

Analisis Sentimen Review Halodoc Menggunakan Naïve Bayes Classifier

Asep Hendra ^{(1)*}, Fitriyani ⁽²⁾

Sistem Informasi, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya,
Bandung

e-mail : ace.hendra07@gmail.com, fitriyani@ars.ac.id.

* Penulis korespondensi.

Artikel ini diajukan 5 Agustus 2020, direvisi 20 September 2020, diterima 13 Oktober 2020, dan dipublikasikan 3 Mei 2021.

Abstract

Healthcare service has the role to help and serve people to access medical services, i.e. providing medicines, medical consultation, or health control. Healthcare service has been transforming to a digital platform. Halodoc is one of the digital platforms that people can use for free or paid, user can also give reviews of Halodoc's performance and services on Google Play Store to give feedback that Halodoc can use to evaluate and improve the app. The Google Play Store review is increasing every day. Therefore an analysis for the review with sentiment analysis for Halodoc's review is needed, first phase of sentiment analysis for the review is preprocessing which has tokenization, transform to lower cases, filter stopword, dan filter token (by length) processes. The data is divided into two positive and negative classes with cross-validation and a k-fold validation value of 10, using Naïve Bayes Classifier algorithm with 81,68% accuracy and AUC 0.756, categorized as fair classification.

Keywords: Naïve Bayes Classifier, Text Mining, Sentiment Analysis, Classification, Halodoc

Abstrak

Layanan kesehatan yang memiliki peran untuk melayani dan membantu masyarakat dalam memanfaatkan layanan kesehatan, baik itu untuk berobat ketika sakit, konsultasi kesehatan atau untuk kontrol kesehatan saja. Layanan kesehatan kini bertransformasi menjadi layanan kesehatan digital. Halodoc merupakan aplikasi layanan digital yang dapat digunakan oleh masyarakat secara gratis maupun berbayar, penggunaannya dapat memberikan ulasan terhadap kinerja atau layanan Halodoc melalui Google Play Store untuk menjadi evaluasi dan peningkatan kinerja Halodoc. Ulasan pada Google Play Store kian hari semakin meningkat, oleh karena itu diperlukan analisis ulasan dengan melakukan analisis sentimen terhadap review Halodoc, tahapan awal dalam analisis sentimen adalah *preprocessing* yang di dalam prosesnya terdapat proses *tokenization*, *transform to lower cases*, *filter stopword*, *filter token (by length)*. Data dibagi menjadi dua kelas yaitu positif dan negatif dengan validasi *cross validation* dan nilai *k-fold validation* 10, algoritma yang digunakan adalah *Naïve Bayes Classifier* dengan hasil akurasi 81.68% dan AUC 0.756, termasuk ke dalam *fair classification*.

Kata Kunci: Naïve Bayes Classifier, Text Mining, Analisis Sentimen, Klasifikasi, Halodoc

1. PENDAHULUAN

Teknologi informasi semakin pesat perkembangannya, teknologi informasi memasuki berbagai bidang seperti: pendidikan, ekonomi, sosial budaya, kesehatan dan lain-lain. Memasuki era 4.0 yang dihadapkan sekarang maka tidak heran masyarakat mau tidak mau harus siap dengan perubahan atau perkembangan zaman. Pengaruh kemajuan teknologi mengantarkan masyarakat akan adanya perubahan dari analog menuju digital, kemajuan teknologi seperti televisi, telepon dan telepon genggam, bahkan internet bukan hanya dapat dinikmati oleh masyarakat kota, namun juga telah dapat dinikmati oleh masyarakat di pelosok desa. Pengguna internet di Indonesia sendiri mencapai 171,71 juta jiwa, dengan penetrasi internet sebesar 64,8% pada tahun 2018. Persentase pertumbuhan pengguna selama satu tahun mencapai 10,12%, kemudian pertumbuhan pengguna internet selama periode 2017-2018 mencapai 27,916 juta lebih dari jumlah penduduk pada tahun 2018 sebesar 264,161 juta jiwa (APJII, 2019).



Di antara berbagai sektor yang terdampak oleh era 4.0, tampaknya sektor kesehatan adalah sektor yang paling mungkin mendapatkan keuntungan dari bergabungnya sistem fisika, digital dan biologi, walaupun sektor ini mungkin juga yang paling tidak siap menerimanya. Hal ini diperkuat dari hasil survei terhadap 622 pemimpin bisnis dari berbagai industri di seluruh dunia oleh *The Economist Intelligence Unit*. Jajak pendapat terhadap para pemimpin bisnis ini menunjukkan bahwa mayoritas yang signifikan dari para eksekutif tersurvei percaya bahwa kesehatan adalah sektor yang akan mendapatkan keuntungan besar dari dampak era 4.0 ini (Tjandrawinata, 2016).

Menurut survei yang dilakukan oleh Ulya (2019), persentase keuntungan dari dampak era 4.0 untuk bidang kesehatan mencapai angka 51,06%, angka ini dinilai cukup besar dan menandakan bahwa masyarakat saat ini dalam memenuhi kebutuhan akan informasi kesehatan melalui *smartphone* masing-masing. Layanan kesehatan digital telah banyak berkembang dalam 2 tahun terakhir, sebut saja Halodoc, Klikdokter, dan beberapa layanan digital yang terintegrasi dengan lembaga kesehatan seperti Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan yang penggunaannya mengalami peningkatan. Pertumbuhan pengguna yang tumbuh signifikan ini merupakan bukti layanan kesehatan digital sudah digemari masyarakat, khususnya di era disrupsi teknologi yang membuat kebiasaan hidup dan perilaku seseorang berubah. Hal tersebut juga diperkuat oleh sebuah survei yang dilakukan oleh *Deloitte* Indonesia bekerja sama dengan Bahar dan *Center for Healthcare Policy and Reform Studies (Chapters)* Indonesia. Berdasarkan survei tersebut, sekitar 84,4% pengguna layanan kesehatan digital mengaku puas dengan layanan yang ada.

Analisis sentimen adalah sebuah metode yang digunakan untuk mengekstrak data opini, memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terkandung dalam sebuah opini (Sari & Wibowo, 2019). Gagasan di balik analisis sentimen adalah untuk terhubung ke ribuan sumber online di internet, mengumpulkan pernyataan tentang merek atau produk, dan analisis dengan cara analisis teks sehubungan dengan sentimen. Hal tersebut menjadi wawasan baru tentang bagaimana cara mengidentifikasi perubahan sentimen dari waktu ke waktu terhadap keberhasilan pemasaran dan bagaimana penyedia layanan dapat meningkatkan reputasi produknya (Hofmann & Klinkenberg, 2013).

Penggunaan analisis sentimen dalam kategori *Fined-grained sentiment analysis* digunakan untuk menganalisis sesuatu atau produk untuk dapat dilihat reputasi dari suatu produk dan untuk meningkatkan kualitas produk ke depannya. Terdapat banyak ulasan pengguna yang tersedia dari ulasan aplikasi Halodoc, ada ulasan positif ada juga ulasan negatif. Semakin banyak ulasan pengguna yang tersedia maka semakin sulit pula calon pengguna untuk menyimpulkan hasil ulasan, sehingga diperlukan adanya klasifikasi ulasan untuk menyimpulkan ulasan dari aplikasi, untuk dapat membantu calon pengguna dalam mengambil keputusan dari ulasan pengguna aplikasi Halodoc. Oleh karena itu, dalam penelitian ini akan dilakukan analisis sentimen dengan menggunakan algoritma *Naïve Bayes Classifier*, algoritma *Naïve Bayes Classifier* dipilih karena *Naïve Bayes Classifier* merupakan metode klasifikasi yang efisien dan sederhana (Nugroho et al., 2020). Selain itu, algoritma *Naïve Bayes Classifier* adalah algoritma dengan performa yang sangat baik dalam beberapa kasus klasifikasi teks (Taufik, 2017). Tujuan dari penelitian ini adalah untuk mengetahui akurasi pada ulasan Halodoc dan mengetahui kinerja algoritma *Naïve Bayes Classifier* dalam melakukan klasifikasi analisis sentimen Halodoc.

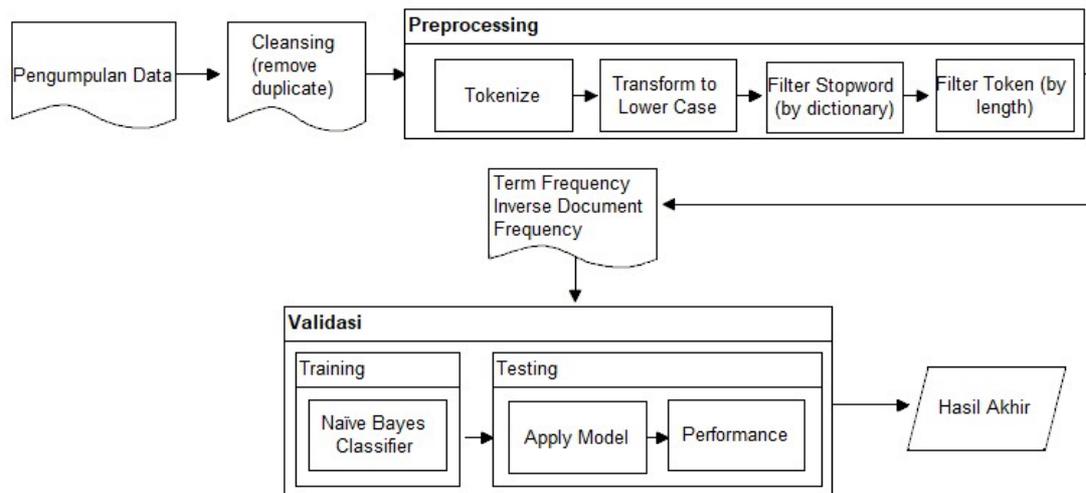
2. METODE PENELITIAN

Penelitian ini mengusulkan metode *Naïve Bayes Classifier* (NBC) sebagai algoritma klasifikasi. Pertama adalah pengumpulan data yang didapat dari laman Google Play Store, kemudian tahap selanjutnya melakukan *preprocessing* dengan *tokenize* ini bertujuan membuat kalimat untuk dijadikan token-token atau pemenggalan kata-kata dengan menggunakan spasi. Kemudian dilakukan *transform to lower case*, tahap ini bertujuan untuk membuat huruf dari suatu kata menjadi seragam, setelah *transform to lower case* selanjutnya menghapus kata yang tidak memiliki relevansi tinggi dengan menggunakan *stopword* dan melakukan penghapusan kata



dalam dokumen yang tidak mencapai batas minimum dan melebihi batas maksimum, dalam penelitian ini menggunakan batas minimum 4 karakter dan batas maksimum 25 karakter.

Setelah tahap *preprocessing* tahap selanjutnya melakukan *term weighting* dengan menggunakan TF-IDF, Kemudian akan dilakukan validasi silang (*Cross Validation*) dengan *k-fold validation* untuk mendapatkan hasil *accuracy* yang maksimal. Desain model yang diusulkan akan melalui pemrosesan data *training* dan data *testing* pada tahap validasi diterapkan algoritma *Naïve Bayes Classifier* pada bagian *training* dan pada bagian *testing* memasukan operator *apply model* dan *performance* untuk menghasilkan performa algoritma yang telah diterapkan, dan yang terakhir adalah evaluasi dan validasi hasil. Berikut gambar desain penelitian pada Gambar 1.



Gambar 1. Desain Penelitian.

2.1. Pengumpulan Data

Penelitian ini menggunakan data *review* Google Play Store pada aplikasi Halodoc sebanyak 950 data sentimen. Data tersebut didapatkan dari laman *review* Google Play Store, kemudian diolah menggunakan Microsoft Excel agar data dapat diolah ke tahap selanjutnya.

2.2. Data Cleansing (Remove Duplicate)

Remove duplicate bertujuan untuk menghilangkan data yang sama. Sehingga dapat mencegah adanya duplikasi data dan dapat mengurangi jumlah *term*. *Remove duplicate* bekerja pada semua jenis atribut.

2.3. Naïve Bayes Classifier

Naïve Bayes Classifier merupakan algoritma *machine learning* yang sederhana dan cepat dalam hal klasifikasi, kinerja yang baik dan mudah dalam penerapannya (Fitriyani, 2018). Kaitan antara *Naïve Bayes Classifier* dengan klasifikasi, kolerasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam *teorema bayes* merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi (Prasetyo, 2012). Konsep dasar yang digunakan oleh *bayes* adalah teorema peluang bersyarat *bayes* aturan *bayes* dapat dinyatakan Samodra et al dalam (Junianto & Riana, 2017):

$$P(C_j) = \frac{P(C_j)P(D|C_j)}{P(D)} \quad (1)$$

Di mana C_j adalah kategori teks yang akan diklasifikasikan, dan $P(C_j)$ merupakan probabilitas dari kategori teks C_j . Sedangkan d merupakan dokumen teks yang dapat direpresentasikan



sebagai himpunan kata (w_1, w_2, \dots, w_n) , dimana w_1 adalah kata pertama, w_2 adalah kata kedua dan seterusnya. Pada saat proses pengklasifikasian dokumen teks, maka pendekatan *bayes* akan menyeleksi kategori teks yang mempunyai probabilitas tinggi (C_{map}) yaitu:

$$C_{map} = \operatorname{argmax}_{C_j} \frac{P(C_j)P(D|C_j)}{P(D)} \quad (2)$$

Nilai $P(D)$ dapat diabaikan karena nilainya adalah konstan untuk semua C_j , sehingga persamaan (2) dapat disederhanakan sebagai berikut:

$$C_{map} = \operatorname{argmax}_{C_j} P(C_j)P(D|C_j) \quad (3)$$

Probabilitas $P(C_j)$ dapat diestimasi dengan cara menghitung jumlah dokumen *training* pada tiap-tiap kategori teks C_j . Kemudian untuk menghitung distribusi $P(D|C_j)$ sangat sulit untuk dilakukan khususnya pada proses pengklasifikasian dokumen teks yang berjumlah besar, karna jumlah *term* $P(D|C_j)$ bisa menjadi sangat besar. Hal ini dikarenakan jumlah *term* tersebut sama dengan jumlah kombinasi posisi kata dikalikan dengan jumlah kategori yang akan diklasifikasikan.

Pendekatan *Naïve Bayes* yang mengasumsikan bahwa setiap kata di dalam setiap kategori adalah tidak bergantung satu sama lain, maka perhitungan dapat lebih di sederhanakan lagi sebagai berikut:

$$P(D|C_j) = \prod_i P(W_i|C_j) \quad (4)$$

Dengan menggunakan persamaan (4) maka persamaan (3) dapat dituliskan menjadi sebagai berikut:

$$C_{map} = \operatorname{argmax}_{C_j} P(C_j) \prod_i P(W_i|C_j) \quad (5)$$

Nilai $P(C_j)$ dan $P(W_i|C_j)$ dihitung ketika proses *training* dijalankan yaitu:

$$P(C_j) = \frac{n(W_j)}{n(\text{Sampel})} \quad (6)$$

$$P(W_i|C_j) = \frac{1+n_j}{|C|+n(\text{kosakata})} \quad (7)$$

Di mana $n(W_j)$ adalah jumlah kata pada kategori j , dan n (sampel) adalah jumlah dokumen sampel yang digunakan dalam proses *training*. Sedangkan n_j adalah jumlah kemunculan dokumen kata W_j pada kategori C_j , $|C|$ adalah jumlah semua kata pada kategori C_j dan $n(\text{kosakata})$ adalah jumlah kata yang unik pada semua data *training*.

2.4. Preprocessing

Tahap *preprocessing* merupakan titik awal dan tahap yang penting dalam klasifikasi (Durairaj & Ramasamy, 2016), pada tahap ini adalah tahap proses untuk mempersiapkan data sebelum masuk ke pemodelan. Tahapan ini meliputi:

2.4.1. Tokenization

Tahap ini adalah proses memecah kumpulan kalimat-kalimat menjadi kata, sekaligus menghilangkan simbol khusus dan tanda baca yang kemudian terbentuk suatu kumpulan kata yang bersifat unik.



2.4.2. Transform case

Setelah tahap mengubah kalimat menjadi pecahan kata, kemudian tahap selanjutnya mengubah semua karakter khususnya huruf akan diubah menjadi huruf non kapital. Tahap ini dimaksudkan agar data yang akan dimasukkan, memiliki susunan kata dengan struktur huruf yang sama, misalnya "Dokter", "dOkter", "doKter", "dokTer", "doktEr", "dokteR", "Dokter", "doktER", dan seterusnya akan diubah menjadi kata yang sama yaitu "dokter".

2.4.3. Filter Token (by length)

Tahap ini adalah proses dilakukannya pemilihan token sesuai yang kita butuhkan, dalam penelitian ini membatasi token dengan ukuran minimal 4 karakter (huruf) dan batas maksimum yang digunakan 25 karakter (huruf). Contohnya kata seperti: "i", "u", "gw", "km", "sdg", "rsp" adalah kata yang kurang dari 4 huruf, selanjutnya kata semisal akan dihapus dalam tahap ini.

2.4.4. Filter Stopword

Tahap ini adalah tahap mengambil kata-kata penting dari hasil *token* dengan membuang kata yang kurang penting atau menyimpan kata yang penting (*wordlist*). *Stopword* biasanya bersifat kata umum yang sering muncul dalam beberapa kalimat dengan jumlah yang besar dan merupakan kata sambung, contohnya: "adalah", "ke", "di", "dari", "dan". Inti dari *stopword* adalah menghapus kata-kata yang memiliki nilai informasi yang rendah atau yang tidak memiliki relevansi dengan isi dari dokumen.

2.5. TF-IDF

Setelah tahap *preprocessing* selesai, maka tahap selanjutnya adalah menerapkan perhitungan bobot menggunakan TF-IDF. Tahap *term weighting* merupakan perhitungan *term frequency*, seberapa sering masing-masing kata muncul dalam suatu dokumen. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus (Maarif, 2015):

$$w = tf_{dt} \times IDF_t \quad (8)$$

Di mana w merupakan *weight* (bobot) dokumen ke- d terhadap dokumen ke- t sama dengan banyaknya kata yang dicari dalam sebuah dokumen dikali dengan *Inverse Document Frequency* kata ke- t dari kata kunci. Nilai IDF didapatkan dari:

$$IDF(t) = \log \frac{D}{df(t)} \quad (9)$$

Di mana D merupakan total dokumen dibagi df atau banyaknya dokumen yang mengandung kata yang dicari.

2.6. Validasi

Pada tahap ini adalah tahapan untuk melakukan evaluasi menggunakan tabel *confusion matrix* untuk mengetahui apakah model yang telah diterapkan menghasilkan sesuai dengan yang diharapkan. Validasi yang dilakukan menggunakan *10 fold validation*, yaitu membagi data secara acak ke dalam 10 bagian kemudian proses pengujian dimulai dengan model yang sudah dibangun dengan data pada bagian pertama, dan model yang terbentuk akan diuji pada 9 bagian dari sisa data. Hasil dari tahap ini adalah *precision*, *recall*, dan nilai *accuracy*. Menurut Bramer, 2007 dalam (Ernawati, 2016). Klasifikasi yang benar diklasifikasikan seperti pada Tabel 1.



Tabel 1. Model Confusion Matrix.

Correct Classification	Classified As	
	+	-
+	TRUE POSITIVE	FALSE NEGATIVE
-	FALSE POSITIVE	TRUE NEGATIVE

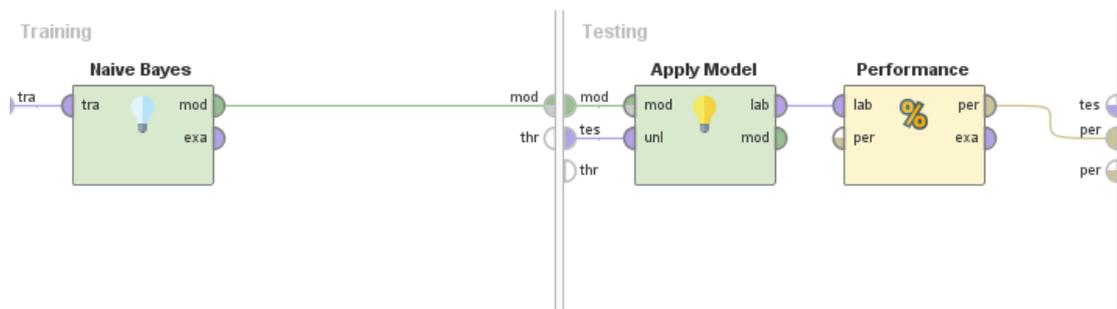
Gorunescu dikutip dalam (Fitriyani & Wahono, 2015) untuk menghitung akurasi digunakan persamaan di bawah ini:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (7)$$

TP merupakan jumlah *true positive* yang berarti dataset positif yang diklasifikasikan ke klasifikasi positif, TN adalah jumlah *true negative* atau jumlah dataset negatif yang diklasifikasikan negatif, FP adalah jumlah *false positive* atau jumlah dataset negatif yang diklasifikasikan positif dan FN adalah jumlah *false negative* atau jumlah dataset positif yang diklasifikasikan negatif.

2.6.1. Cross Validation

Operator proses pada tahap ini dibagi menjadi dua bagian yaitu *Training* dan *Testing*. Pada bagian *training* diterapkan algoritma yang dipakai yaitu *Naive Bayes Classifier* sedangkan pada bagian *esting* terdiri dari operator *Apply Model* untuk menerapkan model pada *dataset* dan *Performance* yang bertugas untuk melihat performa dari model yang diterapkan seperti pada Gambar 2.

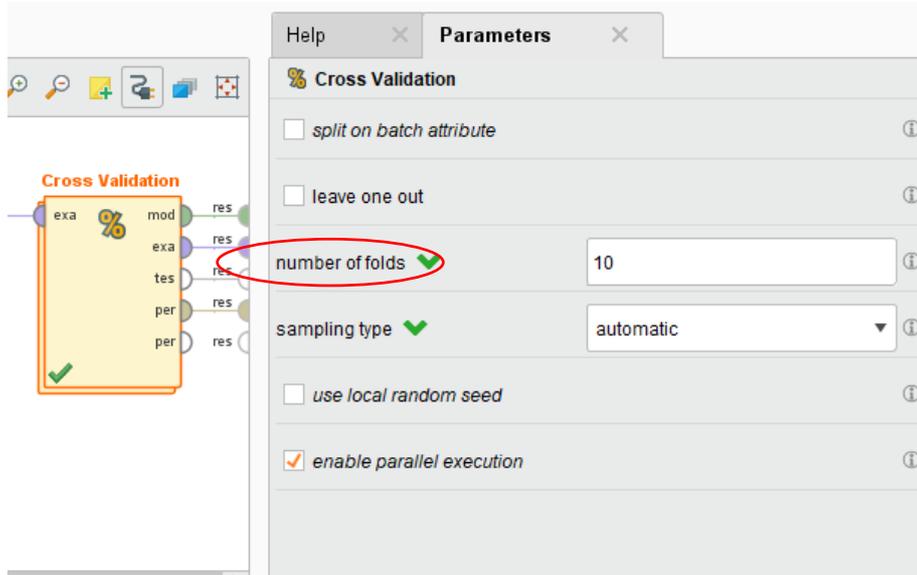


Gambar 2. Tahap Validasi.

2.6.2. K-Fold Validation

Validasi yang dilakukan menggunakan *10 fold validation*, yaitu membagi data secara acak kedalam 10 bagian kemudian proses pengujian dimulai dengan model yang sudah dibangun dengan data pada bagian pertama, dan model yang terbentuk akan diuji pada 9 bagian dari sisa data.





Gambar 3. K-Fold Validation.

3. HASIL DAN PEMBAHASAN

Naïve Bayes Classifier mempunyai kelebihan, di antaranya *Algoritma Naïve Bayes Classifier* dipakai karena *Algoritma Naïve Bayes Classifier* merupakan metode klasifikasi yang efisien dan sederhana (Nugroho et al., 2020). Selain itu, algoritma *Naïve Bayes Classifier* adalah algoritma dengan performa yang sangat baik dalam beberapa kasus klasifikasi teks (Taufik, 2017). Akan tetapi, *Naïve Bayes Classifier* juga memiliki kelemahan di mana sebuah probabilitas tidak bisa mengukur seberapa besar tingkat keakuratan sebuah prediksi. Selain itu, *Naïve Bayes Classifier* juga memiliki kelemahan pada seleksi atribut sehingga dapat mempengaruhi nilai akurasi. (Muhamad et al., 2017). Pada penelitian ini *Algoritma Naïve Bayes Classifier* diterapkan untuk mengetahui akurasi dari analisis sentimen Halodoc dengan data yang digunakan sebanyak 950 data, dengan pengujian data acak 250 data dan 650 data. Dibagi menjadi dua kelas yaitu kelas positif dan kelas negatif. Sebelumnya data tersebut melalui beberapa tahapan proses, berikut penjelasan mengenai proses penelitian yang telah dilakukan, pada Tabel 1 merupakan contoh data yang telah dihasilkan.

Tabel 2. Contoh Data Collect.

No.	Ulasan	Kategori
1.	Aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan, bisa langsung tanya ke dokter.. Harus DI INSTAL Halodoc ini !! Terimakasih Halodoc 😊❤️	Positif
2.	KECEWA. RESPON YANG SANGAT LAMA DAN DOKTER YANG SANGAT JUDES 😞😡😡😡😡	Negatif
3.	Aplikasi yg tdk di rekomendasikan sekali.	Negatif

3.1. Import Data

Import data adalah tahap awal untuk memproses data pada proses-proses yang terjadi pada *Rapidminer*. Data yang akan dimasukkan menggunakan *operator read excel* untuk mengimpor data yang akan dipakai, kemudian diproses pada tahap selanjutnya.



3.2. Preprocessing

Preprocessing adalah tahapan lanjutan setelah pengumpulan data, tahapan tersebut melibatkan beberapa proses, sebagai berikut:

3.2.1. Tokenize

Tahap ini adalah tahap di mana suatu kalimat dipecah menjadi kata-kata yang terpisah, biasanya yang menjadi acuan pemisah antar token adalah spasi atau tanda baca. Contohnya ada pada tabel 2 ini.

Tabel 3. Tokenisasi.

No.	Ulasan	Kategori
1.	Aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan bisa langsung tanya ke dokter Harus DI INSTAL Halodoc ini Terimakasih Halodoc	Positif
2.	KECEWA RESPON YANG SANGAT LAMA DAN DOKTER YANG SANGAT JUDES	Negatif
3.	Aplikasi yg tdk di rekomendasikan sekali	Negatif

3.2.2. Transform Case

Pada tahap ini, semua kata-kata yang didapat kemudian diubah susunan hurufnya menjadi *lower case* atau non kapital, Berikut contohnya seperti pada tabel 3.

Tabel 4. Transform Case.

No.	Ulasan	Kategori
1.	aplikasi yg sangat membantu untuk kita yg awam tentang kesehatan bisa langsung tanya ke dokter harus di instal halodoc ini terimakasih halodoc	Positif
2.	kecewa respon yang sangat lama dan dokter yang sangat judes	Negatif
3.	aplikasi yg tdk di rekomendasikan sekali	Negatif

3.2.3. Filter Token (by length)

Proses ini akan memfilter token berdasarkan batas minimum dan maksimum yang telah ditentukan yaitu batas minimum yang dipakai adalah 4 karakter dan 25 karakter untuk batas maksimum karakter. Seperti terlihat pada tabel 4 Berikut ini.

Tabel 5. Filter Token (by length).

No.	Ulasan	Kategori
1.	aplikasi sangat membantu untuk kita awam tentang kesehatan bisa langsung tanya dokter harus instal halodoc terimakasih halodoc	Positif
2.	kecewa respon yang sangat lama dokter yang sangat judes	Negatif
3.	aplikasi rekomendasikan sekali	Negatif

3.2.4. Filter Stopword

Proses ini bertujuan untuk menghapus kata-kata yang memiliki nilai informasi yang rendah atau yang tidak memiliki relevansi dengan isi dari dokumen, Seperti terlihat pada tabel 6.

Tabel 6. Hasil Preprocessing.

No.	Ulasan	Kategori
1.	aplikasi membantu awam tentang kesehatan, bisa langsung tanya dokter harus instal halodoc terimakasih halodoc	Positif
2.	kecewa respon lama dokter judes	Negatif
3.	aplikasi rekomendasikan sekali	Negatif



3.3. Term weigthing TF-IDF

Setelah semua tahapan *preprocessing* selesai, maka tahap selanjutnya adalah *count vector* atau perhitungan kemunculan kata dari suatu kalimat di dalam dokumen dengan menggunakan pembobotan TF-IDF. Tahap *term weighting* adalah tahap menghitung *term frequencies*. Seberapa sering kata tersebut muncul dalam dokumen. Berikut hasil distribusi TF-IDF Seperti terlihat pada tabel 7 dan tabel 8.

Tabel 7. Hasil Distribusi Term Frequency.

No.	Term	D1	D2	D3	DF	IDF
1	Aplikasi	1	0	1	2	0.176
2	Awam	1	0	0	1	0.477
3	Bisa	1	0	0	1	0.477
4	dokter	1	1	0	2	0.176
5	halodoc	2	0	0	2	0.176
6	harus	1	0	0	1	0.477
7	instal	1	0	0	1	0.477
8	judes	0	1	0	1	0.477
9	langsung	1	0	0	1	0.477
10	rekomendasikan	0	0	1	1	0.477
11	respon	0	1	0	1	0.477
12	sekali	0	0	1	1	0.477
13	tanya	1	0	0	1	0.477
14	tentang	1	0	0	1	0.477
15	terimakasih	1	0	0	1	0.477

Tabel 8. Tabel TF-IDF.

No.	Term	TF*IDF		
		D1	D2	D3
1	aplikasi	0.176	0	0.176
2	awam	0.477	0	0
3	bisa	0.477	0	0
4	dokter	0.176	0.176	0
5	halodoc	0.352	0	0
6	harus	0.477	0	0
7	instal	0.477	0	0
8	judes	0	0.477	0
9	langsung	0.477	0	0
10	rekomendasikan	0	0	0.477
11	respon	0	0.477	0
12	sekali	0	0	0.477
13	tanya	0.477	0	0
14	tentang	0.477	0	0
15	terimakasih	0.477	0	0

3.4. Validasi Hasil

Setelah *preprocessing* dilakukan, selanjutnya menerapkan algoritma pada operator *cross validation* untuk validasi dan untuk mengetahui hasil akurasi dari pemodelan. Pada tahap ini dilakukan perbandingan dengan pengujian data acak. Berikut hasil yang didapat dari pemodelan terlihat pada Tabel 9 dan Tabel 10.

Tabel 9. Perbandingan Hasil Akurasi.

Hasil	250 data	650 data	950 data
Akurasi	81.60%	81.65%	81.68%
AUC	0.688	0.746	0.756



Data acak yang digunakan untuk uji dan perbandingan yaitu 250 data dan 650 data, 250 data dengan proporsi 125 sentimen positif dan 125 sentimen negatif. 650 data dengan proporsi 325 sentimen positif dan 325 sentimen negatif.

Tabel 10. Confusion Matrix.

	<i>True Positif</i>	<i>True Negatif</i>	<i>Class Precision</i>
<i>Pred.Positif</i>	387	86	81.82%
<i>Pred.Negatif</i>	88	389	81.55%
<i>Class Recall</i>	81.47%	81.89%	
<i>Accuracy</i>	81.68% +/- 2.54%		
<i>Micro Average</i>	81.68%		

Dengan rincian perhitungan akurasi sebagai berikut:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

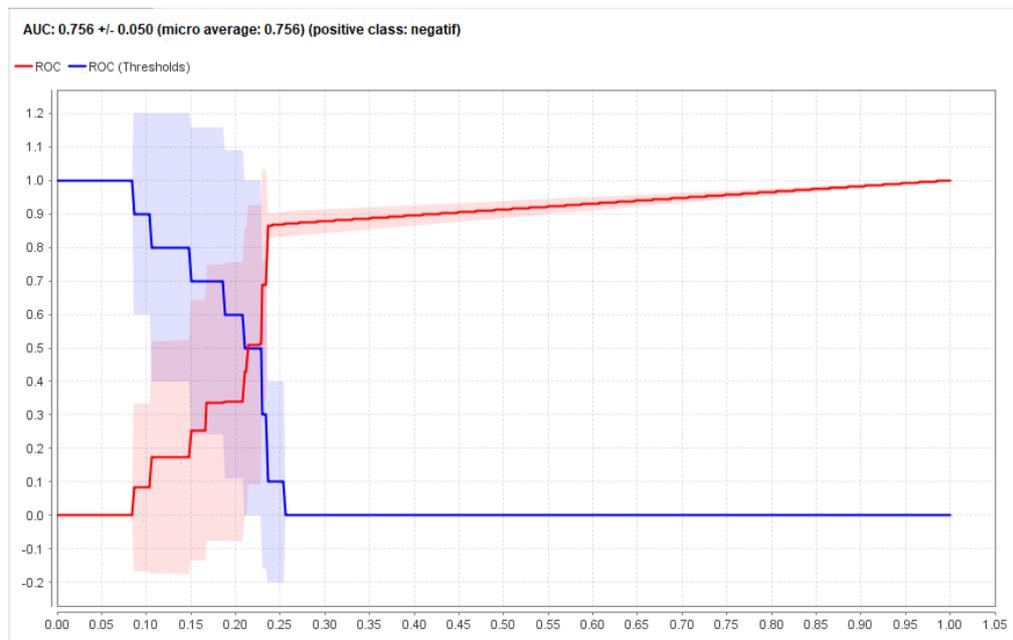
$$accuracy = \frac{387 + 389}{387 + 389 + 88 + 86} \times 100\%$$

$$accuracy = 81.68\%$$

Dari perbandingan hasil akurasi dan AUC yang terlihat pada tabel 10 untuk *Naïve Bayes Classifier*, menunjukkan bahwa pada penelitian ini dengan 950 data memperoleh hasil tertinggi dengan tingkat akurasi sebesar 81.68% dan AUC 0.756.

3.4.1. Kurva ROC

Kurva ROC adalah salah satu cara untuk mengevaluasi hasil akurasi dan klasifikasi dengan visual (Mubarak et al., 2019). Seperti terlihat pada gambar 6, hasil tertinggi menunjukkan nilai AUC sebesar 0.756, dengan artian diagnosa hasil klasifikasi termasuk kedalam *fair classification*.



(sumber: Data Olahan, 2020)

Gambar 4. Kurva ROC.



4. KESIMPULAN

Dari hasil evaluasi dan validasi dapat ditarik kesimpulan bahwa:

- 1) *Naïve Bayes Classifier* terbukti dapat digunakan untuk pengklasifikasian sentimen *review* Halodoc, hasil tertinggi diperoleh dengan menggunakan jumlah data 950 sentimen. *Naïve Bayes Classifier* menghasilkan AUC 0.756 dalam artian bahwa pengklasifikasian pada penelitian ini termasuk ke dalam kategori *fair classification*.
- 2) Algoritma *Naïve Bayes Classifier* yang telah diterapkan untuk analisis sentimen Halodoc dari 475 sentimen positif terdapat 387 sentimen yang tepat di kategorikan sebagai sentimen positif dan sisanya 88 masuk ke dalam kategori sentimen negatif, kemudian dari 475 sentimen negatif 389 sentimen yang dikategorikan tepat sebagai sentimen negatif dan sisanya yaitu 86 sentimen yang dikategorikan sebagai sentimen positif, dengan nilai akurasi 81.68%.
- 3) *Preprocessing* diperlukan untuk mengurangi *term* yang dinilai tidak diperlukan pada saat *count vector*.

Dikarenakan penelitian ini terbatas pada algoritma yang dipakai yaitu *Naïve Bayes Classifier*, dimungkinkan akurasi untuk klasifikasi sentimen *review* Halodoc masih dapat ditingkatkan.

- 1) Beberapa saran untuk penelitian selanjutnya sebagai berikut:
- 2) Peningkatan akurasi dengan seleksi fitur untuk optimasi, misalnya *prune methode*, *genetic algorithm*, dan lain-lain.
- 3) Kemudian pada penelitian selanjutnya dapat digunakan algoritma lain selain *Naïve Bayes Classifier* untuk klasifikasi teks.
- 4) Penelitian *text mining* lain selain analisis sentimen dengan data yang lebih banyak atau sedikit, dan pengimplementasian analisis sentimen ke dalam program atau aplikasi.

DAFTAR PUSTAKA

- APJII. (2019). *Penetrasi & Profil Perilaku Pengguna Internet Indonesia Tahun 2018*.
- Durairaj, M., & Ramasamy, N. (2016). A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. *International Journal of Control Theory and Applications*, 9(27), 255–260.
- Ernawati, S. (2016). Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen Review Perusahaan Penjualan Online Menggunakan Naïve Bayes. *Evolusi: Jurnal Sains dan Manajemen*, 4(1), 45–54. <https://doi.org/10.31294/evolusi.v4i1.605>
- Fitriyani, F. (2018). Metode Bagging Untuk Imbalance Class Pada Bedah Toraks Menggunakan Naive Bayes. *Jurnal Kajian Ilmiah*, 18(3), 278. <https://doi.org/10.31599/jki.v18i3.281>
- Fitriyani, F., & Wahono, R. S. (2015). Integrasi Bagging dan Greedy Forward Selection pada Prediksi Cacat Software dengan Menggunakan Naïve Bayes. *Journal of Software Engineering*, 1(2), 101–108.
- Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman and Hall/CRC.
- Junianto, E., & Riana, D. (2017). Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC. *Jurnal Informatika*, 4(1), 38–45. <https://doi.org/10.31294/ji.v4i1.1810>
- Maarif, A. A. (2015). *Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah*. Universitas Dian Nuswantoro Semarang.
- Mubarok, A., Susanti, S., & Handayani, R. N. (2019). *Optimasi Algoritma Support Vector Machine Menggunakan Particle Swarm Optimization Untuk Analisis Sentimen pada Ulasan Produk Tokopedia*. Universitas Adhirajasa Reswara Sanjaya.
- Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 4(3), 180. <https://doi.org/10.25126/jtiik.201743251>
- Nugroho, K. S., Istiadi, I., & Marisa, F. (2020). Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization. *Jurnal Teknologi dan Sistem Komputer*, 8(1), 21–26. <https://doi.org/10.14710/jtsiskom.8.1.2020.21-26>



- Prasetyo, E. (2012). *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. ANDI.
- Samodra, J., Sumpeno, S., & Hariadi, M. (2019). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes. *SEMINAR NASIONAL ELECTRICAL, INFORMATICS, AND IT'S EDUCATIONS 2019*, 71–74.
- Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 10(2), 681–686. <https://doi.org/10.24176/simet.v10i2.3487>
- Taufik, A. (2017). Optimasi Particle Swarm Optimization Sebagai Seleksi Fitur Pada Analisis Sentimen Review Hotel Berbahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Komputer*, 3(2), 40–47. <https://doi.org/10.31294/jtk.v3i2.1922>
- Tjandrawinata, R. R. (2016). *Industri 4.0: revolusi industri abad ini dan pengaruhnya pada bidang kesehatan dan bioteknologi*. <https://doi.org/10.5281/zenodo.49404>
- Ulya, F. N. (2019, Agustus 19). *Survei: 84,4 Persen Masyarakat Puas dengan Layanan Kesehatan Digital*. Kompas.com. <https://money.kompas.com/read/2019/08/19/134000926/survei--84-4-persen-masyarakat-puas-dengan-layanan-kesehatan-digital?page=all>

