Analisis Sentimen Terhadap *Review E-Commerce* Pada *Google Play Store* Menggunakan Metode *Naïve Bayes Classifier* (NBC) dengan Seleksi Fitur *Information Gain* (IG)

Oleh: Rara Ayu Puspita
Univeristas Muhammadiyah Semarang

Article history
Submission :
Revised :
Accepted :

Keyword:
Analisis Sentimen, *E-Commerce*, *Information Gain*,*Naïve Bayes Classifier*.

Abstract

E-Commerce is a new online-based buying and selling process that already has a mobile application. There are 2 best E-Commerce in Indonesia, namely Tokopedia and Shopee. The importance of conducting a sentiment analysis on the E-Commerce review on the Google Play Store is to help consumers find out the level of security when shopping using mobile applications from the two E-Commerce. In addition, the output of this analysis can be used as an evaluation consideration for the second E-Commerce service provider. This study uses the Naïve Bayes Classifier because it is a method that works very well compared to other classification models. Then the Information Gain (IG) Feature Selection is added so that the classification is more optimal and can improve its accuracy. The accuracy value of the Naïve Bayes Classifier classification method is obtained by 50% for Tokopedia and 40% for Shopee. After adding the Information Gain feature selection using the 100 best features for Tokopedia and 50 best features for Shopee, an accuracy value of 77.5% for Tokopedia and 65% for Shopee was added.

PENDAHULUAN

Pesatnya perkembangan teknologi informasi dan komunikasi berdampak pada perubahan di berbagai bidang, seperti sosial, ekonomi, politik, dan budaya, serta berdampak pada perubahan gaya hidup, termasuk pola konsumsi serta cara berjualan dan berbelanja masyarakat. Di era ini, masyarakat memanfaatkan teknologi informasi dan komunikasi untuk membeli dan/atau menjual barang dan/atau jasa melalui internet. Fenomena ini dikenal dengan perdagangan elektronik atau *E-Commerce*. Fenomena *E-Commerce* menyediakan pilihan cara berbelanja bagi masyarakat dengan tidak perlu datang langsung ke toko.

E-Commerce merupakan suatu konsep baru yang biasa digambarkan sebagai proses jual beli barang atau jasa pada *Word Wide Web* Internet atau prose jual beli atau pertukaran produk, jasa, dan informasi melalui jaringan informasi termasuk internet. *E-Commerce* merupakan kegiatan bisnis yang dijalankan secara elektronik melalui suatu jaringan internet atau kegiatan jual beli barang atau jasa melalui jalur komunikasi digital. (Nugroho, 2006). Indonesia merupakan pasar *E-Commerce* terbesar di Asia Tenggara. Menurut data Wearesocial dan Hootsuite, sekitar 90% pengguna internet di Indonesia pernah berbelanja *online*. Pada tahun 2019, nilai kapitalisasi pasar *E-Commerce* di indonesia mencapai *USD* 21 miliar atau sekitar Rp 294 triliun. Berdasarkan laporan McKinsey, industri *E-Commerce* di indonesia diprediksi akan mencapai nilai *USD* 40 miliar pada tahun 2022.

Banyaknya *E-Commerce* yang ada di Indonesia maka pasti semakin banyak kejahatan *online* yang terjadi, sehingga penilaian masyarakat terhadap suatu *E-Commerce* dapat dijadikan analisa terhadap pasar *online E-Commerce* dapat membantu masyarakat lain supaya lebih berhati – hati dalam melakukan transaksi *online*. Penulis merasa perlu untuk melakukan penelitian ini yaitu dengan membuat sebuah sistem penganalisa opini

atau biasa disebut analisis sentimen masyarakat sehingga bisa mengetahui dan membantu memberikan informasi mengenai analisa sentimen *E-Commerce* masyarakat. Adapun *E-Commerce* yang menjadi objek penelitian ini adalah Tokopedia dan Shopee. Dilansir dari artikel yang ditulis oleh media *online* iNews.id , kedua *e- commerce* tersebut termasuk tiga situs *E-Commerce* terbaik di Indonesia yang paling sering dikunjungi konsumen dengan data kunjungan peringkat pertama yaitu Tokopedia sebesar 1,2 miliar kali, Shopee sebesar 837,1 juta kali.

Review masyarakat saat ini dapat diperoleh di mana saja salah satunya di aplikasi yaitu *Google Play Store* . Berdasarkan survei yang dilakukan pada sebuah situs 1 Februari 2018 menjelaskan bahwa 69% masyarakat Indonesia mengakses internet dengan perangkat mobile. Bersama negara-negara berkembang lain, seperti Brasil, India, Turki, dan Meksiko, pengguna Android di Indonesia menjadi penyumbang pemasukan terbesar di Android Play Store.Tentu saja, informasi yang terkandung dalam *review* ini sangat berharga sebagai alat penentu kebijakan dan ini bisa dilakukan dengan *Text Mining*. *Teks Mining* adalah salah satu Teknik yang dapat digunakan untuk melakukan klasifikasi dokumen dimana *Text Mining* merupakan variasi dari *Data Mining* yang berusaha menemukan pola menarik dari sekumpulan data tekstual yang berjumlah besar (Kurniawan, Effendi, & Sitompul, 2012). Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks *(text categorization)* dan pengelompokan teks (*text clustering)*. Analisis Sentimen atau *Opinion Mining* adalah studi komputasional dari opini – opini orang, sentimen dan emosi melalui entitas atau atribut yang dimiliki yang diekspresikan dalam bentuk teks (Aditya, Hani'ah, Fitrawan, Arifin, & Purwitasari, 2016). Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif, atau netral (Pang & Lee, 2008).

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema *Bayes* . Metode pengklasifikasian dg 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. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dr ” master ” tabel keputusan.

Naive Bayes Classifier bekerja sangat baik dibanding dengan model *classifier* lainnya. Hal ini dibuktikan oleh Xhemali , Hinde Stone dalam jurnalnya “*Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages*” mengatakan bahwa “*Naïve Bayes Classifier* memiliki tingkat akurasi yg lebih baik disbanding model *classifier* lainnya”. Keuntungan penggunan adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yg kecil unt menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai *variable independent*, maka hanya *varians* dr suatu *variable* dalam sebuah kelas yg dibutuhkan unt menentukan klasifikasi, bukan keseluruhan dr matriks kovarians.

Seleksi Fitur *Information Gain* adalah *Information Gain* dalam *machine learning* digunakan untuk mengukur

seberapa relevan / berpengaruh sebuah *feature* terhadap hasil pengukuran. Penggunaan teknik ini dapat mereduksi dimensi *feature* dengan cara mengukur reduksi *Entropy* sebelum dan sesudah pemisahan. *Information Gain (IG)* dikenal juga dengan sebutan *Mutual Information (MI)* dalam kasus untuk mengetahui *dependency* antara dua variable (x,y). Keuntungan menggunakan metode ini yaitu penyederhanaan model, membuatnya lebih mudah untuk diintrepretasi oleh *researcher/user*, mempercepat proses *training*, Menghidari *curse of dimensionality* (beragam fenomena yang timbul saat menganalisa atau mengolah data dengan dimensi tinggi. Dalam *Machine Learning* ini terjadi ketika *high-dimensional feature space* memiliki n *samples* yang terbatas), serta *Enhance generalization* dengan cara menurunkan *overfitting*.

LANDASAN TEORI

Naïve Bayes Classifier (NBC)

Teorema Bayes merupakan teorema yang mengacu pada probabilitas bersyarat (Siang, 2005). Secara umum teorema Bayes dapat dinotasikan pada persamaan berikut.

$$\boldsymbol{P(Aj \mid Bi) = \frac{P(Bj|Ai)P(Aj)}{P(Bi)}} \quad (2.4)$$

Dimana :

P(A|B) : Peluang kategori j, ketika terdapat kemunculan kata i

P(A|B) : Peluang kata i masuk ke dalam kategori j

P (A) : Peluang kemunculan kategori j

P(B) : Peluang kemunculan kata

Terdapat dua tahap dalam klasifikasi tweet. Tahap pertama adalah pelatihan terhadap tweet yang telah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi tweet yang belum diketahui kategorinya (Falahah dan Nur, 2015). Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_3,...,, a_n" dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori tweet. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan ($\boldsymbol{V_{MAP}}$). Adapun persamaan $\boldsymbol{V_{MAP}}$ adalah sebagai berikut

$$\boldsymbol{VMAP = \underset{Aj}{argmax} P(Vj) \prod_{i-1}^{n} P(ai|vj)} \quad (2.5)$$

Nilai ($\boldsymbol{vj}$) dihitung pada saat training, didapat dengan rumus sebagai berikut:

$$\boldsymbol{P(vj) = \frac{|doc\ j|}{|training|}} \quad (2.6)$$

$|\boldsymbol{doc\ j}|$ merupakan jumlah *review* pada kategori j dalam training. Sedangkan |training|merupakan jumlah *review* dalam data yang digunakan untuk training. Setiap probabilitas kata $\boldsymbol{ai}$ pada setiap kategori $\boldsymbol{P(ai|vj)}$, dihitung pada saat training.

$$\boldsymbol{P(ai|vj) = \frac{ni+1}{|n+kosakata|}} \quad (2.7)$$

Di mana,

ni : jumlah kemunculan kata $\boldsymbol{ai}$ dalam *review* yang berkategori $\boldsymbol{vj}$

n : banyaknya seluruh kata *review* dengan kategori $\boldsymbol{vj}$

|kosakata| : banyaknya kata dalam data training

Seleksi Fitur *Information Gain* (IG)

Information Gain merupakan teknik seleksi fitur yang memakai metode scoring untuk nominal ataupun pembobotan atribut kontinu yang didiskritkan menggunakan maksimal entropy.Suatu entropy digunakan untuk mendefinisikan nilai *Information Gain*. Entropy menggambarkan banyaknya informasi yang dibutuhkan untuk mengkodekan suatu kelas. *Information Gain* (IG) dari suatu term diukur dengan menghitung jumlah bit informasi yang diambil dari prediksi kategori dengan ada atau tidaknya term dalam suatu dokumen. (Maulida, Suyatno, & Hatta, 2016)

Teknik seleksi fitur dengan *information gain* artinya adalah memilih simpul fitur dari pohon keputusan berdasar nilai *information gain*. Nilai *information gain* sebuah fitur diukur dari pengaruh fitur tersebut terhadap keseragaman kelas pada data yang dipecah menjadi subdata dengan nilai fitur tertentu. Keseragaman kelas (entropy) dihitung pada data sebelum dipecah dengan persamaan 2.1 dan pada data setelah dipecah dengan persamaan 2.2 berikut ini.

$$Entropy\ (S) = \sum_{i=1}^{k} (Pi)\ log2(Pi) \quad (2.1)$$

Dengan nilai Pi adalah proporsi data S dengan kelas i. K adalah jumlah kelas pada output S.

$$Entropy\ (S,\ A) = \sum_{i=1}^{v} \left(\frac{Sv}{v} * \text{Entropy}\ (Sv)\right) \quad (2.2)$$

Dengan nilai v adalah semua nilai yang mungkin dari atribut A, Sv adalah subset sari S dimana atribut A bernilai v. Nilai *information gain* dihitung dengan persamaan 2.3 berikut ini:

$$\text{Gain}\ (S, A) = \text{Entropy}\ (S) - \text{Entropy}\ (S,A) \quad (2.3)$$

Dengan nilai Gain (S,A) adalah nilai *information gain*. Entropy (S) adalah nilai entropy sebelum pemisah. Entropy(S,A) adalah nilai entropy setelah pemisah. Besarnya nilai *information gain* menunjukkan seberapa besar pengaruh suatu atribut terhadap pengklasifikasian data. (Rasywir & Purwarianti, 2015)

Ukuran Evaluasi Model Klasifikasi

Evaluasi pada suatu klasifikasi pada umumnya dilakukan dengan menggunakan sebuah himpunan data yang diuji, tidak digunakan dalam pelatihan klasifikasi tersebut. Pada tahap ini terdapat sejumlah ukuran yang dapat digunakan untuk menilai kembali atau mengevaluasi model klasifikasi, yaitu accuracy atau tingkat pengenalan, tingkat kesalahan atau kekeliruan klasifikasi, recall atau sensitivity atau true positif, specificity atau true negatif dan precision.

Model klasifikasi yang telah dibuat yaitu pemetaan dari suatu baris data dengan keluaran sebuah hasil prediksi kelas atau target dari data tersebut. Pada klasifikasi ini terdapat dua kelas sebagai luarannya yang disebut klasifikasi biner. Kedua kelas tersebut biasa diinterpretasikan dalam {0,1}, {+1,-1} atau {positif,negatif}.

Pada proses evaluasi klasifikasi terdapat empat kemungkinan yang terjadi yaitu proses pengklasifikasian pada suatu baris data. Jadi, jika data positif dan diprediksi positif maka akan dihitung sebagai true positif, bahkan jika data itu diprediksi negatif maka akan dihitung sebagai false negatif. Jika data negatif dan diprediksi negatif maka akan dihitung sebagai true negatif, tetapi jika data tersebut diprediksi positif maka akan dihitung sebagai false positif. Hasil klasifikasi biner pada suatu dataset yang dipresentasikan dalam bentuk matriks 2×2 yaitu dinamakan confusion matrix. Berikut merupakan contoh dari matriks

Tabel 1. Confusion Matrix

Confusion Matrix		True Class	
		Positive	Negatif
Predicted Class	Positive	*True Postive* (TP)	*False Positive* (FP)
	Negatif	*False Negatif* (FN)	True Negatif (TN)

Confusion Matrix bermanfaat untuk menganalisis kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang ada. TP dan TN menyatakan pada *classifier* mengenali tuple dengan benar , artinya tuple positif dikenali sebagai positif dan tuple negatif dikenali sebagai negatif. Sedangkan, FP dan FN menyatakan bahwa *classifier* salah dalam mengenali tuple, tuple negatif dikenali sebagai positif dan tuple negatif dikenali sebagai positif. Ada pula dalam formula perhitungan performa klasifikasi yaitu nilai akurasi biasa ditampilkan dalam presentase.

Akurasi

Akurasi adalah nilai ketepatan dimana pengguna memprediksi suatu kata sesuai dengan jawaban suatu sistem. Berikut perhitungan nilai akurasi

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

METODE PENELITIAN

Sumber Data

Data yang digunakan adalah data sekunder. Data diambil dengan proses *Scraping* pada kolom komentar Tokopedia dan Shopee di situs *Google Play Store* . Populasi pada penelitian ini adalah seluruh *review* mengenai Tokopedia dan Shopee di *Google Play Store* . Sampel yang digunakan adalah data pada periode 18 Maret – 7 April 2021 didapatkan data *review* Tokopedia sebanyak 80 *record* dan Shopee sebanyak 80 *record*. Data tersebut adalah data yang dibatasi oleh *Google Play Store* .

Variabel dan Struktur Data

Data yang didapatkan dari hasil *scraping* data menggunakan software Web Harvy dibagi menjadi *data training* dan *data testing* menggunakan acuan penelitian terdahulu dengan perbandingan 80% : 20%. Variabel yang digunakan sebanyak tiga atribut dan satu label. Atribut yang digunakan antara lain tanggal *Review*, rating dan komentar. Dari struktur data di bawah, pada penelitian ini hanya menggunakan atribut "Komentar" pada baris ketiga. Atribut "Rating" hanya akan digunakan untuk kebutuhan analisis deskriptif. Namun, dalam analisis klasifikasi sentiment label yang digunakan untuk pengklasifikasian adalah data hasil pelabelan dari analisis sentiment.Jenis data label adalah kategorik yaitu positif dan negatif.

Tabel 2. Atribut

No	Indikator	Jenis Data	Keterangan
1	Tanggal *Review*	Date	Tanggal dibuatnya *review*
2	Rating	Scale	Tingkat kepuasan pengguna *E-Commerce*
3	Komentar	String	Isi *review* pengguna *E-Commerce*

Atribut tanggal *review* berisi data mulai tanggal 18 Maret – 7 April 2021. Rating pada atribut memiliki nilai antara 1-5 dengan kategori dari yang paling rendah ialah "Tidak Suka Sekali" yang diberi skor "1", "Tidak Suka" dengan skor "2", "Lumayan" dengan skor "3", "Suka" dengan skor "4" dan "Suka Sekali" dengan skor "5". Komentar berisi data komentar dari para konsumen Tokopedia dan Shopee yang berbahasa Indonesia

Langkah Penelitian

Langkah-langkah dalam penelitian ini yaitu :

1. *Scraping* data, dibagi menjadi 2 yaitu data *training* dan data *testing*. Dimana data *training* dan data *testing* diambil otomatis dari aplikasi R Studio.
2. *Input* data kedalam aplikasi R Studio, data yang dimasukkan berupa teks berbahasa indonesia
3. *Dataset*, data berupa teks *review* dari *Google Play Store* berbahasa indonesia
4. *Text Preprocessing*
   Setelah data berhasil di *scraping* kemudian dilakukan text preprocessing yang terdiri dari beberapa tahap di antaranya:
   a. Spelling Normalization
      Merupakan tahap awal yang harus dilakukan untuk mendapatkan dokumen data yang baik. Perlakuan yang dilakukan yaitu memperbaiki kata-kata yang terdapat salah ejaan atau disingkat menjadi bentuk tertentu. Proses ini dilakukan dengan bantuan Microsoft Excel dan software Rstudio.

b. Case folding
Merupakan proses penyeragaman bentuk huruf menjadi huruf kecil semua antara "a" sampai dengan "z". Dengan tujuan agar kata yang ditulis dengan huruf awal kapital dan huruf kecil tidak terdeteksi mempunyai arti yang berbeda.

b. Tokenizing
Merupakan proses pemisahan teks menjadi potongan kata yang disebut dengan token. Bertujuan untuk mendapatkan potongan kata yang akan menjadi entitas serta memiliki nilai dalam matriks dokumen teks yang akan dianalisis.

c. Filtering
Kata dan tanda baca yang nantinya tidak bernilai atau tidak berarti akan dieliminasi seperti url, angka, tanda baca, hastag, kata hubung, kata ganti dan lainya. Pemilihan kata yang bermakna menggunakan Stopwords (menghilangkan kata yang kurang penting). Kata penghubung yang akan dihilangkan yaitu:
   a. Penghubung antar kata,seperti dan, atau, serta.
   b. Preposisi, seperti di, ke, pada.

5. Klasifikasi data dengan Metode *Naive Bayes Classifier*
   a. Membagi data training dan data testing
   b. Hitung P(Vj) dan P(ai|vj) dari data training dengan persamaan berikut.

$$P(vj) = \frac{|doc\,j|}{|training|}$$

$$P(ai|vj) = \frac{ni + 1}{|n + kosakata|}$$

   c. Hitung Vmap positif dan Vmap negative data testing

$$VMAP = \underset{Aj}{argmax} P(Vj) \prod_{i-1}^{n} P(ai|vj)$$

   d. Membandingkan nilai Vmap positif apakah lebih besar dari Vmap negative
   e. Data testing terklasifikasi dalam sentiment positif atau sentiment negative
   f. Memperoleh nilai akurasi

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

6. Seleksi Fitur *Information Gain* untuk mendapatkan fitur-fitur terbaik.
   a. Hitung jumlah dokumen positif dan dokumen negative
   b. Hitung entropy dari seluruh dataset

$$Entropy\,(S) = \sum_{i=1}^{k} (Pi)\ log2(Pi)$$

   c. Hitung banyaknya Fitur i
   d. Hitung entropy dari Fitur i

$$Entropy\ (S, A) = \sum_{i=1}^{r} \left(\frac{Sv}{v} * Enteropy\ (Sv)\right)$$

   e. Hitung gain dari Fitur i
   f. Jika i < n maka menghitung nilai gain kembali, namun jika i > n maka dapat dilanjukan
   g. Hitung bobot information gain

$$Gain\ (S, A) = Entropy\ (S) - Entropy\ (S, A)$$

7. Interpretasi dan pengambilan kesimpulan.

HASIL PENELITIAN dan PEMBAHASAN

Scraping Data

Hasil *Scraping* data *review* Shopee dan Tokopedia pada *Google Play Store* pada periode 18 Maret – 7 April 2021 didapatkan data *review* Shopee sebanyak 80 record dan Tokopedia sebanyak 80 record. Data tersebut merupakan data yang dibatasi dari *Google Play Store* .Teknik pengambilan data yang digunakan sebagai berikut :

1. *Scraping* data dengan menggunakan software Web Harvy dengan memasukan URL dari Tokopedia dan Shopee di *Google Play Store* sebagai berikut :
   - https://play.google.com/store/apps/details?id=com.carajualandiTokopedialengkap&hl=in&gl=US&showAllReviews=true
   - https://play.google.com/store/apps/details?id=com.Shopee.id&hl=in&gl=US&showAllReviews=true
2. Mengekspor file hasil *scraping* dalam bentuk dokumen excel (CSV)

Preprocessing Data

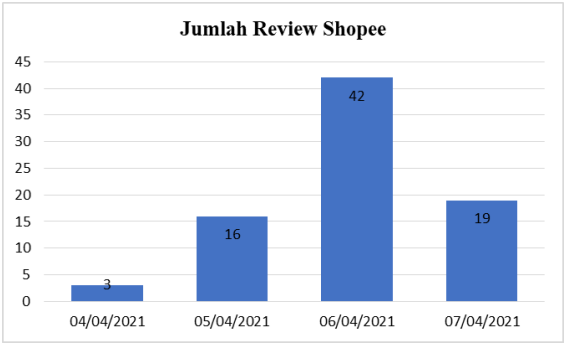
Data dalam penelitian ini merupakan data *review* tanggapan yang diambil dari *Google Play Store* yang memiliki berbagai macam gaya penulisan sehingga data yang diperoleh merupakan data yang tidak terstruktur, oleh karena itu sebelum dilakukan klasifikasi, data perlu diubah menjadi data yang lebih terstruktur. Tahapan untuk mengubah data yang tidak terstruktur menjadi data yang lebih terstruktur disebut tahap preprocessing.

Tabel 3. Preprocessing Data

Hasil Case Folding dan Tokenisasi	
Sangat membantu... Namun dgn perubahan jasa kirim yg otomatis, sangat mengganggu.. rada bingung.., mending pilihan jasa kirim nya manual saja pilihannya seperti dulu.. soalnya kadang ga bisa milih jasa kurir walau status barang sedang dikemas . Pilihan jasa kirim nya mendingan pake sistem yg lama . (Milih sendiri). Itu saja sih.. maaf bintang 3.. tks	sangat membantu. namun dengan perubahan jasa kirim yang otomatis sangat mengganggu, menjadi bingung, lebih baik pilihan jasa kirimnya manual saja seperti dahulu karena terkadang tidak bisa memilih jasa kurir walaupun status barang sedang dikemas. Pilihan jasa kirim lebih baik menggunakan sistem yang lama atau pilih sendiri. Itu saja, maaf saya beri bintang 3. Terima kasih.
Hasil Stemming dan Filtering	
Sangat membantu. Namun dengan perubahan jasa kirim yang otomatis sangat mengganggu, menjadi bingung, lebih baik pilihan jasa kirimnya manual saja seperti dahulu karena terkadang tidak bisa memilih jasa kurir walaupun status barang sedang dikemas. Pilihan jasa kirim lebih baik menggunakan sistem yang lama atau pilih sendiri. Itu saja, maaf saya beri bintang 3. Terima kasih.	sangat membantu. namun dengan perubahan sistem jasa kirim justru mengganggu dan membuat bingung. lebih baik pilihan jasa kirim manual saja seperti dahulu karena kadang tidak bisa pilih jasa kurir walaupun status barang sedang dikemas. pilihan jasa kirim lebih baik menggunakan sistem yang lama yaitu memilih sendiri. maaf saya beri bintang 3, terimakasih.
Hasil Data Bersih	
sangat membantu. namun dengan perubahan sistem jasa kirim justru mengganggu dan membuat bingung. lebih baik pilihan jasa kirim manual saja seperti dahulu karena kadang tidak bisa pilih jasa kurir walaupun status barang sedang dikemas. pilihan jasa kirim lebih baik menggunakan sistem yang lama yaitu memilih sendiri. maaf saya beri bintang 3, terimakasih.	sangat membantu. namun dengan perubahan sistem jasa kirim justru membuat bingung. lebih baik pilihan jasa kirimnya menggunakan sistem yang lama yaitu memilih sendiri. maaf saya beri bintang 3, terimakasih.

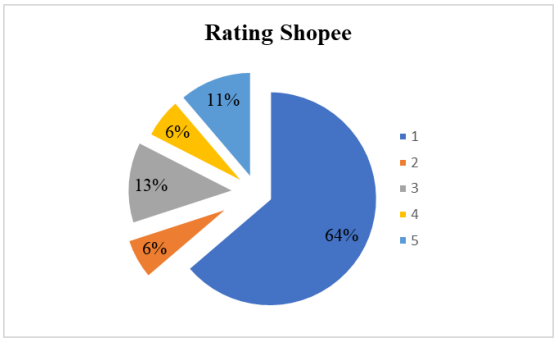
Analisis Deskriptif

Setelah mendapatkan data maka langkah selanjutnya akan dilakukan analisis deskriptif untuk atribut date dan rating. Analisis deskriptif dilakukan untuk mengetahui gambaran umum mengenai *E-Commerce* di laman *Google Play Store* . Informasi yang dapat diambil adalah bagaimana Jumlah *review* yang diberikan oleh konsumen perharinya serta mengetahui tingkat kepuasan konsumen berdasarkan pemberian bintang. Analisis deskriptif dapat dilihat pada grafik dibawah ini :



Gambar 1. Grafik Review Shopee April 2021

Berdasarkan hasil grafik dapat dilihat bahwa komentar terbanyak terjadi pada tanggal 6 April 2021 serta komentar paling sedikit terjadi pada 4 April 2021.

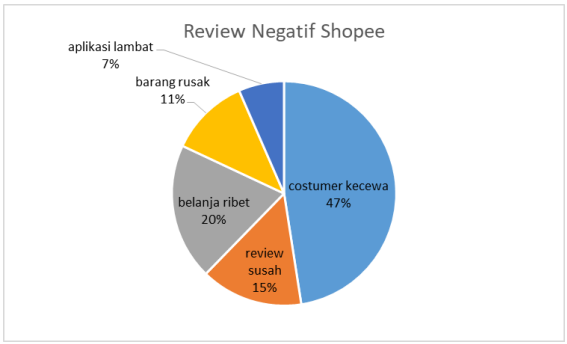


Gambar 2. Grafik Rating Shopee April 2021

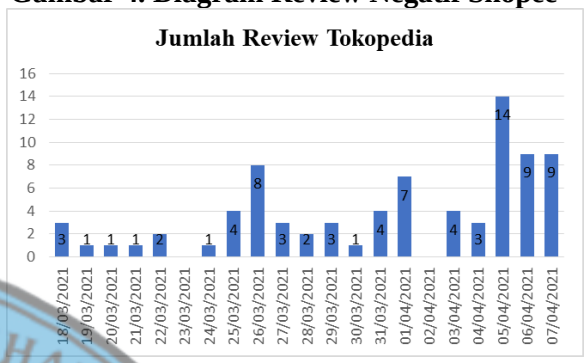
Berdasarkan gambar 4.4 dapat dijelaskan bahwa sebesar 64% konsumen yang memberikan komentar pada Shopee memberikan rating 1 dan selanjutnya 13% konsumen memberikan rating 3 kemudian 11% konsumen memberikan rating 5.

Gambar 3. Visualisasi Word Cloud Shopee

Berdasarkan gambar word cloud diatas menunjukan bahwa semakin besar ukuran tulisan berarti semakin banyak kata tersebut muncul. Dari banyaknya kata yang muncul penulis membuat sebuah grafik yang berisi *review* negatif dari Shopee untuk mengetahui presentase kekurangan dari Shopee yang dianalisis dalam bentuk diagram lingkaran sebagai berikut.

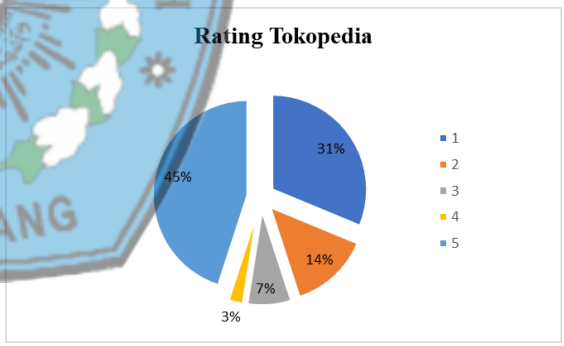


Gambar 4. Diagram Review Negatif Shopee



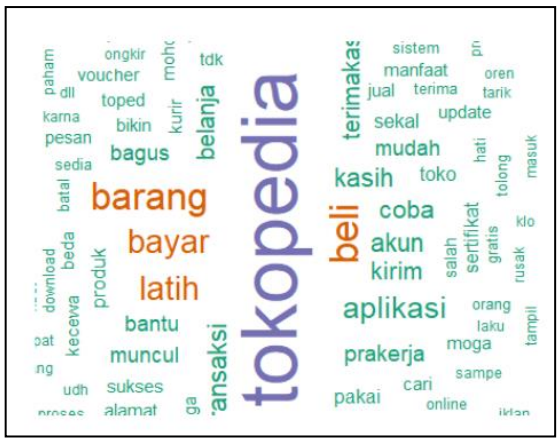
Gambar 5. Grafik Review Tokopedia 18 Maret – 7 April 2021

Berdasarkan hasil grafik dapat diketahui bahwa komentar terbanyak yaitu pada 5, 6, 7 April dan 26 Maret 2021. Kemudian, komentar paling sedikit yaitu 19,20,21, dan 30 Maret 2021.



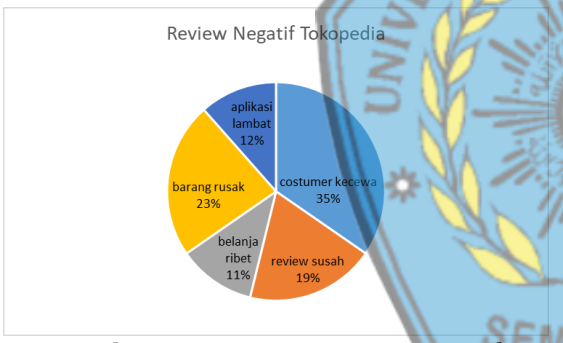
Gambar 6. Grafik Rating Tokopedia 18 Maret – 7 April 2021

Berdasarkan gambar 4.6 dapat digambarkan bahwa sebesar 45% konsumen Tokopedia memberikan rating 5 dan diikuti 31% konsumen yang memberikan rating 1. Kemudian rating yang paling sedikit diberikan yaitu 4.



Gambar 7. Visualisasi Word Cloud Tokopedia

Berdasarkan gambar word cloud diatas menunjukan bahwa semakin besar ukuran tulisan berarti semakin banyak kata tersebut muncul. Dari banyaknya kata yang muncul penulis membuat sebuah grafik yang berisi *review* negatif dari Tokopedia untuk mengetahui presentase kekurangan dari Tokopedia yang dianalisis dalam bentuk diagram lingkaran sebagai berikut.



Gambar 8. Diagram Review Negatif Tokopedia

Data Labeling

Pelabelan pada kelas sentiment hanya menggunankan kategori negatif dan kategori positif adapun hasil dari pelabelan untuk Tokopedia dan Shopee dapat di tunjukan sebagai berikut :

Tabel 4. Jumlah Pelabelan Kelas Sentimen

Sentimen	*Review* Shopee	*Review* Tokopedia
Positif	32	40
Negatif	48	40

Berdasarkan rekap jumlah pelabelan kelas sentiment diperoleh hasil sentiment negatif untuk *review* Shopee sebesar 32 ulasan dan Tokopedia sebesar 40 ulasan. *Review* diklasifikasikan sebagai sentiment positif apabila mengandung pernyataan positif seperti pujian, ungkapan terimakasih, atau testimony positif tentang *E-Commerce*. Suatu ulasan diklasifikasikan sebagai sentiment negatif bila mengandung pernyataan-pernyataan negatif seperti ketidakpuasan, penghinaan, laporan kegagalan layanan, dan sebagainya. Hasil dari pelabelan dapat dilihat pada tabel berikut ini :

Tabel 5. Hasil Pelabelan Sentimen Menggunakan Kamus Lexicon

Sentimen	Skor	*Review*
Positif	**2**	benar bagus sekali
Positif	**7**	pesan mudah ngga ribet simple barang jual terpercaya beli mudah cari barang cari foto barang sesuai datang game hadiah poin lain mudah laku transaksi opsi laku pembayaran mudah beli kalangan diskon barang jual variatif harga variatif harga murah mahal
Negatif	**-5**	komplain s pinjam bayar tagih jatuh tempo ga bayar sesuai tagihan kena denda Shopee jelek
Negatif	**-4**	sangat kecewa sekali akun deteksi akun orang saya tdk kenal menjelaksan Tokopedia care

	kirim lampir e-ktp tdk solusi adakan promo ujung2nya transaksi batal susah usaha Tokopedia sistem ngaco tim it coba cek sistem bikin susah langganan

Contoh perhitungan kelas sentiment berdasarkan *review* "pesan mudah ngga ribet simple barang jual terpercaya beli mudah cari barang cari foto barang sesuai datang game hadiah poin lain mudah laku transaksi opsi laku pembayaran mudah beli kalangan diskon barang jual variatif harga variatif harga murah mahal", terdapat 9 kata positif yakni "mudah","simple","terpercaya","mudah","mudah","variatif","variatif,"murah" dan 2 kata negatif yakni "ribet", dan "mahal". Adapun rumus perhitungan skor sentiment yang digunakan dalam proses pelabelan adalah sebagai berikut:

Skor = (Jumlah kata positif – Jumlah kata negatif)

Sehingga dengan demikian diperoleh perhitungan sebagai berikut :

Skor = (Jumlah kata positif) – (Jumlah kata negatif)

Skor = 9-2=7

Naïve Bayes Classifier

Langkah pertama dalam mengklasifikasikan data *review* adalah melatih model menggunakan data training. Data training dari setiap media yang telah dilakukan preproses digunakan untuk melatih model menggunakan software RStudio. Model yang telah dilatih dengan data training kemudian digunakan untuk mengklasifikasikan data testing ke dalam dua kelas sentimen, positif dan negatif. Pembagian data training dan testing sebesar 80% dan 20%. Klasifikasi dengan metode Naïve Bayes Classifier menghasilkan probabilitas yang digunakan untuk menentukan apakah *review* masuk ke dalam kategori sentimen positif atau negatif. Probabilitas tersebut diperoleh dengan persamaan (2.6) dan persamaan (2.7).

Perhitungan probabilitas setiap kategori *review* seperti yang telah dijelaskan ilustrasi dibawah ini:

Tabel 5. Ilustrasi Struktur Data Naïve Bayes Classifier

Review	Variabel Prediktor									
	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**
1	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
2	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**1**

Keterangan :

Review 1 (Positif) : "rekomendasi cari barang original pilihan toko variatif tampilan aplikasi nyaman tidak pernah ada kendala"

Review 2 (Negatif) : "tidak suka Shopee mohon maaf , kirim ekspedisi barang lambat sampai rumah, barang tersesat ke rumah orang ,kecewa"

Variabel Prediktor :

X1 : rekomendasi
X2 : original
X3 : cepat
X4 : nyaman
X5 : suka
X6 : lancar
X7 : parah
X8 : lambat
X9 : tersesat
X10 : kecewa

Berikut merupakan ilustrasi perhitungan untuk melakukan klasifikasi sentimen menggunakan metode Naïve Bayes Classifier dengan contoh *review* pada Tabel 4.5 sebagai data training dan contoh *review* "barang yang dikirim original tapi opsi jasa kirim pilih alih shoope express lambat sekali tolong ubah sistem kecewa " Perhitungan dilakukan untuk mengklasifikasikan apakah contoh *review* sebagai data testing tersebut memiliki sentimen positif atau negatif. Hal pertama yang dilakukan adalah menghitung probabilitas setiap kelas sentimen dengan persamaan (2.6)

$$P(v1) = \frac{|\boldsymbol{Doc1}|}{|\boldsymbol{Training}|} = \frac{1}{2} = 0{,}5$$

$$P(v2) = \frac{|\boldsymbol{Doc2}|}{|\boldsymbol{Training}|} = \frac{1}{2} = 0{,}5$$

Dimana $\boldsymbol{P(v1)}$adalah probabilitas sentimen positif dan $\boldsymbol{P(v2)}$ adalah probabilitas sentimen negatif. Kemudian dilakukan perhitungan probabilitas kemunculan setiap kata pada masing-masing kategori dengan persamaan (2.7):

Tabel 6. Probabilitas Kemunculan Kata

X1 Rekomendasi $P(v1) = \frac{1+1}{3+10} = 0,1538$ $P(v2) = \frac{0+1}{4+10} = 0,0714$	**X6 Lancar** $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v2) = \frac{0+1}{4+10} = 0,0714$
X2 Original $P(v1) = \frac{1+1}{3+10} = 0,1538$ $P(v2) = \frac{0+1}{4+10} = 0,0714$	**X7 Parah** $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v2) = \frac{0+1}{4+10} = 0,0714$
X3 Cepat $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v2) = \frac{0+1}{4+10} = 0,0714$	**X8 Lambat** $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v1) = \frac{1+1}{4+10} = 0,1428$
X4 Nyaman $P(v1) = \frac{1+1}{3+10} = 0,1538$ $P(v2) = \frac{0+1}{4+10} = 0,0714$	**X9 Tersesat** $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v1) = \frac{1+1}{4+10} = 0,1428$
X5 Suka $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v1) = \frac{1+1}{4+10} = 0,1428$	**X10 Kecewa** $P(v1) = \frac{0+1}{3+10} = 0,0769$ $P(v1) = \frac{1+1}{4+10} = 0,1428$

Selanjutnya adalah mencari probabilitas tertinggi dari *review* yang diujikan. *Review* testing setelah dilakukan praproses teks, maka terdiri dari kata "original", "lambat", "kecewa", Sehingga dicari probabilitas tertinggi dari setiap kata pada *review* tersebut menggunakan persamaan (2.5).

$P(v1)\prod_{i=1}^{n} p(a1|v1)$
$= (0.5)(P(original|v1)x(P(lambat|v1)x(P(kecewa|v1))$
$= (0.5)(0,1538x\ 0,0769\ \_\ 0,0769)$
$= 0,000457$

$P(v2)\prod_{i=1}^{n} p(a1|v2)$
$= (0.5)(P(original|v2)x(P(lambat|v2)x(P(kecewa|v2))$
$= (0.5)(0,0714x\ 0,1428\ x\ 0,1428)$
$= 0,000727$

Nilai probabilitas kata setiap *review* testing yang terbesar adalah probabilitas setiap kata pada sentimen negatif sehingga *review* testing tersebut diklasifikasikan sebagai *review* dengan sentimen negatif.

Tabel 7. Confusion Matrix Klasifikasi Shopee Tanpa Seleksi Fitur Information Gain

	Prediksi Negatif	Prediksi Positif	Jumlah
Kenyataan Negatif	0	48	48
Kenyataan Positif	0	32	32
Jumlah	0	80	80

Berdasarkan tabel Confusion Matriks diatas, sebanyak 40% konsumen menilai positif Shopee serta sebanyak 32 *review* sentimen positif diprediksi benar dengan menggunakkan naïve bayes classifier. Prediksi sentimen positif benar seluruhnya dan prediksi sentimen negatif salah diprediksi seluruhnya yaitu 48 *review*.

Tabel 8. Confusion Matrix Klasifikasi Tokopedia Tanpa Seleksi Fitur Information Gain

	Prediksi Negatif	Prediksi Positif	Jumlah
Kenyataan Negatif	0	40	40
Kenyataan Positif	0	40	40
Jumlah	0	80	80

Berdasarkan tabel Confusion Matriks diatas, sebanyak 50% konsumen menilai positif Tokopedia serta sebanyak 40 *review* sentimen positif diprediksi benar dengan menggunakkan naïve bayes classifier. Prediksi sentimen positif benar seluruhnya dan prediksi sentimen negatif salah diprediksi seluruhnya yaitu 40 *review*.

Tabel 9. Hasil Klasifikasi dengan Menggunakan Naïve Bayes Classifier

E-Commerce	**Accuracy**
Tokopedia	50 %
Shopee	40 %

Berdasarkan tabel hasil klasifikasi Naïve Bayes Classifier dalam analisis sentiment terhadap *review E-Commerce* ini diperoleh nilai akurasi

sebesar 50% untuk Tokopedia dan 40% untuk Shopee.

Seleksi Fitur Information Gain

Setelah dilakukan pelabelan sentimen selanjutnya dilakukan proses seleksi fitur information gain untuk mendapatkan fitur terbaik dan meningkatkan akurasi. Pembahasan berikut merupakan contoh perhitungan seleksi fitur information gain secara manual. Dengan cara menghitung bobot sesuai rumus 2.3 pada bab 2. Metode Information Gain secara sederhana dapat dicontohkan seperti tabel berikut :

Tabel 10. Contoh Koleksi Data

Dokumen	Fitur			Sentimen
	Terima Kasih	Kecewa	Puas	
D_1	Tidak	Ya	Tidak	Negatif
D_2	Ya	Ya	Tidak	Positif
D_3	Tidak	Ya	Tidak	Negatif
D_4	Tidak	Ya	Tidak	Negatif
D_5	Tidak	Ya	Tidak	Negatif
D_6	Tidak	Ya	Tidak	Negatif
D_7	Tidak	Ya	Tidak	Negatif
D_8	Ya	Tidak	Ya	Positif
D_9	Tidak	Ya	Tidak	Negatif
D_{10}	Ya	Ya	Tidak	Positif

Fitur yang terdapat pada tabel 4.4 merupakan potongan kata dari dokumen yang akan dihitung bobotnya. Pada kasus perhitungan bobot information gain penulis mengambil contoh kata “Kecewa” dengan menghitung entropy pada dataset dengan menggunakan peramaan 2.1 , sebagai berikut :

$$Entropy\ (Set) = -\left[\left(\frac{9}{10}\right) log2\left(\frac{9}{10}\right) + \left(\frac{1}{10}\right) log2\left(\frac{1}{10}\right)\right] = 0{,}4689$$

Selanjutnya ambil contoh pada kata “kecewa” yang memiliki value “Ya” atau “Tidak” sehingga bisa dihitung dengan persamaan 2.1 . Setelah itu hitung Entropy (S_{Kecewa}) dengan persamaan 2.2 dan menghasilkan perhitungan sebagai berikut :

$$Entropy\ (Positif) = -\left[\left(\frac{2}{9}\right) log2\left(\frac{2}{9}\right) + \left(\frac{7}{9}\right) log2\left(\frac{7}{9}\right)\right] = 0{,}7642$$

$$Entropy\ (Negatif) = -\left[\left(\frac{1}{1}\right) log2\left(\frac{1}{1}\right) + \left(\frac{0}{1}\right) log2\left(\frac{0}{1}\right)\right] = 0$$

$$Entropy\ (Skecewa) = \left(\frac{9}{10}\right) 0{,}7642 - \left(\frac{1}{10}\right) 0 = 0{,}68778$$

Langkah terakhir untuk mencari nilai information gain menggunakan persamaan 2.3 sebagai berikut :

$$Gain\ (Skecewa) = Entropy\ (Set) - (Entropy\ (S, A)$$

$$Gain\ (Skecewa) = 0{,}4689 - 0{,}68778 = -0{,}2188$$

Dengan bobot Information gain tersebut setiap kata atau fitur akan dirangking dengan cara mengurutkan nilai yang terbesar hingga terkecil dan hasilnya akan didapatkan fitur yang terbaik. Untuk keperluan klasifikasi data testingnya akan diambil fitur dengan akurasi tertinggi.

Tabel 11. Confusion Matrix Klasifikasi Shopee dengan Seleksi Fitur Information Gain

	Prediksi Negatif	Prediksi Positif	Jumlah
Kenyataan Negatif	44	3	47
Kenyataan Positif	25	8	33
Jumlah	69	11	80

Berdasarkan tabel Confusion Matriks diatas, sebanyak 41,25 % konsumen menilai positif Shopee serta sebanyak 8 *review* sentimen positif diprediksi benar dan prediksi sentimen negatif salah diprediksi sebanyak 3 *review*

dengan menggunakkan naïve bayes classifier dengan seleksi fitur *Informatin Gain*.

Tabel 12. Confusion Matrix Klasifikasi Tokopedia dengan Seleksi Fitur Information Gain

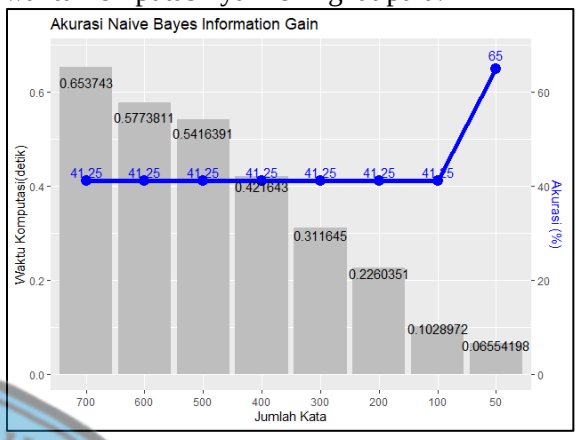
	Prediksi Negatif	Prediksi Positif	Jumlah
Kenyataan Negatif	25	12	37
Kenyataan Positif	6	37	43
Jumlah	31	49	80

Berdasarkan tabel Confusion Matriks diatas, sebanyak 53,75% konsumen menilai positif Tokopedia serta sebanyak 6 *review* sentimen positif salah diprediksi dan prediksi sentimen negatif benar diprediksi sebanyak 25 *review* dengan menggunakkan naïve bayes classifier dengan seleksi fitur *Informatin Gain*.

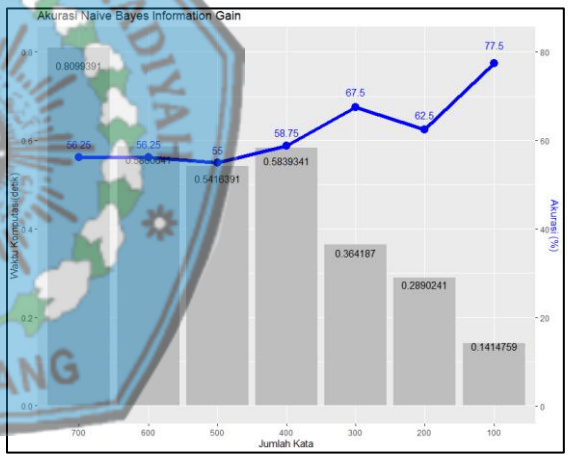
Tabel 13. Hasil Klasifikasi Naïve Bayes Classifier dengan Seleksi Fitur Information Gain

E-Commerce	**Jumlah Kata Ranking Teratas**							
	7000	**6000**	**5000**	**4000**	**3000**	**2000**	**100**	**50**
Tokopedia	56,26	56,26	55	58,75	66,75	62,5	77,5	-
Shopee	41,25	41,25	41,25	41,25	42,5	41,5	43,75	65

Berdasarkan tabel hasil klasifikasi Naïve Bayes Classifier dengan Seleksi Fitur Information Gain diatas dapat dijelaskan bahwa dengan adanya kombinasi seleksi fitur information gain dapat meningkatkan akurasi dari Naïve Bayes Classifier. Berdasarkan jumlah kata dengan rangking teratas menunjukkan bahwa 100 fitur terbaik pada Tokopedia dan 50 fitur terbaik pada Shopee yang memiliki nilai Information Gain tertinggi, memperoleh tingkat akurasi yang paling besar sedangkan untuk 700 kata dengan ranking teratas menunjukkan bahwa hampir tidak ada perubahan dengan tingkat akurasinya karena hal ini sama saja dengan pengujian tanpa menggunakan seleksi fitur sebab 700 kata tersebut adalah hampir semua kata yang ada pada data. Adapun grafik yang mengilustrasikan tingkat akurasi dari metode *Naïve Bayes Classifier* dengan Seleksi Fitur *Information gain* untuk analisis sentiment terhadap *review E-Commerce* sebagai berikut. Grafik dibawah ini dapat dijelaskan bahwa semakin sedikit fitur (rangking teratas) yang digunakan akurasinya semakin meningkat dan waktu komputasinya meningkat pula.



Gambar 9. Grafik Shopee Naïve Bayes Classifier dengan Seleksi Fitur Information Gain



Gambar 10. Grafik Tokopedia Naïve Bayes Classifier dengan Seleksi Fitur Information Gain

Pembahasan

Berdasarkan analisis sentiment yang dilakukan hampir sebagian memberikan penilaian positif kepada Tokopedia dan Shopee. Sebanyak 50% pengguna Tokopedia memberikan penilaian positif. Tokopedia juga mendapatkan rating bintang 5 sebanyak 45% dibandingkan Shopee yang hanya 11% saja. Hasil klasifikasi yang dilakukan dengan Naïve Bayes Classifier diperoleh akurasi 50% untuk Tokopedia dan 40% untuk Shopee. Kemudian setelah dikombinasikan dengan Seleksi Fitur *Information Gain* akurasinya meningkat

sebesar 77,5% untuk Tokopedia dan 65% untuk Shopee dengan menggunakan fitur terbaik dengan 100 rangking teratas pada Tokopedia dan 50 rangking teratas untuk Shopee. Kata-kata yang paling banyak ditulias pada *review* negatif Tokopedia antara lain costumer kecewa sebesar 35% dan barang rusak sebesar 23% sedangkan untuk Shopee 47% costumer kecewa dan 20% belanja ribet. Berdasarkan hasil klasifikasi dari asosiasi teks yang dilakukan, secara umum dapat diketahui bahwa pengguna aplikasi Tokopedia dan Shopee mayoritas membicarakan menganai barang, transaksi karena selalu muncul pada kelas sentiment positif maupun negatif. Secara umum metode asosiasi teks yang digunakan menunjukan hasil ulasan negatif yang sering muncul diantaranya konsumen kecewa dengan barang yang rusak, transaksi yang ribet dan aplikasi yang lambat.

KESIMPULAN DAN SARAN

Kesimpulan

1. Hasil deskripsi review terkait Tokopedia dan Shopee pada Google Playstore dengan jumlah *review* atau tanggapan sebanyak 80 *review* untuk Tokopedia dan 80 *review* untuk Shopee dari 18 Maret – 7 April 2021. Dalam *review* tersebut Metode Naïve Bayes Classifier benar memprediksi sentiment positif untuk Tokopedia sebanyak 40 *review*. Sedangkan untuk Shopee sentiment positif diprediksi benar sebanyak *32 review* Kemudian setelah dikombinasikan dengan seleksi fitur Information Gain pada Tokopedia hanya salah diprediksi sebesar *6 review* untuk sentiment positif dan *12 review* untuk sentiment negative. Pada Shopee salah diprediksi sebesar *25 review* untuk sentiment positif dan *3 review* untuk sentiment negatif
2. Tingkat akurasi yang diperoleh dari hasil klasifikasi menggunakan metode Naïve Bayes Classifier dengan pembagian data training dan data testing yaitu 80%:20% diperoleh nilai akurasi sebesar 50% untuk Tokopedia dan 40% untuk Shopee, sedangkan setelah dikombinasikan dengan Seleksi Fitur Information Gain akurasinya meningkat sebesar 77,5% untuk Tokopedia dan 65% untuk Shopee dengan menggunakan fitur terbaik dengan 100 rangking teratas untuk Tokopedia dan 50 fitur teratas untuk Shopee.

Saran

1. Penambahan koleksi kamus pada *database* kata gaul, karena pada *review Google Play Store* terlalu banyak Bahasa yang kurang baku.
2. Data yang digunakan hanya data yang dibatasi oleh *Google Play Store* sehingga perlu ditambahkan data agar hasil klasifikasi sentimen yang lebih baik.
3. Pada penelitian selanjutnya peneliti menyarankan menambahkan metode yang bisa meningkatkan akurasi dari algoritma *Naïve Bayes Classifier* dan Seleksi Fitur *Information Gain.*
4. Pembagian data training menggunakan *K-fold Cross Validation* agar nilai akurasi tidak berubah.

DAFTAR PUSTAKA

Aini, S H A, Yuita Arum Sari & Achmad Arwan. 2018.*Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes* . Program Studi Teknik Informatika. Fakultas Ilmu Komputer. Universitas Brawijaya. Malang.

Feldman, R., & Sanger, J. 2007. The Text Mining Handbook Advanced Approaches In Analyzing Unstructured Data. New York : Cambridge http://repository.unimus.ac.id 50 University Press.

Khotimah, N. 2019. *Analisis Sentimen Terhadap Review E-Commerce Dengan Metode Stochastic Gradient Descent.*Skripsi. Program Studi Statistika Universitas Muhammadiyah Semarang. Semarang.

Liu, B. 2012. *Sentiment Analysis and Subjectivity*. Synthesis Lectures on Human Language Technologies. USA: Morgan & Claypool Publishers.

Liu, Y., Loh, H. T., & Sun, A. 2009. Imbalanced text classification: A term weighting approach. Expert Systems with Applications, 36(1), 690–701. https://doi.org/10.1016/j.eswa.2007.10.042

Maulida, I., Suyatno, A., & Hatta, H. R. (2016). *Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain*. *JSM STMIK*

Mujilahwati, S. *2016. Preprocessing Text Mining pada Twitter. Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA)* .

Saraswati, N.S. 2011. *Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis*. Skripsi. Program Studi Teknologi Informasi Fakultas Teknik UGM Yogyakarta.

Syakuro, A. 2017. *Analisis Sentimen Masyarakat Terhadap E-Commerce Pada Media Sosial Menggunakan Metode Naïve Bayes Classifier (Nbc) Dengan Seleksi Fitur Information Gain (Ig)*. Skripsi. Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim. Malang.

Yulianita, Tanti. 2020. *Analisis Sentimen Publik Terhadap Kebijakan Pemerintah Dalam Penanganan Covid-19 Di Indonesia Menggunakan Naïve Bayes Classifier Dan K-nearst Neighbour*. Skripsi. Program Studi Statistika Universitas Muhammadiyah Semarang. Semarang.

Yosmita, Ditia. 2018. *Analisis Sentimen Online Review Pengguna E-Commerce Menggunakan Metode Support Vector Machine dan Maximum Entropy*. Skripsi. Program Studi Statistika Universitas Islam Indonesia. Yogyakarta.

Zaki, M. J., & Meira, W. J. 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms, 562. Retrieved from https://books.google.com/books?hl=en&lr=&id=Gh9GAwAAQBAJ&pgis=1