KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 2, No 4, Februari 2022
Hal 148-155
https://djournals.com/klik

Analisa sentimen Pengguna Instagram Di Indonesia Pada Review Smartphone Menggunakan Naive Bayes

Kairil Anwar

Program Studi Teknik Informatika, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Budi Darma,
Jalan Sisingamanganraja No. 338,Medan, Sumatera Utara, Indonesia
Email: nuar30568@gmail.com

Abstrak-Semakin majunya teknologi informasi dan taraf hidup masyarakat mengakibatkan semakin meningkatnya tuntutan masyarakat terhadap kualitas pelayanan dan produk yang digunakan. Keperluan smartphone ini telah menjadi gaya hidup yang dianggap penting bagi sebagian masyarakat saat ini.sebuah Fenomena tersebut mendukung munculnya banyak sebuah smartphone yang menawarkan produk untuk memenuhi keperluan masyarakat akan teknologi dalam hal berkomunikasi. Pada penelitihan ini merek smartphone yang digunakan adalah vivo_Indonesia, oppo indonesia. Mechine learning adalah aplikasi kecerahan buatan (AI) yang menyediakan kemampuan sistem untuk belajar dan meningkatakn kemampuannya secara otomatis dari pengalaman tanpa harus diprogram secara eksplisit. Mechine learning memberi cara-cara baru dalam menggali wawasan serta membantu badan penelitian memecahkan masalah-masalah. Salah satu contohnya adalah analisa sentimen yang dilakukan secara otomatis. Analisa sentimen sendiri perlu dilakuakn kerna pengguna media sosial di masyarakat semakin meningkat sehingga mempengaruhi perkembangan opini publik. Oleh kerena itu hal ini dapat dimanfaatkan untuk menganalisa opini publik tersebut dengan mengaplikasikan data sciene, salah satunya adalah Text Mining atau dikenal dengan istilah text analytics. Tahapan keseluruhan metode yang digunakan pada penelitian ini adalah dengan menggunakan Text Mining pada video di instagram mengenai review smarphone dengan metode scraping, labelisasi, preproseccing (case folding, tokenisasi, filtering), perhitungan frekuensi kemunculan kata (tf), dan klasifikasi sentimen, yang digunakan yaitu term frequency (tf) dan Gaussian naive bayes. Hasil dari penelitian ini adalah mengklasifikasikan komentar pada video instagram dalam sentimen positif, negatif, netral dan mengetahui kualitas dari setiap proses analisa sentimen yang diambil dan membandingkan algoritma Multinomnal Naive Bayes. Hasil dari akurat F-score yang didapat adalah 73% pada percobaan dengan menggunakan algoritma Gaussian Naive Bayes sementara pada percobaan dengan menggunakan algoritma Multinomnal Naive Bayes akurasi yang didapat sebesar 83% pendekatan ini diharapkan akan sangat berguna bagi pengembangan analisa sentimen pada penelitian selanjutnya.

Kata Kunci: Machine Learning; Analisis Sentimen; Text Mining; Term Frequency (TF); Multinomial Naive Bayes

Abstract-The more advanced information technology and people's standard of living have resulted in the increasing demands of society on the quality of services and products used. The need for this smartphone has become a lifestyle that is considered important for some people today. This phenomenon supports the emergence of many smartphones that offer products to meet people's needs for technology in terms of communication. In this study, the smartphone brands used were Vivo_Indonesia, Oppo Indonesia. Machine learning is an artificial brightness (AI) application that provides the ability for systems to learn and improve automatically from experience without having to be explicitly programmed. Machine learning provides new ways to gain insight and help research bodies solve problems. One example is sentiment analysis which is done automatically. Sentiment analysis itself needs to be done because social media users in the community are increasing so that it affects the development of public opinion. Therefore, this can be used to analyze public opinion by applying scientific data, one of which is Text Mining or known as text analytics. The overall stages of the method used in this research is to use Text Mining on videos on Instagram regarding smartphone reviews with the scraping, labeling, preproseccing (case folding, tokenization, filtering) methods, calculating the frequency of word occurrences (tf), and sentiment classification, which are used namely term frequency (tf) and Gaussian naive bayes. The results of this study are to classify comments on Instagram videos in positive, negative, neutral sentiments and find out the quality of each sentiment analysis process taken and compare the Multinominal Naive Bayes algorithm. The results of the accurate F-score obtained are 73% in the experiment using the Gaussian Naive Bayes algorithm while in the experiment using the Multinomial Naive Bayes algorithm the accuracy obtained is 83%. This approach is expected to be very useful for the development of sentiment analysis in further research.

Keywords: Machine Learning; Sentiment Analysis; Text Mining; Term Frequency (TF); Naive Bayes Multinomial

1. PENDAHULUAN

Smartphone merupakan salah satu keperluan masyarakat saat ini yang bisa menunjang aktifitas sehari-hari. Keperluan ini juga begitu diperhatikan oleh perusahaan elektronik sehingga bermunculan banyak berbagai merek-merek smartphone. Semakin majunya teknologi informasi dan taraf hidup masyarakat mengakibatkan semakin meningkatnya tuntutan masyarakat terhadap kualitas pelayanan dan produk yang digunakan. Keperluan smartphone ini telah menjadi gaya hidup yang dianggap penting bagi sebagian masyarakat saat ini.sebuah Fenomena tersebut mendukung munculnya banyak sebuah smartphone yang menawarkan produk untuk memenuhi keperluan masyarakat akan teknologi dalam hal berkomunikasi. Pada penelitihan ini merek smartphone yang digunakan adalah vivo_Indonesia, oppo indonesia[1].

Dengan kemajuan teknologi, membaca komik dapat dilakukan secara online yaitu dengan melalui situs-situs web Saat ini smartphone merk Oppo dan Vivo di pasar indonesia sudah cukup terkenal,sudah banyak pengguna kedua merk smartphone ini,dan tentu sudah tidak asing jika anda ke sebuah dealer atau toko retail pasti si penjual memperkenalkan dan manawarkan anda membeli kedua merk smartphone tersebut di banding dengan merk smartphone lainnya,saya rasa hal ini terjadi hampir di semua penjual retail, keuntungan yang di dapatkan ketika menjual satu unik smartphone Oppo dan Oivo seharga 1.4 juta sekitar 200 sampai 300 ribu rupiah,untuk smartphone 5 jutaan keuntungan penjual bisa mencapai 500-1,5 juta rupiah,keuntungan sebesar itu hanya bisa di dapatkan dengan menjual kedua merk di

atas,untuk merk lain profit atau keuntungan di dapatkan jauh lebih kecil ,oleh karena itu banyak sekali penjual lebih fokus menjual kedua merk tersebut,tidak heran apabila baru masuk di indonesia sudah banyak dealer yang mempromosikan kedua merk di banding merk lainnya.

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes , yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes . Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian[2].

Dalam sebuah data *science* istiah *machine learning* dan data *visualization* bisa menjadi hal yang yang sering dibicarakan. Tidak hanya itu data *science* juga mencangkup bidang yang jauh lebih luas serta banyak sekali istilah-istilah teknologi yang didalam bidang data *science* ini. Salah satu adalah *text mining* atau bisa dikenal dengan istilah *text analytics.* Pada dasar *text mining* merupakan teknilogi (Al) *Artifical intelligence* yang mungkin penggunanya untuk mengubah sebuah konten inti dari dalam sebuah dokumen teks akan menjadi sebuah data kuantitatif secara cepat, datakuantitatif tersebut nantinya tapat digunakan dan diproses lagi sesuai keinginan pengguna.

Pada penelitian terdahulu (Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, Fazal Masud Kundi, 2015) yang berjudul *Sentiment Analysis on* YouTube: *A Brief Survey* menyimpukan bahwa analisis sentimen pada YouTube dapat digunakan untuk memecahkan masalah seperti kamus sentimen yang terbatas, gaya bahasa yang tidak resmi yang digunakan oleh pengguna dan menetapkan label untuk nilai setiap kata. Sedangkan pada penelitian (Evasaria M. Sipayung, Herastia Maharani, Ivan Zefanya, 2016) yang berjudul Perancangan Sistem Analisis Sentimen Komentar Pelanggan menggunakan Metode Naïve Bayes *Classiffier* menyimpulkan bahwa Naïve Bayes *classiffier* dapat menganalisis sentimen dari 175 data dan terbukti memiliki nilai akurasi sentimen sebesar 75,42%, sehingga mampu mengembalikan dokumen dan kecocokan data yang tinggi[3].

2. METODOLOGI PENELITIAN

2.1 Text Mining

Text Mining merupakan proses penambangan data berupa teks dan sumber datanya didapat dari dokumen dan bertujuan untuk mencari kata-kata yang mewakili isi dari dokumen sehingga dapat dianalisa keterhubungan antar dokumen [4]. Tujuan dari *Text Mining* adalah mengekstrak informasi yang berguna dari sumber data. Jadi, sumber data yang digunakan pada *text minig* adalah sekumpulan dkumen yang memiliki format yang tidak terstruktur melalui identifikasi dan eksplorasi pola yang menarik.

2.2 Algoritma Naive Bayes

Algoritma Naïve Bayes *classification* bertujuan untuk melakukan klasifikasi data pada kelas tertentu. Unjuk kerja pengklasifikasi diukur dengan nilai predictive accuracy[5]. Tahapan dari proses algoritma Naïve Bayes adalah sebagai berikut:

1. Menghitung jumlah *class*/label.
2. Menghitung jumlah kasus yang sama dengan *class* yang sama.
3. Mengalikan semua variabel *class*.
4. Membandingkan hasil semua variabel.

Algoritma Naïve Bayes *classification* memiliki persamaan yang dapat dijadikan acuan untuk menghitung nilai probabilitas dalam pengambilan suatu keputusan, adapun persamaannya adalah sebagai berikut.

$$P(\mathrm{X}|\mathrm{H}) = \frac{\mathrm{P}(\mathrm{X}|\mathrm{H}),\mathrm{P}(\mathrm{X})}{\mathrm{P}(\mathrm{H})} \qquad (1)$$

Dimana :

X : Data dengan *class* yang belum dikrtahui
H : Hipotesis data X merupakan suatu *class* spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X
P(H) : Probabilitas hipotesis H
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

2.3 Gaussian Naive Bayes

Gaussian Naïve Bayes dapat digunakan untuk memperkirakan distribusi data karena hanya perlu memperkirakan *mean* dan *standard deviasi* dari data latih.

$$\overline{\mathrm{X}} = {}^{1}/_{\mathrm{n}} \sum \mathrm{X} \qquad (2)$$

Dimana:

$\overline{\mathrm{X}}$ = *Mean* (X).

n = Jumlah contoh.
X = Nilai untuk variabel input dalam data latihan.

Sedangkan persamaan untuk menghitung *standard deviasi* dapat dilihat di bawah ini.

$$S = \sqrt{1/n \sum(X\iota - \overline{X}^2)} \quad (3)$$

Dimana :
S = *Standard deviasi.*
N = Jumlah contoh.
$\overline{X}$ = Rata-rata nilai untuk variabel input dalam data latihan.
Xi = Nilai spesifik dari variabel x

2.4 Multinomial Naïve Bayes

Probabilitas dari dokumen d yang ada di kelas c dapat menghitung dengan persamaan 4 di bawah ini:

$$(c|d) \propto P(c)\prod 1 \leq k \leq nd\ P(tk|c) \quad (4)$$

Dimana:
P(c) = *prior probability* dari sebuah dokumen yang terdapat dalam kelas *c*.
$<t_1,t_2,...,tn_d>$ = kumpulan *token* dalam dokumen *d* yang merupakan bagian dari *vocabulary* yang digunakan untuk mengklasifikasi dan n_d adalah jumlah token tersebut di dalam dokumen *d*. Untuk memperkirakan *prior probability* P(c) digunakan persamaan 2.5 sebagai berikut:

$$P(C) = \frac{Nc}{N} \quad (5)$$

Dimana:
Nc = Jumlah dokumen *training* dalam kelas c.
N = Jumlah dari dokumen *trainin*g seluruh kelas.

Bisa memperkirakan *conditional probability* **P(t|c)** persama yang digunakan, iyalah:

$$P(t_k|c = \frac{Tc_t}{\sum t' \epsilon\, v\, T_{c\,t'}} \quad (6)$$

Dimana:
T_ct = Jumlah kemunculan *term* t dalam sebuah dokumen *training* dari kelas c.
$\sum t' \in V\ T_ct'$ = Jumlah total dari keseluruhan *tem* yang terdapat dalam sebuah dokumen *training* dari kelas c.

Masalah dari proses perkiraan nilai *conditional probabilities* pada persamaan (2.77) adalah terdapat nilai nol dari sebuah kombinasi *(term|class)* yang tidak terdapat dalam data *trainin*g. Berdasarkan contoh diatas, bila *term WTO* dalam data training hanya terdapat dalam dokumen China, maka perkiraan untuk kelas-kelas lainnya, misalnya UK, akan bernilai nol (0):

$$(WTO|UK) = 0 \quad (7)$$

Untuk menghilangkan nilai nol, digunakan *add-one* atau *Laplace smoothing.* Proses ini menambahkan nilai satu (1) pada setiap nilai Tct dari perhitungan *conditional probabilities*. Sehingga persamaan untuk *conditional probabilities* menjadi:

$$P(t_k|c = \frac{Tc_t+1}{(\sum t' \epsilon\, v\, T_{c\,t'})+B'} \quad (8)$$

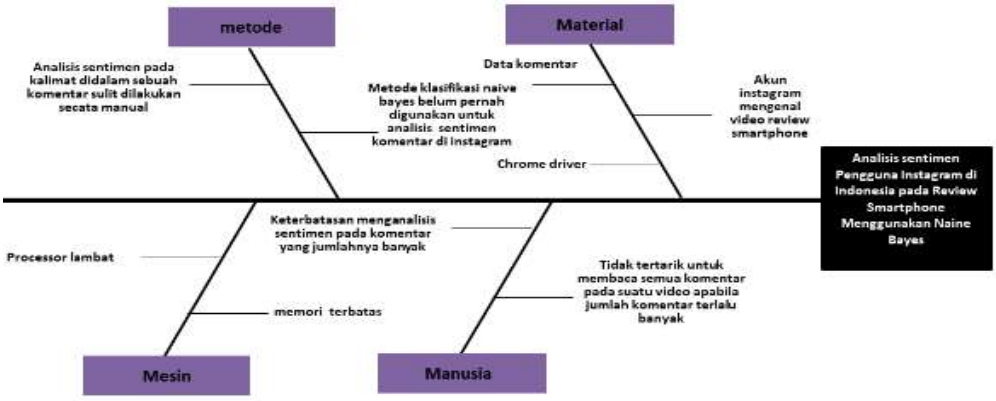
.

3. HASIL DAN PEMBAHASAN

3.1 Analisa Masalah

Dalam penelitian ini, permasalahan yang akan dibahas adalah bagaimana membangun sebuah sistem yang dapat melakukan analisa sentimen terhadap komentar pengguna pada aplikasi instagram mengenai video *review smartphone* menggunakan metode klasifikasi Naïve Bayes secara otomatis, cepat dan tepat.

Setelah melakukan analisis masalah, selanjutnya penulis mengidentifikasi dan menjelaskan penyebab masalah yang terjadi untuk kemudian dicari solusinya. Kemudian proses dari analisis dan identifikasi masalah yang telah dilakukan dapat dijelaskan melalui *Ishikawa Diagram (fishbone)* yang bisa dilihat dari gambar 3.1 dibawah.



Gambar 1. *Ishikawa Diagram*

Pada gambar 1. diatas merupakan *Ishikawa Diagram* yang dapat menjelaskan penyebab atau landasan masalah dalam penelitian ini sehingga dibangunlah sebuah sistem yang akan menjadi solusi dari masalah tersebut. Dari permasalahan yang terdapat pada sistem penelitian ini dibagi menjadi 4 komponen yaitu metode, material, mesin dan manusia. Pada metode dijelaskan mengenai bagaimana sistem dirancang dan dikembangkan berdasarkan metode-metode yang belum pernah digunakan pada penelitian-penelitian yang sudah pernah dilakukan. Sedangkan pada material dijelaskan mengenai bagian-bagian utama dan penting yang diperlukan untuk menjalankan sistem pada penelitian ini. Kemudian pada mesin dijelaskan mengenai hal apa saja yang kemungkinan akan mempengaruhi kinerja dari perangkat keras (*hardware*) dalam menjalankan sistem pada penelitian ini. Dan pada manusia dijelaskan mengenai berbagai hal yang kemungkinan akan diterima oleh *user* dalam menjalankan sistem pada penelitian ini.
Adapun struktur data yang digunakan dalam penelitian ini yaitu:

1. ***Scraping***
   Scraping dilakukan untuk mengumpulkan data yang selanjutnya digunakan sebagai pembelajaran bagi *Machine Learning*. Alamat *url* instagram. Instagram dipilih sebagai tempat untuk melakukan *scraping data* karena banyak *video* pada instagram yang membahas tentang *review* pada sebuah *smartphone* dan tentu saja para pengguna instagram yang aktif berinteraksi sehingga dapat dikumpulkan data komentar yang jumlahnya bisa memadai untuk digunakan sebagai acuan pada *machine learning*. *Scraping* dilakukan secara otomatis menggunakan *software* Chrome Driver. Pada umumnya data komentar yangdihasilkan dari proses *scraping* berupa data asli yang ada pada elemen *XML atau HTML* pada halaman-halaman *website*. Kemudian data dari hasil *scraping* tersebut akan dijadikan sebagai *dataset* yang jumlahnya sekitar 3.000 komentar semi terstruktur. Adapun ekstensi *file dataset* yang digunakan dalam proses ini adalah csv.
2. ***Labelisasi***
   Labelisasi dilakukan untuk memberikan identitas pada setiap data komentar dilakukan secara manual sesuai dengan kebutuhan sistem sebagai data latih dimana ada tiga kategori sentimen yaitu sentimen positif, negatif, dan netral. Hasil dari labelisasi ialah data berbentuk .csv yang sudah terlabel. Contoh labelisasi terhadap komentar dapat dilihat pada Tabel 1. di bawah ini.

Tabel 1. Contoh Labelisis Terhadap Komentar

Id user	Komentar	Sentimen
0	Dari semua teman2 yang udah make hp nya, semua pada puas dengan spesifikasi	Positif
1	Emang sih bagus tapi desain nya masih perlu banyak perubahan	Nitral
2	Baterainya cepat habis,jadi ribet kemana- mana mesti bawak pd	Negatif
3	Harganya mahal banget gak sesuai dengan fiturnya	Negatif
4	Gua udah beli ni smartphone pokoknya harganya menjamin kualitas	Positif

3.2 Analisa Data

1. **Preprocessing**
   Pada penelitian ini *preprocessing* dilakukan untuk mengolah data yang ada sehingga peneliti dapat menghindari gangguan pada data-data yang tidak konsisten. Tujuannya agar hasil *output* dari klasifikasi memiliki tingkat keakuratan yang tinggi. Tahapan dari *preprocessing* meliputi *case folding*, *tokenizing*, dan *filtering.* Adapun penjelasan dari masing-masing tahapan tersebut antara lain:
   a. *Case Folding* merupakan tahap mengubah semua huruf yang terdapat pada komentar menjadi huruf kecil. Hanya huruf "a-z" yang dapat diterima.
      Contoh: Review Pribadi Gw pake ZF Max M2 Ga ada Lag buat game. Hasil: review pribadi gw pake zf max m2 ga ada lag buat game.
   b. *Tokenizing* adalah tahap pemotongan *string input* berdasarkar tiap kata yang menyusunnya. Secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata.

Contoh: review pribadi gw pake zf max m2 ga ada lag buat game.
Hasil:[review][pribadi][gw][pake][zf][max][m2][ga][ada][lag][buat][game]

c. *Filtering* adalah tahap mengambil kata-kata yang dianggap penting. Pada penelitian ini digunakan fungsi regex untuk proses *filtering* yaitu:
   1) Menghapus semua karakter khusus.
   2) Menghapus semua karakter tunggal.
   3) Menganti beberapa spasi menjadi spasi tunggal.
   4) Konversi huruf menjadi huruf kecil.
   5) *Stopword* (menghilangkan dari sambing, kata depan atau kata tidak terkait hubunganya dari anaslisa sentimen).

   Contoh: review pribadi saya memakai zf max m2 tidak ada kelambatan untuk bermain game.
   Hasil: review pribadi memakai zf max m2 kelambatan bermain game.

Tabel 2. Perbandingan Hasil *Preprpcessing*

Sebelum *preprocessing*	Sesudah *preprocessing*
Review Pribadi Gw pake ZF Max M2 Ga ada lag buat game.	review pribadi pakai zf max m2 lambat main game

2. Term Frequency (TF)

Term frequency adalah suatu komentar yang bisa melewati tahap *preprocessing*. *Term* merupakan kata atau frase yang dapat digunakan untuk mengetahui konteks dari dokumen. Karena setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, maka dilakukan setiap kata tersebut diberikan sebuah indikator, yaitu *term weight.* Berikut ini merupakam contoh perhitungan *term frequency* (TF). Data kalimat dapat dilihat pada Tabel 3 di bawah ini.

Tabel 3. Data Kalimat

No.	Komentar
1.	harga mahal
2.	saya menyesal beli handphone ini
3.	handphone ini kamera bagus
4.	oppo di indonesia murah
5	saya baru pakai oppo

Pada Tabel 3. terdapat 5 kalimat, untuk mencari *term frequency* makalangkah paling awal yang hasur dilakukan adalah dengan memisahkan kata dalam kalimat untuk dijadikan kumpulan kata seperti yang dilihat dari tabel 4. di bawah.

Tabel 4. Kumpulan Kata pada Kalimat

Harga	Handphone	Bagus	Saya
Mahal	Ini	Oppo	Baru
Saya	Handphone	Di	Pakai
Menyesal	Ini	Indonesia	Oppo
Beli	Kamera	Murah	

Pada Tabel 4. terdapat daftar kata-kata setelah dilakukannya pemisahan kata pada kalimat selanjutnya disusun, apabila terdapat kata yang muncul lebih dari 1 kali maka cukup ditulis 1 kali saja. Setelah membuat daftar kata maka akan dihitung jumlah kemunculan kata pada setiap dokumen (*term frequency*) pada kalimat yang terdapat pada Tabel 3. Berikut merupakan penyusunan daftar kata dan penghitungan *term frequency* yang dapat dilihat pada Tabel 5. di bawah ini.

Tabel 5. Penyusunan Daftar Kata Perhitungan Nilai *Term Frequency* (Tf)

Kata	**D1**	**D2**	**D3**	**D4**	**D5**	**TF**
Harga	1	0	0	0	0	1
Mahal	1	0	0	0	0	1
Saya	0	1	0	0	1	2
Menysal	0	1	0	0	0	1
Beli	0	1	0	0	0	1
Handphone	0	1	1	0	0	2
Ini	0	1	1	0	0	2
Kamera	0	0	1	0	0	1
Bagus	0	0	1	0	0	1
Oppo	0	0	0	1	1	2
Di	0	0	0	1	0	1
Indonesia	0	0	0	1	0	1
Merah	0	0	0	1	0	1

Baru	0	0	0	0	1	1
Pakai	0	0	0	0	1	1

Pada Tabel 5. dapat dilihat hasil dari proses penyusunan daftar kata dan perhitungan jumlah kata yang muncul dalam suatu kalimat atau dokumen. D1 merupakan dokumen ke-1 seperti yang dapat dilihat pada Tabel 3, sedangkan D2 merupakan dokumen ke-2 dan seterusnya. Setiap dokumen akan di seleksi berdasarkan daftar kata, bila kata tersebut terdapat pada dokumen akan diberi nilai 1 sedangkan bila kata tersebut tidak ada akan diberi nilai 0. Pada kolom TF merupakan jumlah dari kemunculan kata pada total dokumen yang dijadikan sebagai contoh. Setelah mendapatkan nilai TF maka langkah selanjutnya adalah menghitung nilai IDF, menggunakan rumus di bawah ini:

$$idf = log\ (\frac{total\ jumlah\ dokumen}{tf})$$

Berikut ini merupakan contoh perhitungan IDF yang dilihat dari tabel 6. di bawah.

Tabel 1. Perhitungan Nilai IDF

Kata	TF	IDF
Harga	1	0.69897
Mahal	1	0.69897
Saya	2	0.39794
Menyesal	1	0.69897
Beli	1	0.69897
Handphone	2	0.39794
Ini	2	0.39794
Kamera	1	0.69897
Bagus	1	0.69897
oppo	2	0.39794
Di	1	0.69897
Indonesia	1	0.69897
Murah	1	0.69897
Baru	1	0.69897
Pakai	1	0.69897

Dari tabel 6 bisa dilihat apabila ingin menghitung IDF dari kata "handphone" yaitu dengan mencari total jumlah dokumen dimana pada Tabel 2. terdapat 5 dokumen sedangkan kemunculan kata "handphone" pada seluruh dokumen terdapat 2 kali, sehingga perhitungan manualnya yaitu menggunakan rumus di bawah ini:

$$log\left(\frac{total\ jumlah\ dokumen}{tf}\right)$$

Dan dapat dituliskan log $\left(\frac{5}{2}\right) = 0.39794$ dan seterusnya untuk menghitung nilai IDF dari kata yang lain. Setelah mendaparkan nilai IDF langkah selanjutnya adalah mencari nilai TF-IDF seperti yang bisa kita lihat di tabel 7. di bawah.

Tabel 7. perhitungan nilai TF-IDF

Kata	D 1	D 2	D 3	D 4	D 5	TF	IDF	TF-IDF				
								D1	D2	D3	D4	D5
Harga	1	0	0	0	0	1	0.69897	0.69897	0	0	0	0
Mahal	1	0	0	0	0	1	0.69897	0.69897	0	0	0	0
Saya	0	1	0	0	1	2	0.39794	0	0.39794	0	0	0.39794
Menyesal	0	1	0	0	0	1	0.69897	0	0.69897	0	0	0
Beli	0	1	0	0	0	1	0.69897	0	0.69897	0	0	0
Handphone	0	1	1	0	0	2	0.39794	0	0.69897	0.69897	0	0
Ini	0	1	1	0	0	2	0.39794	0	0.39794	0.39794	0	0
Kamera	0	0	1	0	0	1	0.69897	0	0	0.69897	0	0
Bagus	0	0	1	0	0	1	0.69897	0	0	0.69897	0	0
Oppo	0	0	0	1	1	2	0.39794	0	0	0	0.39794	0.39794
Di	0	0	0	1	0	1	0.69897	0	0	0	0.69897	0
Indonesia	0	0	0	1	0	1	0.69897	0	0	0	0.69897	0
Merah	0	0	0	1	0	1	0.69897	0	0	0	0.69897	0
Baru	0	0	0	0	1	1	0.69897	0	0	0	0	0.69897
Pakai	0	0	0	0	1	1	0.69897	0	0	0	0	0.69897

Nilai TF-IDF yang terdapat pada Tabel 7. inilah yang akan dijadikan *input* dalam proses klasifikasi menggunakan algoritma Naïve Bayes.

3.3 Klasifikasi Menggunakan Algoritma Naive Bayes

Sebuah data sudah memastikan proses *preprocessing* dan *term frequency* merupakan data latih yang akan menjadi *input* pada proses pengujian. Berlajut dari data uji bisa melalui tahap pengklasifikasian menggunakan algoritma. Berikut merupakan contoh kalimat yang akan diklasifikasikan menggunakan algoritma Naive Bayes yang dapat dilihat pada Tabel 8. di bawah ini.

Tabel 3.8 Contoh Kalimat Klasifikasi

No.	Komentar	*Class* Kata		Label
		Positif	Negatif	
1.	harga mahal	TidakAda	Ada	Negatif
2.	saya menyesal beli handphone ini	Tidak Ada	Ada	Negatif
3.	handphone ini kamera bagus	Ada	Tidak Ada	Positif
4.	samsung di indonesia murah	Ada	Tidak Ada	Positif
5.	saya baru pakai Samsung	?	?	?

Pada Tabel 8. dapat dilihat contoh kalimat 1 sampai dengan 4 sudah diberi label (sentimen). Sedangkan pada contoh kalimat 5 belum diberi label (sentimen). Maka dari itu untuk memberi label (sentimen) pada contoh kalimat 5 akan digunakan metode klasifikasi algoritma Naïve Bayes. Adapun tahapan-tahapan klasifikasi dalam algoritma Naïve Bayes adalah sebagai berikut.

1. Menghitung jumlah *class*/label:
   P(C = Positif) = 2/4 (jumlah kalimat dengan kategori "positif")
   P(C = Negatif) = 2/4 (jumlah kalimat dengan kategori"negatif")
   P(C = Netral) = 0/4 (jumlah kalimat dengan kategori"netral")
2. Menghitung jumlah kasus yang sama dengan *class* yang sama:
   P(*Class* Kata = Positif | C = Positif) =2/2 (jumlah *Class* Kata berstatus positif dengan keterangan positif, dibagi dengan jumlah data yang positif).
   P(*Class* Kata = Positif | C = Negatif) =0/2
   P(*Class* Kata = Positif | C = Netral) =0/0
   P(*Class* Kata = Negatif | C = Positif) =2/0
   P(*Class* Kata = Negatif | C = Negatif) =2/2
   P(*Class* Kata = Negatif | C = Netral) =0/0
3. Mengalikan semua hasil variabel:
4. Variabel "Positif" (P(*Class* Kata = Positif | C = Positif)) * (P(*Class*Kata = Negatif | C = Positif)) * (P(C = Positif)) = 2/2 * 2/0 * 2/4 = ~
5. Variabel "Negatif" (P(*Class* Kata = Positif | C = Negatif)) * (P(*Class* Kata = Negatif | C = Negatif)) * (P(C = Negatif)) = 0/2 * 2/2 * 2/4 = 0/16
6. Variabel "Netral" (P(*Class* Kata = Positif | C = Netral)) * (P(*Class*Kata = Negatif | C = Netral)) * (P(C = Netral)) = 0/0 * 0/0 * 0/4 = 0/0
7. Membandingkan hasil semua variabel:

Dari hal di atas, penilai probabilitas tertinggi ada di variabel "Netral", sehingga bisa disimpulkan bahwa kalimat tersebut termasuk kedalam kategori komentarnetral. Maka pada contoh kalimat 5 diberi label (sentimen) "Netral".

3.4 Pengujian Sistem

Dari Pengujian sistem dapat dilakukan untuk memeriksa suatu sistem sudah dibangun, akan dilakukan suatu penguji sistem terdiri dari proses pelatihan atau pengujian terhadap data komentar. Bisa tepat Pengujian Sistem dari F-*score*, bisa melakukan dengan 2 kali peroses, total peroses yang dilakukan dimasukan untuk membandingkan akurasi yang paling baik akan menghasilkan setiap algoritma. ini hasil dari setiap penguji ditampilakan oleh Tabel 9 dan Tabel 10 berikut ini.

Tabel 9. Hasil Pengujian Sistem dengan Gaussian NB

	Gaussian NB			
	Precision	*recall*	f1-score	Support
Positif	0.92	0.67	0.78	401
Netral	0.93	0.63	0.75	789
Negatif	0.52	0.96	0.67	425
Weigted/avg	0.82	0.73	0.74	1615

Tabel 10. Hasil Pengujian Sistem dengan *Multinomial* NB

Multinomial NB

	Precision	*recall*	1-score	Support
Positif	0.92	0.67	0.78	401
Netral	0.93	0.63	0.75	789
Negatif	0.52	0.96	0.67	425
Weigted/avg	0.82	0.73	0.74	1615

Dari Tabel 9 dan Tael 10 dapat dilihat bahwa dengan menggunakan algoritma Gausisian Naïve Bayes sistem dapat menghasilkan akurasi sebesar 73% dari 8074 *dataset* yang telah di *split* menjadi data latih dan data uji sementara pada algoritma *Multinomial* Naïve Bayes sistem dapat menghasilkan akurasi ketepatan hingga 81%.

4. KESIMPULAN

Berdasarkan hasil kesimpulan dari hasil dan pembahasan pada penelitian ini dimana dengan melakukan scraping data, analisa sentimen dapat dilakukan secara otomatis dan cepat. Preprocessing pada komentar terbukti efektif menghasilkan kalimat yang penting terhadap proses analisa sentimen. Dimana hasil pengujian analisa sentimen pada komentar menggunakan Gaussian Naive Bayes bisa menghasilkan akurasi sebesar 73% sementara dengan menggunakan Multinomial Naïve Bayes menghasilkan akurasi sebesar 81%. Sehingga dari pengujian tersebut yang dilakukan Multinomial Naïve Bayes menghasilkan akurasi akan lebih baik dibandingkan dengan menggunakan Gaussian Naïve Bayes..

REFERENCES

[1] "Algoritma Naive Bayes," 2019. https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/.
[2] "No Title," file:///C:/Users/khairil/Downloads/46479-121-107793-1-10-20190506.pdf. .
[3] Balya, "Analisis Sentimen Pengguna Youtube di Indonesia pada Review Smartphone Menggunakan Naïve Bayes," 2019.
[4] A. Raharjo, "Cybercrime; Pemahaman dan Upaya Pencegahan Kejahatan Berteknologi," Citra Aditya Bakti, pp. 11–36, 2002.
[5] "machine-learning," https://www.advernesia.com/blog/data-science/machine-learning-adalah/. .
[7] "Teori Text Mining,"No Title," https://informatikalogi.com/text-preprocessing/. .
[8] Bustami, "Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah," TECHSI J. Penelit. Tek. Inform., vol. 4, pp. 127–146, 2010.