebesar 86,43% dari data pengujian menggunakan Naïve Bayes [8].

Tidak hanya itu, klasifikasi Naïve Bayes juga digunakan dalam mendeteksi istopolog dalam data *liquid chromatography high-resolution mass spectrometry* (LC-HRMS). LC-HRMS adalah alat yang ideal untuk screening secara kuantitatif/kualitatif, dengan specifity yang jauh lebih baik dari level triple quadrupole, Q-TOF, atau Q-Trap. Identifikasi atau penghilangan isotopolog merupakan langkah yang diperlukan untuk mengurangi jumlah fitur yang akan diidentifikasi dalam sampel yang dianalisis dengan analisis tidak bertarget. Pendekatan yang tersedia saat ini bergantung pada pola isotop yang diprediksi atau toleransi massa arbitrer, masing-masing memerlukan informasi tentang rumus molekul atau kesalahan instrumental. Oleh karena itu, model klasifikasi isotopolog Naïve Bayes telah dikembangkan yang tidak bergantung pada informasi ambang atau rumus molekul apa pun. Naïve Bayes menggunakan cacat massa unsur dari enam rasio unsur dan berhasil mengidentifikasi isotopologi untuk kedua pola isotop teoretis dan sampel influen air limbah, mengungguli salah satu pendekatan yang paling umum digunakan (yaitu, metode perbedaan massa 1,0033 Da - CAMERA). Untuk isotopologi teoretis, model klasifikasi mengungguli metode perbedaan massa "in-house" dengan tingkat positif sejati (TPr) sebesar 99,0% dan tingkat positif palsu (FPr) sebesar 1,8% dibandingkan dengan TPr sebesar 16,2% dan FPR sebesar 0,02%, dengan asumsi tidak ada kesalahan. Untuk sampel influen air limbah, model klasifikasi, dengan TPr 99,8% dan tingkat deteksi palsu (FDr) 0,5%, kembali tampil lebih baik daripada metode perbedaan massa, dengan TPr 96,3% dan FDr 4,8%. Oleh karena itu, dapat disimpulkan bahwa model klasifikasi dapat digunakan untuk
identifikasi isotopolog, tidak memerlukan
ambang atau informasi tentang rumus molekul
[9].

4. Perbandingan dengan algoritma lain

Perbandingan dilakukan dengan tujuan mencari dan mendapatkan, metode algoritma mana yang lebih efektif. Berbagai algoritma sudah dipilih untuk dibandingkan dengan algoritma Naïve Bayes, diantarnya adalah perbandingan antara Algortima K-Nearest Neighbors dan Naïve Bayes dalam pengklasifikasian data diagnosis dari penyakit diabetes melitus [10], dan perbandingan antara Algoritma Naïve Bayes, Decision Tree, dan Random Forest. untuk Klasifikasi Sentimen Anti-LGBT pada Media Twitter [8].

Hasil dari perbandingan yang pertama, yaitu K-Nearest Neighbors dan Naïve Bayes dalam klasifikasi data diagnosis penyakit diabetes melitus akan dipaparkan disini. Sebelumnya, data yang digunakan yaitu berasal dari website Kaggle.com, dan dataset berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases*. Total data yang digunakan adalah 200 data. Metodologi penelitian yang digunakan adalah Knowledge Discovery in Database (KDD). Dengan dilakukannya pengolahan data dengan cara *selection, preprocessing, transformation, mining, dan evaluation,* maka didapatkan sebuah keluaran (*output*) yaitu yes dan no. Untuk keluaran yes, maka berdasarkan ciri-ciri yang didapatkan dari pengumpulan data (diagnosis) bahwa orang tersebut menderita diabetes. Begitu juga dengan sebaliknya, jika keluaran yang didapatkan adalah no, maka orang tersebut tidak menderita diabetes. Setelah semua data didapatkan, kemudian dilakukan pemodelan dengan menggunakan software Jupyter Notebook dengan bahasa pemrograman Python. Pemodelan pada penelitian ini adalah klasifikasi data dengan algoritma *Naïve Bayes* dan *K-Nearest Neighbor*. Nilai k untuk algoritma KNN adalah 6. Selain itu, data akan dievaluasi dengan menghitung nilai akurasi, recall, dan precision dari algoritma KNN dan Naïve Bayes. Berikut Tabel dari hasil nilai akurasi dari KNN dan Naïve Bayes [10].

Tabel 2 Nilai akurasi

Pembagian Data (Testing – Training)	**Akurasi Naïve Bayes**	**Akurasi KNN**
10 – 90	80%	75%
20 – 80	78%	75%
30 – 70	80%	71%
40 – 60	74%	66%
50 – 50	78%	65%

Selanjutnya melakukan evaluasi dengan menghitung nilai recall dari masing- masing algoritma, berikut adalah tabel dari hasil penghitungan nilai recall.

Tabel 3 Nilai recall

Pembagian Data (Testing – Training)	**Recall Naïve Bayes**	**Recall KNN**
10 – 90	0.86	0.86
20 – 80	0.89	0.89
30 – 70	0.86	0.92
40 – 60	0.82	0.86
50 – 50	0.85	0.85

Dan evaluasi terakhir yang dilakukan adalah menghitung nilai precision dari masing-masing algoritma, hasilnya dapat dilihat pada tabel berikut.

Tabel 4 Nilai precision

Pembagian Data (Testing – Training)	Precision Naïve Bayes	Precision KNN
10 – 90	0.86	0.80
20 – 80	0.81	0.78
30 – 70	0.82	0.75
40 – 60	0.77	0.68
50 – 50	0.80	0.66

Dengan semua perbandingan yang dilakukan antara dua algoritma yaitu KNN dan Naïve Bayes dalam mengklasifikasikan diagnosis penyakit diabetes melitus. Dapat dilihat bahwa nilai akurasi dari Naïve Bayes lebih tinggi dibandingkan KNN. Dimana nilai akurasi yang paling tinggi yang didapatkan dari algoritma Naïve Bayes yaitu sebesar 80%. Sedangkan algoritma KNN nilai akurasi tertinggi yaitu sebesar 75%. Selain itu, diketahui bahwa nilai recall paling tinggi dihasilkan oleh algoritma KNN yaitu sebesar 0.92. Dan untuk nilai presisi lebih tinggi dihasilkan oleh algoritma Naïve Bayes yaitu 0.86. Dengan dilakukan penelitian ini dapat memberikan informasi bermanfaat mengenai penyakit diabetes melitus dan algoritma yang lebih unggul diantara algoritma KNN dan Naïve Bayes sehingga dapat menjadi acuan referensi serta pengembangan ilmu pengetahuan untuk penelitian selanjutnya [10].

Hasil dari perbandingan yang kedua membasas topik tentang LGBT. Masyarakat di dunia pada umumnya dan Indonesia pada khususnya menggunakan media sosial sebagai alat untuk mengungkapkan perasaannya pikiran, minat, dan pendapat tentang berbagai hal. Studi kasus yang diambil peneliti adalah kampanye Anti LGBT. Mengambil studi kasus kampanye Anti-LGBT untuk analisis sentimen karena gerakan ini banyak dibicarakan oleh masyarakat Indonesia di media sosial. Berdasarkan fakta tersebut, dapat disimpulkan bahwa kampanye Anti-LGBT menarik banyak orang untuk diperbincangkan. Hal ini dibuktikan dengan banyaknya komentar di media sosial, khususnya Twitter, oleh masyarakat Indonesia terhadap kampanye Anti-LGBT. Berdasarkan komentar di Twitter, analisis sentimen dapat dilakukan [8].

Salah satu cara untuk melakukan analisis sentimen dapat dilakukan dengan menggunakan data dari media sosial. Ribuan pengiriman terjadi setiap hari di setiap media sosial. Setiap orang dapat mengutarakan pendapatnya melalui media sosial dengan bebas. Opini tersebut mengandung sentimen positif, negatif dan netral terhadap suatu topik. Sentimen positif mengungkapkan pendapat baik tentang konteks, sentimen negatif mengungkapkan pendapat buruk dalam konteks, sedangkan sentimen netral mengungkapkan hal-hal yang tidak mendukung baik atau buruk. Algoritma yang digunakan dalam melakukan analisis sentimen pada penelitian ini adalah Naïve Bayes, Decision Tree, dan Random Forest [8].

Data yang digunakan dalam penelitian ini adalah data dari media sosial Twitter dengan komentar diskusi tentang Anti-LGBT kampanye. Kata kunci yang digunakan menggunakan hashtag yang berada dalam lingkup kasus yang sedang dibahas. Data yang digunakan adalah data relevan yang sudah dibersihkan yaitu sebanyak 3744 komentar. Data diberi label secara manual sebagai sentimen positif, negatif, dan netral. Ada beberapa contoh pelabelan tweet secara manual menurut sentimen yang berbeda [8]. Maka, didapatkan hasil komparasi sebagai berikut :

Dari hasil pertama algoritma Naive Bayes, memperoleh data dengan total 936 komentar/tweet, diperoleh 102 komentar dinyatakan positif, 4 komentar dinyatakan negatif, dan 703 komentar dinyatakan netral. Dari hasil tersebut, maka rata-rata persentase masing-masing Recall, Precision dan F1-Measure antara 56% dan 65%. Pengujian algoritma Naïve Bayes pada penelitian ini menggunakan tool RapidMiner menghasilkan akurasi sebesar 83,43%. Waktu perhitungan algoritma ini sekitar 15 detik [8].

Hasil kedua algoritma Decision Tree, memperoleh data dengan total 936 komentar/tweet, diperoleh 0 komentardinyatakan positif, 0 komentar dinyatakan negatif, dan 775 komentar dinyatakan netral. Dari hasil tersebut, maka rata-rata persentase masing-masing Recall, Precision dan F1-Measure adalah antara 27,66% dan 33,33%. Pengujian algoritma Decision Tree pada penelitian ini menggunakan tool RapidMiner menghasilkan akurasi sebesar 82,91%. Waktu perhitungan algoritma ini adalah sekitar 13 detik [8].

Dan yang terakhir algoritma Random Forest, memperoleh data dengan total 936 komentar/tweet, mendapat 0 komentar dinyatakan positif, 0 komentar dinyatakan negatif, dan 775 komentar dinyatakan netral. Dari hasil tersebut, maka rata-rata persentase masing-masing Recall, Precision dan F1-Measure adalah antara 27,66% dan 33,33%. Pengujian algoritma Random Forest pada penelitian ini menggunakan tool RapidMiner menghasilkan akurasi sebesar 82,91% . Waktu perhitungan algoritma ini sekitar 1,27 menit [8].

Berdasarkan data komentar yang diperoleh dari twitter tentang kampanye Anti LGBT kecenderungan komentar disampaikan berisi komentar netral. Pengguna Twitter netral tentang masalah ini. Dari 936 data pengujian, terdapat 703 komentar dengan sentimen netral, kemudian 102 komentar dengan sentimen positif, dan 4 komentar dengan sentimen negatif. Di sini dapat disimpulkan bahwa pengguna media sosial di Indonesia bersikap netral terhadap kampanye anti-LGBT, namun lebih banyak yang mendukung kampanye Anti-LGBT daripada yang menolak. Berdasarkan hasil akurasi algoritma Naïve Bayes sebesar 83,43% yang lebih tinggi dari akurasi Algoritma Decision Tree dan Algoritma Random Forest, dapat disimpulkan bahwa algoritma Naïve Bayes sangat baik digunakan untuk melakukan analisis sentimen pada kasus kampanye Anti-LGBT di media sosial Twitter [8].

Kesimpulan

Kesimpulan dari beberapa hasil dan pembahasan yang sudah dipaparkan diatas adalah, bahwa metode klasifikasi Naïve Bayes menjadi metode pengklasifikasian data yang paling efisien dan akurat. Terbukti dengan hasil penelitian dan penggunaannya dalam beberapa contoh klasifikasi data yang sudah dilakukan dan perbandingan dengan metode klasifikasi lain, sehingga menjadikan Naïve Bayes sebagai salah satu metode klasifikasi yang lebih unggul dari pada metode lainnya. Manfaatnya, kita bisa mendapatkan hasil data yang akurat dan terpercaya, yang mana data tersebut bisa dijadikan acuan untuk dikemudian hari.

Daftar Rujukan

[1] N. Deepa, J. Sathya Priya, and T. Devi, "Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy," *Mater. Today Proc.*, vol. 62, pp. 4795–4799, Jan. 2022, doi: 10.1016/j.matpr.2022.03.345.

[2] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *J. Masy. Inform.*, vol. 13, no. 2, pp. 85–98, 2022, doi: 10.14710/jmasif.13.2.47983.

[3] M. Sudhakar and K. P. Kaliyamurthie, "Effective prediction of fake news using two machine learning algorithms," *Meas. Sensors*, vol. 24, Dec. 2022, doi: 10.1016/j.measen.2022.100495.

[4] T. A.M and A. Yaqin, "Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter," *InComTech J. Telekomun. dan Komput.*, vol. 12, no. 1, p. 01, 2022, doi: 10.22441/incomtech.v12i1.13642.

[5] M. F. Andriansyah, D. Yusup, and A. Voutama, "Menggunakan Metode Naïve Bayes Berbasis Website Web-Based Expert System of Covid-19 Early Detection Using Naïve Bayes Method," *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 2, pp. 446–455, 2021.

[6] R. Hayami, Y. Fatma, O. T. Antoni, and H. Mukhtar, "Analisa Efektifitas Kebijakan PPKM terhadap Pertumbuhan Kasus COVID-19 Menggunakan Algoritma Naïve Bayes," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1649, 2022, doi: 10.30865/mib.v6i3.4356.

[7] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character

Recognition and Naïve Bayes Classifier," in *Procedia Computer Science*, 2021, vol. 179, pp. 498–506, doi: 10.1016/j.procs.2021.01.033.

[8] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," in *Procedia Computer Science*, 2019, vol. 161, pp. 765–772, doi: 10.1016/j.procs.2019.11.181.

[9] D. van Herwerden, J. W. O'Brien, P. M. Choi, K. V. Thomas, P. J. Schoenmakers, and S. Samanipour, "Naive Bayes classification model for isotopologue detection in LC-HRMS data," *Chemom. Intell. Lab. Syst.*, vol. 223, Apr. 2022, doi: 10.1016/j.chemolab.2022.104515.

[10] N. M. Putry, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *EVOLUSI J. Sains dan Manaj.*, vol. 10, no. 1, 2022, doi: 10.31294/evolusi.v10i1.12514.