ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

SENTIMEN ANALISIS BELAJAR ONLINE DI TWITTER MENGGUNAKAN NAÏVE BAYES

Oktavia Putri Zusrotun[1], Alif Catur Murti[2], Rina Fiati[3]

[1, 2, 3] Program Studi Teknik Informatika, Fakultas Teknik,
Universitas Muria Kudus

email: 201851024@std.umk.ac.id [1], alif.catur@umk.ac.id [2],
rina.fiati@umk.ac.id [3]

Abstrak

Aktivitas belajar *online* kini menjadi pemandangan yang biasa kita jumpai di masa pandemi. Karena penyebaran virus yang cepat, lembaga pendidikan terpaksa mengubah metode pembelajaran yang semula dilakukan secara tatap muka dengan pembelajaran *online*. Pembelajaran *online* memiliki beberapa kelemahan yaitu penggunaan internet memerlukan infrastruktur yang memadai, membutuhkan biaya yang banyak, komunikasi melalui internet memiliki berbagai kendala atau lambat. Dengan kekurangan dan perubahan mendadak seperti ini, menimbulkan pro dan kontra di masyarakat, khususnya bagi para pelaku pendidikan. Media sosial, khususnya *Twitter*, kini menjadi salah satu wadah dimana para siswa dapat secara efektif dan efisien menyuarakan keluhan mereka tentang situasi dan kondisi pendidikan saat ini. Untuk mengetahui pandangan masyarakat terhadap pembelajaran *online*, disini penulis mencoba melakukan analisis sentimen berdasarkan sentimen masyarakat melalui *Twitter*, baik itu pandangan positif, negatif atau netral dengan menggunakan algoritma *Naïve Bayes*. Analisis akan dilakukan menggunakan *Microsoft Excel* dan *RapidMiner* dengan bahasa pemrograman *Python*. Uji model dilakukan dengan menggunakan library *python* yaitu *MultinomialNaiveBayes* dengan akurasi diperoleh sebesar 74,08%. Dalam proses uji model, besarnya data tes diambil 30% dari data training yang dilakukan secara acak. Evaluasi model yang dilakukan pada penelitian ini menggunakan 15 *fold cross validation* dengan hasil akurasi 76,39%.

Kata kunci: Belajar *Online*, *Twitter*, *Naïve Bayes*, *Phyton*, *RapidMiner*

Abstract

Online learning activities are now a common sight we encounter during a pandemic. Due to the rapid spread of the virus, educational institutions were forced to change the method of learning that was originally done face-to-face with online learning. Online learning has several weaknesses, namely the use of the internet requires adequate infrastructure, requires a lot of costs, communication via the internet has various obstacles or is slow. With shortages and sudden changes like this, it creates pros and cons in society, especially for education actors. Social media, especially Twitter, is now a place where students can effectively and efficiently voice their complaints about the current situation and condition of education. To find out the public's view of online learning, here the author tries to do a sentiment analysis based on public sentiment through Twitter, whether it is a positive or negative view using the Naïve Bayes algorithm. The analysis will be carried out using Microsoft Excel and Rapid Miner with the Python programming language. The model test was carried out using the python library, namely MultinominalNB with an accuracy of 74.08%. In the model testing process, the amount of testing data is taken 30% of the training data which is done randomly. The evaluation of the model carried out in this study used 15 fold cross validation with an accuracy of 76.39%.

Keywords : *Online Learning, Twitter, Naïve Bayes, Python, Rapid Miner*

Diterima Redaksi: 29-06-2022 | **Selesai Revisi:** 28-10-2022 | **Diterbitkan Online:** 27-12-2022
DOI: https://doi.org/10.23887/janapati.v11i3.49160

PENDAHULUAN

Dunia, termasuk Indonesia, prihatin dengan penyebaran COVID-19. COVID-19 merupakan virus jenis baru, sehingga banyak pihak yang belum mengetahui atau memahami cara penanganan virus tersebut. Dalam situasi belum ada pengobatan atau vaksin, dunia disibukkan dengan berbagai upaya pencegahan

COVID-19 untuk menahan peningkatan kasus positif. Terkait virus COVID-19, pemerintah telah melakukan beberapa langkah untuk mencegah penyebaran virus tersebut, seperti memblokir penyebaran virus yang sudah masuk zona merah atau melakukan tindakan isolasi fisik untuk mencegah penyebarannya. virus. kontak tubuh. Pemerintah Republik Indonesia juga menerbitkan berbagai protokol kesehatan. Protokol ini diterapkan di seluruh Indonesia oleh pemerintah di bawah kepemimpinan terpusat Kementerian Kesehatan RI (2020).

Peningkatan kasus COVID-19 tidak hanya berdampak pada perekonomian global tetapi juga sektor pendidikan. Ketika kebijakan *physical distancing* yang kemudian menjadi dasar pelaksanaan *home education* menggunakan teknologi informasi tiba-tiba diterapkan, guru, siswa, siswa dan orang tua tidak siap dan terkejut. Pendidik awalnya terkejut dengan kebutuhan untuk mengubah cara mereka mengajar secara langsung, tetapi sekarang semua pelatihan dilakukan secara *online*. Kebijakan menghadirkan pendidikan *online* ke sekolah dan universitas di Indonesia sebagai respon terhadap pandemi COVID-19 yang hampir melanda dunia.

Kementerian Pendidikan dan Kebudayaan Republik Indonesia merekomendasikan pengenalan kursus pembelajaran *online*. Hal itu sesuai dengan Surat Edaran Kementerian Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 tentang Pencegahan Virus Corona (Covid19) di Lembaga Pendidikan dan surat dari Sekretaris Jenderal Kementerian Pendidikan. dan Kebudayaan No.35492/A.A5/HK/2020 (Tanggal 12 Maret 2020 Tentang Pencegahan Penyebaran Corona Virus Disease (Covid19)). Selain mematuhi surat edaran dan imbauan domisili perguruan tinggi di masing-masing pemerintah daerah.[1]

Pesatnya perkembangan dunia teknologi informasi dan komunikasi menjadi solusi dari pembelajaran daring. Hal ini juga tidak lepas dari penyedia layanan yang menyediakan informasi yang berbeda. Informasi menghasilkan tambahan data, di antaranya sebagian besar dalam bentuk data, teks dapat menjadi dalam sumber, adalah sangat potensial hingga ekstrak ditambah secara mendalam. Salah satu dari misalnya adalah data teks diambil dari *Twitter*. *Twitter* adalah jaringan sosial yang berfokus pada komunikasi cepat. Lebih dari 140 juta pengguna aktif memposting lebih dari 400 juta *Tweet* ikonik setiap hari. *Twitter* telah menjadi alat komunikasi penting di semua tingkatan. *Twitter* telah memainkan peran penting di berbagai acara sosial dan politik. *Twitter* menyediakan antarmuka pemrograman aplikasi (API) untuk mengumpulkan data sentimen. Ada dua jenis API yang tersedia: RESTAPI dan *StreamingAPI*. RESTAPI digunakan untuk mengakses status dan timeline pengguna. *Streaming* API digunakan untuk mengakses kata kunci, tagar, ID pengguna, dan lokasi.

Perumusan masalah yang di dapat pada uraian diatas adalah bagaimana opini sentimen masyarakat terhadap belajar *online* yang sedang banyak diperbincangankan di media sosial *Twitter* kemudian mengklasifikasikannya dengan algoritma *Naïve Bayes*.

Tujuan yang ingin dicapai dalam penelitian ini adalah:

a. Untuk mengetahui perspektif masyarakat terhadap belajar *online* berdasarkan sentimen masyarakat melalui *Twitter*.
b. Dapat mengklasifikasikan polarisasi sentimen positif dan negatif mengenai belajar *online.*
c. Mengetahui seberapa besar tingkat akurasi yang didapat dari hasil prediksi klasifikasi algoritma *Naïve Bayes* tentang belajar *online*.
d. Membuat model klasifikasi dengan menggunakan metode *Naïve Bayes* tentang belajar *online.*

Penelitian Terkait

Penelitian terkait digunakan sebagai bahan pertimbangan dalam penelitian yang sedang dilakukan. Berikut penelitian yang digunakan:

Penelitian yang dilakukan oleh [2] Di masa pandemi, pemerintah telah menetapkan bahwa proses pembelajaran akan dilakukan secara *online*, yang berarti semua siswa harus melalui proses yang berbeda dari pembelajaran sebelumnya ketika belajar tatap muka di kelas. Para mahasiswa memiliki sudut pandang yang berbeda terhadap strategi pembelajaran *online* yang dilakukan sehingga mengeluarkan pendapat pribadi melalui *Twitter* dengan memberikan evaluasi yang menurut pengalaman pribadi berlangsung selama proses pembelajaran *online* di kampus tempat mereka belajar. Berdasarkan sentimen siswa terhadap strategi pembelajaran *online* yang disediakan melalui jejaring sosial *Twitter*, penelitian dilakukan dengan menganalisis klasifikasi sentimen menggunakan metode *Bayesian*. Hasil analisis klasifikasi *mood* mahasiswa dengan metode *Bayesian* memberikan hasil negatif atau positif untuk strategi pembelajaran *online* yang diterapkan mahasiswa dalam proses pendidikan

di perguruan tinggi, dimana nilai klasifikasinya adalah presisi = 80%, Recall = 80% dan Akurasi = 80%.

Penelitian yang dilakukan oleh [3] menyimpulkan bahwa data dari WHO menyebutkan bahwa pada minggu kedua November 2020, lebih dari 52 juta orang dinyatakan positif Covid-19 dan 1,2 juta meninggal. Namun, pembelajaran *online* yang awalnya merupakan strategi, menjadi kontroversial karena proses orientasi yang singkat. Pergeseran yang cepat dari pembelajaran tatap muka ke pembelajaran *online* skala besar telah menimbulkan reaksi beragam di masyarakat. Studi yang dilakukan dengan cara mengekstrak teks berbasis dokumen dari *Twitter* dan menganalisisnya menggunakan *Naïve Bayes*. Hasil penelitian menunjukkan bahwa selama periode November 2020, sentimen positif dari pembelajaran *online* adalah 30%, sentimen negatif adalah 69%, dan sentimen netral adalah 1%.

Penelitian yang dilakukan oleh [4] menyimpulkan karena strategi pandemi Covid-19 dengan pembatasan sosial telah memaksa semua institusi untuk menghentikan proses pembelajaran dan menggantinya dengan pembelajaran *online*. Kementerian Pendidikan dan Kebudayaan Republik Indonesia menekankan kebijakan pembelajaran *online* di masa pandemi dengan mengeluarkan kebijakan *Learning From Home*. Penelitian dilakukan dengan tujuan untuk menganalisis opini masyarakat tentang pembelajaran *online* di masa pandemi Covid-19. Penelitian dilakukan dengan menerapkan algoritma *Naïve Bayes* untuk klasifikasi sentimen. Ketidakpuasan masyarakat terhadap pembelajaran *online* menimbulkan persepsi negatif. Analisis sentimen menggunakan data *Twitter* dengan kata kunci “belajar daring” dalam bahasa Indonesia. Hasil dari penelitian menunjukkan 50.75% sentimen positif, 83.3% sentimen negatif.

Penelitian yang dilakukan oleh [5] menyimpulkan pada pekan terakhir September 2020, Covid-19 di Indonesia telah menginfeksi lebih dari 252.000 orang. Virus ini menyebar melalui kontak fisik, sehingga semua negara menggunakan jarak sosial dan jarak fisik untuk mengurangi interaksi. Pandemi Covid-19 berdampak signifikan terhadap cara lembaga pendidikan beroperasi. Semua lembaga pendidikan harus menghentikan proses pembelajaran dan menggantinya dengan pembelajaran *online*. Pembelajaran daring ini ditegaskan oleh Menteri Pendidikan dan Kebudayaan Republik Indonesia, Nadiem Makarim yang selanjutnya mengeluarkan kebijakan pembelajaran daring untuk mencegah penyebaran darurat virus Covid-19. Penelitian dilakukan dengan *text mining* dan sentimen berbasis dokumen pada data *Twitter* yang dianalisis menggunakan metode *Naïve Bayes*. Analisis selama periode Oktober 2020 mengungkapkan 25% sentimen positif, 74% sentimen negatif, dan 1% sentimen netral. Beberapa *tweet* menunjukkan bahwa kata "stres" dan "covid" adalah kata yang paling banyak diucapkan bulan September 2020.

Penelitian yang dilakukan oleh [6] menyimpulkan bahwa pandemi COVID-19 merupakan penyakit yang menyebar di seluruh dunia, termasuk Indonesia. Banyak bidang, termasuk pendidikan, terkena dampak pandemi ini. Indonesia saat ini sedang menerapkan strategi pembelajaran *online* yang menuai banyak opini publik. Analisis sentimen di cabang *text mining* digunakan untuk mengklasifikasikan entitas dalam dokumen teks yang terdiri dari dua kelas, positif dan negatif. Nilai diperoleh dengan mengkategorikan catatan judul dan konten berita yang terkait dengan pembelajaran *online*. Tujuan dari penelitian yang dilakukan adalah untuk memprediksi pendapat orang tua tentang pembelajaran *online* dan menggunakan algoritma pengklasifikasi *Naïve Bayes* untuk menemukan nilai keakuratan pendapat tersebut. Pada artikel ini, kita akan menggunakan metode TF-IDF untuk melakukan pemerataan teks. Hasil pada artikel ini menunjukkan bahwa algoritma *Naïve Bayes* classifier memiliki nilai akurasi 65% berdasarkan nilai 62,5% positif dan 37,5% negatif, 100 berita positif dan 100 berita positif.

Penelitian yang dilakukan oleh [7] menyimpulkan bahwa Covid-19 dinyatakan sebagai pandemi oleh Organisasi Kesehatan Dunia (*World Health Organization*) pada 11 Maret 2020. Oleh karena itu, pemerintah mengimbau masyarakat untuk tidak melakukan aktivitas di luar rumah guna memutus mata rantai penyebaran wabah Covid-19. Kegiatan yang tiba-tiba dihentikan berdampak negatif bagi masyarakat. Ini dimulai dengan penurunan pertumbuhan ekonomi dan menyebabkan keterlambatan dalam penelitian mahasiswa. Pandemi Covid-19 mempengaruhi berbagai perspektif kehidupan manusia saat ini, khususnya di bidang pendidikan. Semoga proses pendidikan dan pembelajaran terus berlanjut. Semua siswa harus belajar *online* atau di rumah secara *online*. Namun karena pembelajaran *online* menimbulkan pro dan kontra dari masyarakat umum, maka diperlukan kajian untuk menganalisis opini publik tentang pembelajaran online dan menggunakan *Twitter* sebagai sumber data penjelajah untuk

mengetahui efektifitas pembelajaran *online* di masa pandemi Covid-19. Berdasarkan hasil analisis sentimen masyarakat terhadap pembelajaran *online* menggunakan algoritma *Naïve Bayes classifier* menggunakan *Rapidminer* sebagai perangkat lunak yang digunakan untuk pengolahan data, akurasi 60,00%, akurasi 65,67%, dan recall 53,30%.

Penelitian ini mengimplementasikan metode *Naive Bayes Classifier* dalam sistem untuk mengklasifikasikan data menjadi perasaan positif dan perasaan negatif berdasarkan data, ulasan, komentar mengenai belajar *online* yang dikumpulkan dari *Twitter*. Klasifikasi yang digunakan menggunakan dua kelas yaitu sentimen positif dan negatif. Sehingga diharapkan dapat menghasilkan rancangan *prototype analytical* data dengan metode *Naïve Bayes*. Algoritma *Naïve Bayes* digunakan untuk mengklasifikasi berdasarkan komentar di *Twitter* dan juga melalui klasifikasi komentar pada penelitian ini diharapkan kita dapat melihat seberapa besar penyebaran atau pengaruh belajar *online* melalui sosial media *Twitter*. Dalam data *training*, *tweet* yang berisi opini tentang pembelajaran *online* ditandai sebagai positif dan *tweet* yang tidak berisi opini tentang pembelajaran *online* ditandai sebagai negatif. Analisis ini dilakukan dengan *Microsoft Excel* dan *Rapidminer* menggunakan bahasa pemrograman *Python*. Hasil yang diharapkan dari survei ini adalah untuk mengetahui opini publik tentang pembelajaran *online* berbasis opini publik melalui analisis sentimen di jejaring sosial dari *Twitter* menggunakan algoritma *Naive Bayes*.Peneliti memberikan nilai akurasi sebesar 74%.

Studi Literatur

A. Pembelajaran *Online*

Kegiatan pembelajaran *online* ini dilakukan untuk menggantikan kegiatan pembelajaran tatap muka. Pembelajaran *online* memiliki beberapa kelemahan yaitu penggunaan jaringan internet yang membutuhkan infrastruktur yang memadai dan biaya yang tidak sedikit. Walaupun terdapat berbagai kendala, dapat dikatakan pembelajaran *online* efektif apabila siswa dapat mencapai tujuan pembelajaran dan aktif berinteraksi dengan guru atau dosen.[8]

B. *Twitter*

Jejaring sosial adalah sarana di mana pengguna dapat mengekspresikan diri. *Twitter* sangat populer dan tersebar luas di Indonesia, pengguna *Twitter* dapat berekspresi dan mencoba sendiri tanpa batasan. Pengguna *Twitter* dapat menggunakan *tweet* dalam bentuk ucapan dan keinginan.[9]

C. *Scraping*

Scraping adalah teknik ekstraksi data *Twitter* yang menyediakan antarmuka program aplikasi (API) yang digunakan untuk mengakses informasi yang terkandung di dalamnya. *Scraping* dilakukan dengan *tools Python* dan *RapidMiner*.[10]

D. *Preprocessing*

Preprocessing menjadi proses awal sebelum melakukan klasifikasi. Proses ini digunakan untuk membersihkan data dari *noise* dan siap untuk digunakan pada proses selanjutnya, berikut merupakan tahap-tahap *preprocessing*:

a. *Cleaning* pada proses ini kalimat dibersihkan dari *hashtag*, *mention*, dan juga tanda baca.
b. *Case Folding* merupakan proses penggantian huruf dari huruf yang bercampur (*lowercase* dan *uppercase*) menjadi semua huruf kecil.
c. *Tokenization* ialah proses perubahan kalimat menjadi kata.
d. *Stopword* merupakan penghilangan kata yang sering berbobot namun sering muncul.
e. *Stemming* adalah proses mengubah kata menjadi bentuk sederhana. *Stemming* dapat mengurangi semua variasi dari satu kata menjadi bentuk sederhana yang sama.
f. Pembobotan kata atau *labeling* dilakukan dengan cara menghitung bobot kata dalam dokumen. *Weighting word* adalah [11]

E. Analisis sentimen

Analisis sentimen adalah teknik penggalian informasi berupa sikap seseorang terhadap suatu topik atau peristiwa dengan mengklasifikasikan polaritas suatu teks. Pengelompokan dilakukan untuk melihat apakah teks tersebut positif, negatif, atau netral. Analisis sentimen dapat digunakan untuk menentukan opini publik tentang suatu topik. Penulis mengevaluasi data yang diambil dari *Twitter* dengan bantuan metode *web scraping. Web scraping* adalah teknik mengekstrak data dan informasi dari sebuah situs web dan kemudian menyimpannya dalam format tertentu.[12]

F. Klasifikasi *Bayesian*

Klasifikasi *Bayesian* adalah klasifikasi statistik yang dapat digunakan untuk memprediksi kemungkinan keanggotaan kelas. Klasifikasi *Bayesian* didasarkan pada teorema *Bayes*, yang memiliki kemampuan klasifikasi yang mirip dengan pohon keputusan dan jaringan saraf. Klasifikasi *Bayesian* telah terbukti memiliki akurasi dan kecepatan tinggi ketika diterapkan pada database dengan jumlah data yang besar. Metode *Bayes* adalah pendekatan statistik untuk menarik inferensi induktif dalam masalah klasifikasi. Pertama, konsep dasar dan

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

definisi dalam teorema *Bayes* dibahas, kemudian teorema ini digunakan untuk melakukan klasifikasi dalam *data mining* [13].

METODE

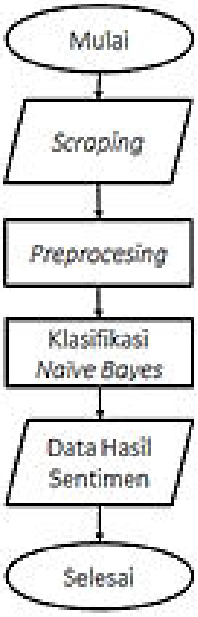
Penelitian ini mengembangkan analisis sentimen untuk mengklasifikasikan *tweet* yang dibuat oleh pengguna *Twitter* sebagai positif, negatif atau netral. Metode yang digunakan untuk menganalisis sentimen dari *tweet* belajar *online* adalah metode *Naïve Bayes*.

A. Metode Pengumpulan Data

Data yang diperoleh untuk melakukan penelitian ini adalah melalui *scraping* yaitu data yang diambil langsung dari *tweet* menggunakan *Twitter* API. Pengumpulan data dilakukan dengan menggunakan kata kunci untuk belajar *online*. Jumlah data yang diperoleh sekitar 1000 data.

B. Perancangan Program Analisis

Diagram alir secara umum pada program dapat dilihat pada Gambar 1 diagram alir

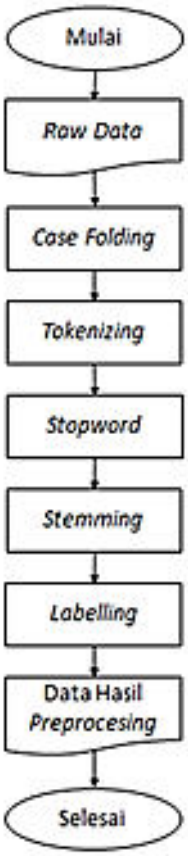


berikut:

Gambar 1. Diagram Alir Program

Gambar 1 menjelaskan bahwa alur program dimulai dengan *scraping* data dari *Twitter* menggunakan *Python* dan *RapidMiner*. Kemudian data tersebut diolah dengan melakukan *preprocessing*. Pada tahap *preprocessing*, data diolah menggunakan *Microsoft Excel*. Selanjutnya masuk ke tahap klasifikasi *Naïve Bayes* dimana hasil sentimen akan positif, negatif atau netral.

1) *Preprocessing* Dilakukan dengan mengambil data dari hasil *scraping* dan kemudian diolah terlebih dahulu sebelum dapat diklasifikasikan. *Flowchart* fase *preprocessing* terlihat pada diagram alir berikut:



Gambar 2. Diagram Alir Tahap *Preprocessing*

Pada Gambar 2 dijelaskan bahwa alur dari langkah *preprocessing* dimulai dari data mentah, kemudian dilakukan langkah *folding case* yaitu proses penghilangan tanda baca. Setelah itu ada fase *tokenization*, yaitu proses pemisahan teks menjadi potongan-potongan (token) untuk dianalisis. Selanjutnya adalah fase *stopword*, yaitu proses perolehan kata-kata penting dari hasil token. Jadi langkah *stemming* adalah proses menghilangkan kata yang dibubuhkan pada bentuk dasarnya. Selanjutnya adalah proses pembobotan atau pelabelan kata, yang dilakukan dengan menghitung bobot kata (*term*) menggunakan perhitungan TF-IDF di Excel. Persamaan *TF-IDF* (*Term Frequency-Inverse Document Frequency*) dapat dilihat pada Persamaan 1:

$$\mathbf{wi,j} = \mathbf{tfi,j\ x\ log}\ \frac{\mathbf{N}}{\mathbf{dfi}} \quad \textbf{(1)}$$

Keterangan:

$tf_{i,j}$: nilai kemunculan dari token i (kata) dalam dokumen j

df_i : nilai dokumen yang memuat token i

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

N : nilai total dokumen

Persamaan 1 menjelaskan bahwa rumus perhitungan *TF-IDF* menghitung bobot kata (*term*) dari data. *Term frequency* menunjukkan seberapa sering (tingkat frekuensi) suatu istilah muncul dalam dokumen. Sedangkan *document frequency* adalah jumlah dokumen di mana istilah muncul.

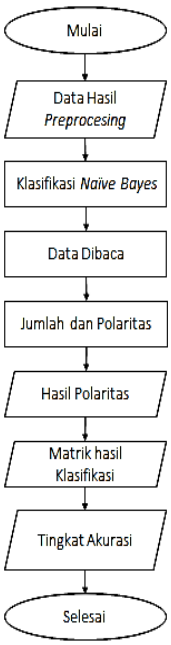
2) Klasifikasi *Naïve Bayes* umumnya digunakan untuk melakukan prediksi sentimen yang muncul pada data yang masih belum memiliki sentimen. Metode *Naïve Bayes* dapat mengelola data dalam jumlah besar dengan hasil akurasi yang tinggi.[14] Persamaan umum yang dimiliki pada algoritma klasifikasi *Naïve Bayes* dapat dilihat pada Persamaan 2:

$$\mathbf{P(H|X)} = \frac{P(X|H) \times P(H)}{P(X)} \quad \textbf{(2)}$$

Keterangan:

H : Hipotesis data suatu class spesifik.
X : Data dengan kelas yang masih belum diketahui.
P(H) : Probabilitas H.
P(X) : Probabilitas X.
P(H|X) : Probabilitas hipotesis berdasar kondisi.
P(X|H) : Probabilitas berdasar kondisi pada hipotesis.

Setelah data dilakukan *preprocessing* maka akan dihasilkan nilai polaritas data dan akurasi klasifikasi, serta matriks sentimen. Adapun *flowchart* proses klasifikasi dapat dilihat pada Gambar 3:



Gambar 3. Diagram Alir Proses Klasifikasi

Gambar 3. menggambarkan aliran data yang setelah dilakukan *preprocessing* kemudian dilakukan pada tahap klasifikasi menggunakan algoritma *Naïve Bayes*. Setelah data terbaca, akan didapatkan bilangan polaritas, hasil polaritas, matriks hasil klasifikasi dan tingkat akurasi.

Dalam proses klasifikasi yang berjalan akan menghasilkan nilai kinerja. Ada beberapa indikator untuk menghitung nilai manfaat. Indikator tersebut meliputi nilai presisi, *recall* dan *F1-Score*. Perhitungan nilai presisi dapat dilihat pada Persamaan 3:

$$\boldsymbol{Precision} = \frac{PB}{PB+PS} \quad \textbf{(3)}$$

Keterangan:
PB : Positif Benar
PS : Positif Salah

Persamaan 3 menjelaskan bahwa perhitungan nilai *precision* menunjukkan nilai. Ketelitian dari klasifikasi yang telah dilakukan. Sedangkan perhitungan nilai *recall* dapat dilihat pada Persamaan 4:

$$\mathbf{Recall} = \frac{\mathbf{PB}}{PB+NS+NeS} \quad \textbf{(4)}$$

Keterangan:
PB : Positif Benar
NS : Negatif Salah
NeS : Netral Salah

Persamaan 4 menjelaskan bahwa perhitungan nilai *recall* menunjukkan hasil ketelitian klasifikasi yang benar. Selanjutnya yaitu perhitungan nilai *F1-Score* dapat dilihat pada Persamaan 5:

$$\mathbf{F1 - Score} = \frac{\mathbf{2 * Precision * Recall}}{\mathbf{Precision + Recall}} \quad \textbf{(5)}$$

Persamaan 5 tersebut menjelaskan bahwa nilai dari *f1-score* yang didapat menunjukkan nilai kinerja dari algoritma *Naïve Bayes*. Sedangkan *macro average* atau biasa disebut dengan rata-rata makro dapat dilihat pada Persamaan 6:

$$\mathbf{Macro\ Average} = \mathbf{matrix\ setiap\ variabel} \quad \textbf{(6)}$$

Persamaan 6 tersebut menjelaskan persamaan dari perhitungan *macro average* menunjukkan bahwa perhitungan rata-rata makro menghitung matriks dari setiap label. *Macro average* mungkin adalah metode yang paling mudah diantara banyaknya metode untuk menemukan nilai rata-rata. *Macro f1-score* dihitung menggunakan rata-rata aritmatika (*unweighted mean*) dari semua kelas *f1-score*. Metode ini memperlakukan semua kelas secara

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

setara terlepas dari nilai *support*nya. Selanjutnya yaitu perhitungan *weighted average* atau disebut dengan rata-rata tertimbang dapat dilihat pada Persamaan 7:

$$\textbf{Weighted Average} = \textbf{jumlah instance setiap variabel} \quad (7)$$

Persamaan 7 tersebut menjelaskan bahwa perhitungan *weighted average* atau disebut dengan rata-rata tertimbang juga merupakan *f1-score* yang memiliki nilai rata-rata tertimbang yang mirip terhadap variabel. *Weigted average f1 score* dihitung dengan mengambil rata-rata semua kelas *f1-score* sambil mempertimbangkan nilai *support* masing-masing kelas. *Support* mengacu pada jumlah kejadian aktual kelas dalam *dataset* (kumpulan data). ”Bobot” pada dasarnya mengacu pada proposi dukungan masing-masing kelas relatif terhadap jumlah semua nilai *support*. Dengan *weighted averaging*, rata-rata keluaran akan memperhitungkan kontribusi setiap kelas yang dibobot dengan jumlah contoh kelas yang diberikan. Berikutnya akan dilakukan perhitungan yang menghasilkan nilai akurasi dari klasifikasi, selain dari keempat indikator yang telah dijelaskan tersebut. Persamaan perhitungan akurasi dapat dilihat pada Persamaan 8:

$$Akurasi = \frac{NB + NeB + PB}{NB + NS + NeB + NeS + PB + PS} \quad (8)$$

Keterangan:
NB : Negatif Benar
NS : Negatif Salah
NeB : Netral Benar
NeS : Netral Salah
PB : Positif Benar
PS : Positif Salah

Persamaan 8 menjelaskan bahwa perhitungan akurasi akan menghitung nilai dari klasifikasi yang dilakukan. Selain akurasi, klasifikasi juga akan menghasilkan validasi data. Hasil validasi data dapat dilihat pada *confusion matrix*. *Confusion matrix* memberikan hasil data yang positif ditebak benar, positif ditebak salah, negatif ditebak benar, negatif ditebak salah, netral ditebak benar, dan netral ditebak salah. *Confusion matrix* dapat dilihat pada Gambar 4:

	Negatif	**Netral**	**Positif**
Negatif	**NB**	**NS**	**NS**
Netral	**NeS**	**NeB**	**NeS**
Positif	**PS**	**PS**	**PB**

Gambar 4. *Confusion Matrix*

Pada gambar 4. dijelaskan bahwa bentuk dari *confusion matrix* memberikan prediksi dari hasil klasifikasi benar ditunjukkan pada variabel NB (Negatif Benar), NeB (Netral Benar), dan PB (Positif Benar). Sedangkan prediksi dengan hasil yang salah diunjukkan oleh NS (Negatif Salah), NeS (Netral Salah), dan PS (Positif Salah).[15]

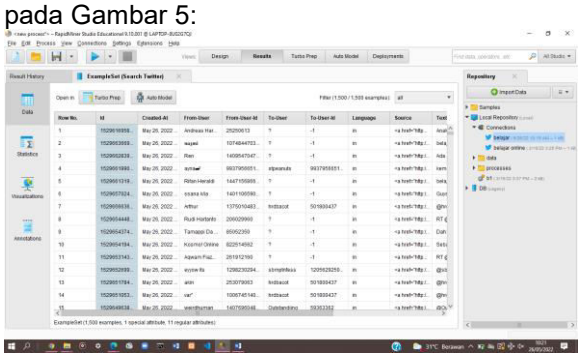
HASIL DAN PEMBAHASAN

1) *Scraping* adalah teknik mengekstrak data dan informasi dari sebuah situs web dan kemudian menyimpannya dalam format tertentu. Dengan cara mengambil data dari *Twitter* yang menyediakan API.[10] *Scraping* dilakukan dengan menggunakan *tools* yaitu *Python* dan *RapidMiner*. *Keyword* untuk analisa data adalah “Belajar *Online*”. Dengan *scraping* didapatkan didapatkan data sejumlah 1186 data. Hasil data *scraping* dapat dilihat pada Tabel 1:

Tabel 1. Data Hasil Scraping

Created-At	From-User	Text
2022-05-25 12:33:01	Boraeonni	Sekarang aku ga perlu takut jaringan leletketika belajar online. Semenjak pakai #PinterMilihTemen smartfreen internetan juga makin gampang, tanpa gangguan. #UnlimitedBebasWorry deh pokoknya. https://t.co/RKDM9g8iiH

Scraping dilakukan dengan menggunakan *tools RapidMiner* dengan bahasa pemrograman *Python.* Hasil dari proses *scraping* dapat dilihat pada Gambar 5:



Gambar 5. Data Hasil *Scraping*

Pada gambar 5 tersebut menjelaskan bahwa pengambilan data atau proses *scraping* pada media sosial *twitter.* Kemudian data tersebut disimpan ke dalam *excel*.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

2) *Preprocessing* menjadi proses awal sebelum melakukan klasifikasi. Proses ini digunakan untuk membersihkan data dari *noise* dan siap untuk digunakan pada proses selanjutnya. Langkah-langkah *preprocessing* yang dilakukan disesuaikan berdasarkan kondisi dari data komentar mengenai "Belajar *Online*".Proses *preprocessing* dapat dilihat pada Tabel 2:

Tabel 2. Proses *Preprocessing*

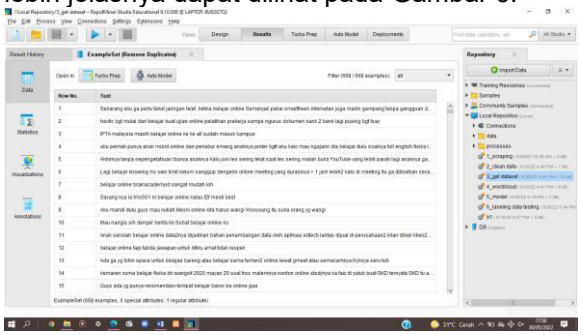
Dataset Awal	Sekarang aku ga perlu takut jaringan lelet ketika belajar online. Semenjak pakai #PinterMilihTemen smartfreen internetan juga makin gampang, tanpa gangguan. #UnlimitedBebasWorry deh pokoknya. https://t.co/RKDM9g8iiH
Dataset Cleansing	Sekarang aku ga perlu takut jaringan lelet ketika belajar online Semenjak pakai smartfreen internetan juga makin gampang tanpa gangguan deh pokoknya
Dataset Case Folding	sekarang aku ga perlu takut jaringan lelet ketika belajar online semenjak pakai smartfreen internetan juga makin gampang tanpa gangguan deh pokoknya
Dataset Tokenizing	['sekarang', 'aku', 'ga', 'perlu', 'takut', 'jaringan', 'lelet', 'ketika', 'belajar', 'online', 'semenjak', 'pakai', 'smartfreen', 'internetan', 'juga', 'makin', 'gampang', 'tanpa', 'gangguan', 'deh', 'pokoknya']
Dataset Stopword	['sekarang', 'aku', 'ga', 'perlu', 'takut', 'jaringan', 'lelet', 'belajar', 'online', 'semenjak', 'pakai', 'smartfreen', 'internetan', 'makin', 'gampang', 'gangguan', 'deh', 'pokoknya']
Dataset Stemming	sekarang aku ga perlu takut jaringan lelet belajar online semenjak pakai smartfreen internetan makin gampang gangguan deh pokoknya
Data Akhir Preprocessing	takut jaringan lelet belajar online semenjak pakai smartfreen internetan gampang gangguan pokoknya

Hasil akhir dari data *preprocessing* secara keseluruhan dapat dilihat pada Tabel 3:

Tabel 3. Data *Preprocessing*

Dataset Awal	Sekarang aku ga perlu takut jaringan lelet ketika belajar online. Semenjak pakai #PinterMilihTemen smartfreen internetan juga makin gampang, tanpa gangguan. #UnlimitedBebasWorry deh pokoknya. https://t.co/RKDM9g8iiH
Data Akhir Preprocessing	takut jaringan lelet belajar online semenjak pakai smartfreen internetan gampang gangguan pokoknya

Sama seperti proses *scraping*, pada proses *preprocessing file* hasil dari *preprocessing* yang telah dilakukan disimpan ke dalam excel. Untuk lebih jelasnya dapat dilihat pada Gambar 6:



Gambar 6. Data Hasil *Preprocessing*

3) Ekstraksi Fitur Setelah terbentuknya *file* yang akan dijadikan *dataset*, maka selanjutnya data tersebut akan dibentuk menjadi sebuah model klasifikasi. Namun sebelum membentuk model, ada beberapa tahapan yang harus dilakukan agar terbentuknya suatu model yang baik. Yang pertama dilakukan adalah membaca *file* xlsx dan kemudian dilakukan tokenisasi terhadap seluruh dokumen dalam *file* tersebut. Berdasarkan hasil tokenisasi yang dilakukan, maka penulis juga ingin mengetahui frekuensi kata yang banyak diperbincangkan oleh pengguna *twitter*, untuk itu penulis memvisualisasikannya dalam bentuk *wordcloud* pada Gambar 7:

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

Gambar 7. Visualisasi Kata Terpopuler dengan *Wordcloud*

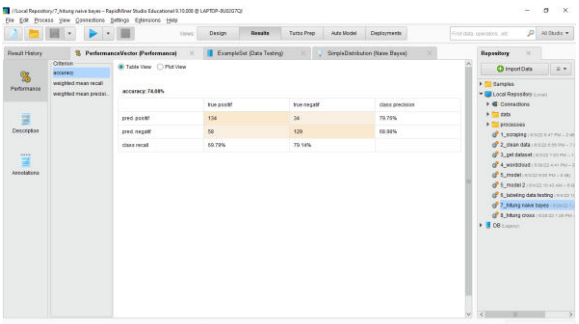
4) Klasifikasi algoritma *Naïve Bayes* yang dilakukan dalam analisa ini menggunakan perbandingan 70:30 antara data *training* dengan data *testing*. Dengan melakukan *training* pada *dataset* yang ada akan menghasilkan sebuah model pada data *testing*. Data yang telah selesai dilakukan klasifikasi akan menghasilkan polaritas dan sentimen analisis seperti pada Tabel 4:

Tabel 4. Data Hasil Klasifikasi

Sentimen	**Polaritas**
Positif	597
Negatif	589

Penelitian ini menggunakan *tools Rapid Miner studio* 9.10.1.0. Perhitungan akurasi menggunakan *RapidMiner* dilakukan dengan metode *Naive Bayes* dan *Cross Validation* dapat dilihat pada Gambar 8:

a. Perhitungan Akurasi dengan *Naive Bayes*



Gambar 8. Hasil Pengukuran Evaluasi Performa

Perhitungan akurasi klasifikasi menggunakan metode *Naive bayes* dengan fitur TF-IDF diperoleh sebesar 74,08%. Hasil presisi, *recall* dan f1- *score* di setiap kelasnya dapat dilihat pada Tabel 5:

Tabel 5. Nilai presisi, *recall* dan f1- *score*

Jenis Klasifikasi	Presisi	Recall	f1-score
Positif	79,76%	69,79%	74,44%
Negatif	68,98%	79,14%	73,71%

Hasil dari evaluasi model dapat dilihat nilai presisi dan *recall* di setiap kelasnya dapat dikatakan tingkat kemampuan sistem dalam mencari ketepatan antara informasi yang diminta oleh pengguna untuk kelas positif sebesar **79,76%**, untuk kelas negatif sebesar **68,98%**. Sedangkan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi untuk kelas positif sebesar **69,79%**, untuk kelas negatif sebesar **79,14%.** Artinya kinerja sistem keberhasilan sistem dalam menemukan kembali sebuah informasi yang bernilai positif negatif dalam dokumen sangatlah rendah.

b. Perhitungan Akurasi dengan *Cross Validation*

Untuk menentukan hasil uji dan evaluasi yang maksimal dilakukan pengujian kedua dengan menggunakan *k-fold cross validation*. Dimana jumlah nilai k yang digunakan pada penelitian ini yaitu 15. Hal ini dikarenakan nilai tertinggi yang didapatkan untuk mencapai hasil akurasi yang optimal dibutuhkan 15 *fold*. Dalam 15 *fold CV*, data dibagi menjadi 15 *fold* berukuran kira-kira sama, sehingga memiliki 15 *subset* data untuk mengevaluasi kinerja model atau algoritma.

Gambar 9. Evaluasi Model dengan *Cross Validation*

Gambar 9 memperlihatkan performansi akurasi metode *Cross Validation* dengan menggunakan 15 *fold cross validation* dengan fitur TF-IDF yaitu sebesar 76,39%. Untuk hasil presisi, *recall* dan f1-*score* di setiap kelasnya dapat dilihat pada Tabel 6:

Tabel 6. Nilai Presisi, *Recall* dan f1-*Score* Evaluasi Model

Jenis Klasifikasi	Presisi	Recall	f1-score
Positif	81,95%	72,23%	76,78%
Negatif	71,34%	81,28%	75,99%

Hasil dari evaluasi model dapat dilihat nilai presisi dan recall di setiap kelasnya dapat dikatakan tingkat kemampuan sistem dalam

mencari ketepatan antara informasi yang diminta oleh pengguna untuk kelas positif sebesar **81,95%**, untuk kelas negatif sebesar **71,34%**. Sedangkan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi untuk kelas positif sebesar **72,23%**, untuk kelas negatif sebesar **81,28%**.

KESIMPULAN

Berdasarkan hasil pengujian algoritma *Naive Bayes Classifier* yang telah dilakukan ada beberapa hal yang dihasilkan, antara lain yaitu pada penelitian ini, algoritma *Naive Bayes Classifier* terbukti algoritma yang akurat karena menghasilkan nilai akurasi sebesar 74,08%. Untuk memastikan hasil penelitian ini, dilakukan juga pengujian dengan *K-Fold Cross Validation* dengan k sebesar 15 yang hasil nilai akurasinya sebesar 76,39%. Saran yang dapat diberikan dalam penelitian ini adalah proses *labeling* atau pelabelan pada tahap *preprocessing* sebenarnya sudah dilakukan secara otomatis oleh *rapid miner*, namun hasil yang didapatkan belum sepenuhnya sesuai dengan harapan peneliti terbukti dari hasil akurasi yang nilainya cukup rendah. Kedepannya kami berharap hal ini bisa dieksplorasi lebih jauh untuk melakukan proses pelabelan secara otomatis dan tepat.

REFERENSI

[1] Kemendikbud, “Surat Edaran Nomor 1 Tahun 2020 tentang Pencegahan Penyebaran Corona Virus Disease (Covid019) Di Perguruan Tinggi, Kementrian Pendidikan dan Kebudayaan,” 2020.

[2] A. Syahputri dan M. Zarlis, “Analisis Klasifikasi Sentimen Mahasiswa Terhadap Strategi Pembelajaran Online Pada Media Sosial Twitter Menerapkan Metode Naïve Bayes,” vol. 4, no. 1, 2020, doi: 10.30865/komik.v4i1.2567.

[3] S. Samsir, A. Ambiyar, U. Verawardina, F. Edi, dan R. Watrianthos, “Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes,” *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 1, hlm. 157, Jan 2021, doi: 10.30865/mib.v5i1.2580.

[4] S. Yana Nursyi’ah, A. Erfina, dan C. Warman, “ANALISIS SENTIMEN PEMBELAJARAN DARING PADA MASA PANDEMI COVID-19 DI TWITTER MENGGUNAKAN ALGORITMA NAÏVE BAYES,” 2021.

[5] S. H. Sahir, R. S. Ayu Ramadhana, M. F. Romadhon Marpaung, S. R. Munthe, dan R. Watrianthos, “Online learning sentiment analysis during the covid-19 Indonesia pandemic using twitter data,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1156, no. 1, hlm. 012011, Jun 2021, doi: 10.1088/1757-899x/1156/1/012011.

[6] F. Sidik, I. Suhada, A. Haikal Anwar, dan F. Noor Hasan, “Analisis Sentimen Terhadap Pembelajaran Daring dengan Algoritma Naïve Bayes Classifier,” *J. Linguist. Komputasional JLK*, vol. 5, no. 1, hlm. 34–43, 2022.

[7] R. Pinka *dkk.*, “SENTIMENT ANALYSIS OF PUBLIC OPINIONS ON THE EFFECTIVENESS OF ONLINE LEARNING USING NAÏVE BAYES ALGORITHM,” *J. Inf. Syst. Inform. Comput. Issue Period*, vol. 6, no. 1, hlm. 273–279, 2022, doi: 10.52362/jisicom.v6i1.822.

[8] Hermanto dan A. Noviriandini, “ANALISA SENTIMEN TERHADAP BELAJAR ONLINE PADA MASA COVID-19 MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE BERBASIS PARTICLE SARM OPTIMAZATION,” *J. Inform. KaputamaJIK*, vol. V, no. 1, hlm. 129–136, 2021.

[9] D. Rustiana dan N. Rahayu, “ANALISIS SENTIMEN PASAR OTOMOTIF MOBIL: TWEET TWITTER MENGGUNAKAN NAÏVE BAYES,” *J. SIMETRIS*, vol. VIII, no. 1, hlm. 113–120, 2017.

[10] M. Y. Aldean, M. D. Hilmawan, R. Indriyati, J. Lasama, dan A. Junaidi, “Analisa Relevansi Tweet terhadap Hashtag dengan Metode Logistic Regression,” *Conf. Electr. Eng. Telemat. Ind. Technol. Creat. 2019*, hlm. 25–38, 2019.

[11] M. Syarifuddin, “ANALISIS SENTIMEN OPINI PUBLIK TERHADAP EFEK PSBB PADA TWITTER DENGAN ALGORITMA DECISION TREE-KNN-NAÏVE BAYES,” *INTI Nusa Mandiri*, vol. XV, no. 1, hlm. 87–94, 2020, doi: 10.33480/inti.v15i1.1433.

[12] M. I. Aditama, R. I. Pratama, K. U. W. Hafizzana, dan N. A. Rakhmawati, “Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19,” *JIEET J. Inf. Eng. Educ. Technol.*, vol. VI, no. 2, hlm. 90–92, 2020.

[13] H. Annur, “Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes,” *Ilk. J. Ilm.*, vol. 10, no. 2, hlm. 160–165, Agu 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[14] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” *Integer J.*, vol. II, no. 1, hlm. 32–41, 2017.
[15] F. Ratnawati, “Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter,” *J. INOVTEK POLBENG -SERI Inform.*, vol. III, no. 1, hlm. 50–59, 2018.