

# ANALISIS SENTIMEN PADA *REVIEW* APLIKASI KESEHATAN HALODOC MENGGUNAKAN METODE *MAXIMUM ENTROPY*

Oleh: Tri Budi Setiyana  
Univeristas Muhammadiyah Semarang

---

Article history	Abstract
Submission :	Advances in technology encourage various developments in all aspects of people's lives. The existence of the internet is the main driving factor for the development of digital information technology which has led to the emergence of new media such as online health portals. Halodoc is one of the leading online health portals in Indonesia. In order to maintain the health services presented are very important. The public assement can be seen from the Google Play website in the review column. Sentiment analysis can be used to analyze these reviews by classifying positive and negative sentiments. Data from Halodoc user reviews during 2020 was then labeled and analyzed using the Maximum Entropy algorithm method. The test results obtained an accuracy value of 96,39%, precission on positive sentiment of 97% and negative 94% and recall on positive sentiment of 97% and negative 94%. Furthermore, from the text association, information is obtained related to topics/words that are often discussed by users of the Halodoc application, namely doctors, applications and good, for positive sentiment classes including positive, doctor, and fast. Meanwhile, the negative sentiment class that often complained about were negative, doctor, bad, and slow.
Revised :	
Accepted :	
<b>Keyword:</b> Aplikasi Halodoc Analisis Sentimen, <i>Maximum Entropy</i> Asosiasi Teks	

---

## Pendahuluan

Kualitas layanan merupakan sesuatu yang wajib dan harus selalu dikedepankan oleh setiap para pelaku bisnis, pelayanan yang diberikan dengan kualitas yang baik akan membuat konsumen memberikan respon positif terhadap perusahaan, baik dalam pelayanan jasa maupun produk. Kualitas layanan merupakan salah satu bentuk produk perusahaan berupa tindakan atau perbuatan yang dapat ditawarkan oleh suatu pihak kepada pihak lain atau pelanggan yang pada dasarnya bersifat *intangible* (tidak terwujud fisik) namun hasilnya bisa dirasakan oleh pelanggan.

Seiring dengan berkembangnya zaman, perkembangan teknologi informasi dan komunikasi berkembang sangat pesat. Terobosan-terobosan teknologi terbaru dibuat

untuk mempermudah dalam kehidupan sehari-hari, baik dalam menyelesaikan pekerjaan maupun memperoleh informasi. Keberadaan internet menjadi salah satu faktor pendukung utama perkembangan teknologi informasi dan komunikasi yang menyebabkan dunia menjadi tanpa batas dan menyebabkan perubahan social secara signifikan berlangsung secara cepat. New media merupakan dari perkembangan teknologi informasi dan komunikasi yang semakin maju yang bersifat baru dan digital, dengan fitur andalannya, membuat informasi dan komunikasi semakin cepat berpindah dari satu sumber kepada khalayak luas, hal ini pula yang semakin di manfaatkan oleh perusahaan untuk mengembangkan bisnisnya kearah digital atau *online*.

Penetrasi internet Indonesia yang menunjukkan peningkatan disetiap tahunnya. Berdasarkan data survei yang diselenggarakan oleh APJII (Asosiasi Penyelenggara Jasa

Internet Indonesia) pada tahun 2018, pengguna internet di Indonesia sebanyak 171,17 juta jiwa dari total populasi 246,16 juta jiwa. Sedangkan untuk perangkat yang digunakan oleh pengguna internet melalui *mobile* yaitu 93,9% mengakses internet setiap hari, sedangkan yang menggunakan computer adalah 9% setiap hari. Dengan data yang menunjukkan angka pengguna yang besar pada pemakaian internet lewat *mobile* atau *smartphone* menjadi kesempatan perusahaan dalam memanfaatkan kesempatan ini untuk menggunakan strategi *mobile* untuk mengembangkan usahanya dengan menciptakan aplikasi *mobile*.

Seiring kemajuan teknologi digital yang semakin canggih. Kolaborasi antara teknologi digital dan kesehatan dibuktikan dengan terciptanya berbagai aplikasi kesehatan melalui gawai (*gadget*) untuk menjaga dan memantau kesehatan dengan mudah dan praktis. Apalagi masyarakat tidak bisa lepas dari *gadget* dan kesehatan adalah suatu hal penting yang tidak dapat lepas dari kehidupan manusia. Pada tahun 2016, MHealth Indonesia membuat terobosan baru dengan memanfaatkan kecanggihan teknologi tersebut dengan menghadirkan Halodoc konsultasi dokter sebagai aplikasi kesehatan yang membantu dalam melayani kesehatan masyarakat luas dengan lebih mudah. Dengan meluncurkan aplikasi Halodoc, pengguna/pelanggan dapat dengan mudah dan cepat dalam berselancar di dunia virtual untuk mencari informasi kesehatan dan berkomunikasi dengan dokter.

Pandemi virus corona Covid-19 membuat masyarakat di dunia harus terbiasa dengan tatanan kehidupan yang baru atau biasa disebut New Normal. Salah satu hal yang mulai dilakukan selama pandemic adalah pelayanan telemedicine yang berfungsi untuk mengurangi layanan kesehatan secara tatap muka demi menekan penyebaran covid-19. Layanan telemedicine dapat menghemat waktu dan biaya serta membantu pasien memutuskan apakah cukup hanya di rumah saja atau perlu mengunjungi dokter. Dengan menggunakan telemedicine, pasien dapat berkonsultasi secara *online* dengan spesialis terqualifikasi untuk membicarakan pilihan layanan kesehatan, dimana pasien akan mendapatkan informasi kesehatan dengan efisien dan cepat.

Aplikasi Halodoc yang terdapat disitus *Google Play* masih tercatat sebagai salah satu top posisi sebagai aplikasi kesehatan. *Google Play* sendiri merupakan layanan distribusi digital

yang terpraktasai oleh *Google* yang berfungsi sebagai toko aplikasi resmi dari sistem *android*. Salah satu fitur dari *Google Play* adalah keberadaan kolom komentar bagi pengguna untuk memberikan penilaian keberadaan kolom komentar bagi pengguna untuk memberikan penilaian kepada aplikasi berupa *rating* atau *review*. Hal ini dimanfaatkan untuk melihat penilaian public berdasarkan ulasan positif seperti apresiasi dan saran maupun ulasan negatif seperti keluhan, untuk mencerminkan mengenai bagaimana kinerja Halodoc dimata para pengguna.

## Landasan Teori

### Data Mining

Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies dkk, 2004). *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data. Menurut Macleman dkk (2009), fungsi data mining dibagi menjadi sebagai berikut:

#### 1. Classification

Classification untuk mencari model atau fungsi yang menggambarkan dan membedakan kelas-kelas atau konsep data untuk mengklasifikasikan target class ke dalam kategori yang dipilih disebut classification. Dalam klasifikasi, terdapat target variabel kategori, misalnya pendapatan tinggi, sedang, dan rendah.

#### 2. Clustering

Clustering berguna untuk mencari pengelompokan atribut ke dalam segmentasi-segmentasi berdasarkan similaritas. Cluster berbeda dengan classification karena tidak adanya variabel target dalam cluster. Algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok yang memiliki kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai maksimal.

#### 3. Association

Association berguna untuk mencari keterkaitan antara atribut atau item set, berdasarkan jumlah item yang muncul dari association rule yang ada. Tugas asosiasi dalam data mining adalah

menemukan atribut yang muncul dalam satu waktu.

#### 4. Regression

Regression berguna untuk mencari prediksi dari suatu pola yang ada, fungsinya hamper menyerupai dengan classification.

#### 5. Forecasting

Forecasting berguna untuk peramalan waktu yang akan datang berdasarkan trend yang telah terjadi di waktu sebelumnya.

#### 6. Sequence Analysis

Sequence analysis berguna untuk mencari pola urutan dari ringkasan kejadian.

#### 7. Deviation Analysis

Deviation analysis berguna untuk mencari kejadian abnormal yang sangat berbeda dari keadaan umumnya.

### **Text Mining**

Menurut Feldman & Sanger (2007), text mining merupakan proses penggalan informasi secara insentif yang bekerja menggunakan alat dan metode tertentu untuk menganalisis suatu kumpulan dokumen. Text mining digunakan untuk mendeskripsikan sebuah kronologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, sedangkan data mining digunakan dalam pengolahan data yang bersifat terstruktur.

Adapun yang paling membedakan antara text mining dan data mining berada pada sumber daya yang digunakan. Kesamaan antara keduanya yakni menggunakan data besar dan data berdimensi tinggi dengan struktur yang terus berubah. Pada text mining, pola-pola yang diekstrak dari data tekstual yang tidak terstruktur. Sedangkan pada data mining, data yang diolah umumnya sudah terstruktur dari proses warehousing. Sehingga text mining biasanya lebih sulit dari data mining karena berkaitan langsung dengan masyarakat dimana memiliki struktur teks yang kompleks, struktur yang tidak lengkap, Bahasa yang berbeda, dan arti yang tidak standar. Maka dari itu digunakan Natural Language Processing untuk analisis teks yang tidak berstruktur tersebut. Secara umum tahap-tahap pada text mining dapat dibagi atas text preprocessing dan feature selection.

### **Pembobotan Kata (*Term Weigthing*)**

Pembobotan kata (*term weighting*) merupakan salah satu tahapan yang perlu diperhatikan dalam mencari informasi dari koneksi dokumen yang heterogen. Dalam dokumen umumnya terdapat kata. Frase, atau unit indeks lainnya yang menunjukkan konteks dari dokumen tersebut, hal inilah yang disebut *term*. *Term weigthing* digunakan untuk memberikan indikator dan setiap kata sesuai dengan tingkat kepentingan masing-masing kata dalam dokumen. Salah satu metode pembobot *term* terbaru yang paling banyak digunakan adalah metode *Term Frequency – Inverse Document Frequency* (TF-IDF). Dalam perhitungan TF-IDF, perhitungan bobot *term* dari sebuah doukmen dilakukan dengan menghitung masing-maing niali *Term Frequency* dan *Inverse Document Frequency*.

Menurut Zafikri (2008), perhitungan nilai TF-IDF dapat dilakukan dengan menggunakan rumus berikut:

#### *1. Term Frequency (TF)*

*Term Frequency* merupakan faktor yang menentukan perhitungan bobot *term* berdasarkan jumlah dan bentuk kemunculan kata pada dokumen. Pada dasarnya dapat dikatakan bahwa semakin besar nilai jumlah kemunculan suatu term, maka semakin besar juga niali bobot term tersebut dalam dokumen. Adapun perhitungan nilai *term frequency* dapat dilakukan dengan beberapa cara sebagai berikut:

- 1) TF biner, pemberian bobot *term* dilihat berdasarkan ada tidaknya suatu kata dalam dokumen. Jika terdapat kata tersebut maka diberi nilai satu, jika tidak diberi nilai nol.
- 2) TF murni atau *raw TF*, pemberian bobot *term* dilihat berdasarkan jumlah kemunculan suatu kata dalam dokumen. Missal jika kata tersebut muncul tiga kali maka akan diberi bobot tiga.
- 3) TF logaritmit, pemberian bobot *term* pada dokumen yang memiliki sedikit

kata dalam *query*, namun mempunyai frekuensi yang tinggi.

$$tf = 1 + \log (tf)$$

- 4) TF normalisasi, pemberian bobot *term* diperoleh dengan membandingkan frekuensi sebuah kata dengan jumlah seluruh kata dalam dokumen

$$tf = 0,5 + 0,05x \left( \frac{tf}{\max tf} \right)$$

## 2. Inverse Document Frequency (IDF)

*Inverse Document Frequency* merupakan proses mengurangi dominasi *common term* yang sering muncul dalam dokumen, *common term* perlu dihilangkan umumnya kurang bernilai sehingga sering menyebabkan analisis kurang maksimal. Selain itu, IDF juga bertujuan untuk menjaga faktor dengan menghitung nilai faktor kebalikan dari frekuensi dokumen yang mempunyai suatu kata. Adapun perhitungan nilai *inverse Document Frequency* dapat dilakukan dengan  $idf_j = \log \left( \frac{D}{df_j} \right)$

dimana

$D$  : jumlah keseluruhan dokumen

$df_j$  : jumlah dokumen yang mempunyai *term*  $t_j$ .

Adapun nilai TF-IDF diperoleh dari perkalian nilai *Term Frequency* dengan nilai *Inverse Document Frequency*. Maka pada perhitungan TF-IDF untuk *raw TF* digunakan rumus sebagai berikut:

$$W_{ij} = tf_{ij} \times idf_j$$

$$W_{ij} = tf_{ij} \times \log \frac{D}{df_j}$$

Dengan

$W_{ij}$  : bobot *term*  $t_j$  terhadap dokumen  $d_i$

$tf_{ij}$  : jumlah kemunculan *term*  $t_j$  dalam dokumen  $d_i$

## Analisis Sentimen

Menurut Lee dan Pang (2008), analisis sentiment merupakan proses memperoleh informasi dengan cara

memahami, mengekstrak, dan mengolah data tekstual secara otomatis. Analisis sentiment mulai terkenal pada tahun 2013 sebagai salah satu cabang *text mining* dan juga dikenal dengan *opinion mining*. Pada dasarnya, analisis sentimen digunakan untuk mengetahui tanggapan dan sikap dari suatu kelompok atau individu terhadap suatu topik bahasan kontekstual keseluruhan dokumen. Tanggapan dan sikap tersebut dapat berupa pendapat, penilaian atau evaluasi (teori appraisal), keadaan efektif (keadaan emosional penulis saat menulis) atau komunikasi emosional (efek emosional yang sampai pada pembaca).

Secara umum domain yang sering membutuhkan analisis sentiment antara lain produk konsumen, layanan dan jasa, peristiwa social dan politik yang memerlukan opini publik. Analisis sentiment lebih memiliki kecenderungan terhadap penelitian mengenai pernyataan suatu pendapat yang mempunyai suatu sentiment baik positif atau negatif. Pendapat memiliki pengaruh yang tinggi kepada perilaku seseorang, maka dapat dikatakan semua aktivitas seseorang terwakili dari pendapat tersebut. Hal ini terlihat dalam proses pengambilan keputusan dimana umumnya diambil dari pendapat orang-orang. Pada sektor bisnis dan organisasi, pendapat dan opini publik menjadi sangat penting terhadap penilaian suatu produk dan jasa.

Bagi sektor bisnis, analisis sentiment dapat berguna dalam proses pelacakan produk, jasa, merek, dan target konsumen di pasar. Selain itu analisis sentimen juga dapat menilai keunggulan dan kelemahan suatu produk dan jasa. Secara umum analisis sentimen digunakan untuk mendeteksi keluhan, persepsi produk atau layanan baru, dan persepsi dari suatu merek tertentu.

## Klasifikasi

Menurut Prasetyo (2012), klasifikasi adalah proses pengelompokan teramati dari suatu objek data ke dalam suatu kelas

tertentu berdasarkan kelas-kelas yang ada. Teknik klasifikasi lebih efektif digunakan dalam proses prediksi dan penggambaran suatu kumpulan data untuk jenis kategori biner atau nominal dibandingkan dengan kategori ordinal.

Menurut Ham dan Kamber (2006), data klasifikasi strategi menjadi dua proses tahapan. Tahap pertama merupakan *learned model* dimana dilakukan pembangunan model hasil analisis *record database* dari serangkaian kelas data yang ada. Masing-masing *record* diasumsikan mempunyai *predefined class* yang didasarkan pada atribut kelas label, karena masing-masing *record* memiliki kelas label maka klasifikasi termasuk ke dalam *supervised learning*. Hal inilah yang membedakan antara klasifikasi dan *clustering learning* yang tidak memerlukan kelas label (*unsupervised learning*). Tahap ini juga sering disebut dengan tahap pembelajaran atau pelatihan. Pelatihan dilakukan dengan menganalisis data latih hingga diperoleh informasi yang dibutuhkan untuk membangun suatu model algoritma klasifikasi. Proses pembangunan tersebut dapat dilihat sebagai proses pembentukan dan penataan fungsi  $y = f(x)$  dengan  $y$  yaitu kelas label hasil prediksi dan  $x$  yaitu *record* yang akan diprediksikan.

Menurut Ham dan Kamber (2006), untuk memperoleh hasil klasifikasi yang baik diperlukan beberapa persiapan sebagai berikut:

1. Pembersihan data  
Pembersihan data digunakan untuk mengurangi kecacatan data terutama dalam proses pembangunan model. Pembersihan yang bisa dilakukan adalah menghilangkan data noise, melengkapi data yang hilang, dan seterusnya.
2. Analisa relevansi  
Dalam proses klasifikasi terdapat atribut-atribut yang memiliki tingkat kemampuan tinggi dan sering kali saling berhubungan kuat satu dengan lainnya. Sehingga atribut-atribut

tersebut perlu dihilangkan agar tidak mempengaruhi tingkat keoptimalan klasifikasi.

### Ukuran Evaluasi Model

Proses evaluasi dilakukan dengan menghitung suatu ukuran tertentu terhadap himpunan data uji, yakni data yang tidak digunakan dalam proses pembuatan model klasifikasi tersebut. *Confusion matrix* merupakan matrix yang berisi informasi mengenai klasifikasi actual yang akan diprediksi oleh sistem klasifikasi (Kohavi & Provost, 1998). Sistem klasifikasi dibentuk dari pemetaan suatu baris data dan *output* suatu hasil prediksi kelas dari data tersebut. Pada suatu klasifikasi baris data dapat menghasilkan empat kemungkinan yang digunakan untuk menilai dan mengevaluasi proses klasifikasi. Apabila data positif dan tepat di prediksi positif maka disebut *true positive*, namun jika salah dan terprediksi negatif maka disebut *false negative*. Apabila data negatif dan tepat diprediksi negatif maka disebut *true negative*, namun jika salah dan terprediksi positif maka disebut *false positive*.

Kelas	Positif	Negatif
Positif	<i>True Positive</i> ( <i>TP</i> )	<i>False Negative</i> ( <i>FN</i> )
Negatif	<i>False Positive</i> ( <i>FP</i> )	<i>True Negative</i> ( <i>TN</i> )

### K-FOLD Cross Validations

*Cross validation* adalah teknik validasi tingkat keakuratan sebuah model dari suatu dataset tertentu. Dataset terbagi

menjadi data latih sebagai data yang digunakan untuk membangun model dan data uji sebagai data yang digunakan untuk memvalidasi model tersebut. Model klasifikasi digunakan dalam proses klasifikasi atau prediksi suatu data baru yang tidak termasuk dalam data pembangunan model.

Menurut Refaeilzadeh dkk (2009), salah satu metode dari *cross validation* yang umumnya digunakan dalam perhitungan akurasi prediksi suatu sistem adalah *K-fold cross validation*. Proses dalam *K-fold cross validation* dilakukan dengan membagi dataset menjadi *K* segmen yang hamper sama ukuran proporsinya. Kemudian salah satu segmen *K* diambil sebagai data uji sedangkan *K-1* segmen lainnya digunakan sebagai data latih dari pembentukan model baru. Proses pelatihan dan penilaian ini dilakukan sebanyak *K* kali iterasi. Nilai *K-fold cross validation* diperoleh dari rata-rata hasil iterasi yang dilakukan. Jumlah *K* yang umumnya digunakan dalam *K-fold cross validation* yakni 5, 7, 10 dan 15

Iterasi	Data latih	Data uji
1	K2,K3,K4,K5	K1
2	K1,K3,K4,K5	K2
3	K1,K2,K4,K5	K3
4	K1,K2,K3,K5	K4
5	K1,K2,K3,K4	K5

### Maximum Entropy

Menurut Nigam (1999), Maximum Entropy merupakan salah satu *machine learning* yang menggunakan proses pengestimasiian probabilitas distribusi dalam pengklasifikasian data. Dalam metode

*Maximum Entropy* dinyatakan bahwa untuk dataset yang tidak diketahui informasi mengenai distribusinya, maka data tersebut akan diasumsikan untuk mengestimasi berbagai *natural language* taks seperti *language modelling* pelabelan *part of speech* dan segmentasi pada teks lainnya.

Menurut Anggreini (2008), *Maximum Entropy* adalah metode klasifikasi berbasis probabilitas yang termasuk dalam kelas model eksponensial. Prinsip dari *Maximum Entropy* didasarkan pada distribusi  $p(a/b)$  yang akan memberikan nilai *entropy* maksimum. *Maximum Entropy* didefinisikan sebagai rata-rata nilai informasi yang maksimum untuk suatu himpunan kejadian *X* dengan distribusi nilai probabilitas yang seragam. Distribusi nilai probabilitas seragam dimaksud adalah distribusi yang menggunakan faktor ketidakpastian yang minimum atau dapat disebut sebagai distribusi yang memakai asumsi seminimal mungkin. Dengan menggunakan asumsi yang minimal, maksimal distribusi yang didapatkan merupakan distribusi yang paling mendekati kenyataan. Pencarian disribusi probabilitas yang paling memberikan nilai *entropy* yang maksimum dilakukan dengan tujuan mencari probabilitas terbaik.

### Definisi Entropy

*Entropy* merupakan rata-rata dari himpunan informasi yang terdapat dalam suatu kumpulan kejadian  $x = (x_1, x_2 \dots x_n)$ . Himpunan informasi yang terdapat pada suatu kejadian dapat dinyatakan dengan

$$h(x) = \log \frac{1}{p(x)}$$

dimana  $h(x)$  adalah himpunan informasi dari suatu kejadian *x* yang dinyatakan dengan ukuran *bit*. Sedangkan  $p(x)$  adalah probabilitas dari kemunculan kejadian merepresntasikan himpunan informasi suatu kejadian. Semakin besar nilai  $h(x)$ , maka semakin besar pula informasi yang dapat dinyatakan sebagai berikut

$$H(p) = -\sum_{x \in \epsilon} p(x) \log p(x)$$

Dengan

$$\epsilon \in A \text{ dan } b \in B$$

$$x = (ab)$$

$$\epsilon = A \times B$$

$p(x)$  yakni peluang kelas a terdapat pada dokumen b.

Adapun hasil yang diharapkan dari metode algoritma *Maximum Entropy* yakni mendapatkan nilai ( $p$ ) yang paling maksimal. Nilai *entropy* yang maksimal akan terpenuhi pada distribusi seragam sehingga mengakibatkan  $(x) = \frac{1}{|x|}$ , dengan nilai  $|X|$  merupakan kardinilitas dari  $X$ . kardinilitas suatu himpunan merupakan nilai ukuran banyaknya elemen yang terdapat dalam suatu himpunan. Sehingga proses untuk mendapatkan nilai maksimal yang seragam dalam klasifikasi tidak semudah membagi nilai 1 dengan nilai kardinilitas  $X$ . Selain itu pencarian distribusi probabilitas tentunya harus memenuhi batasan-batasan sesuai dengan data yang diteliti.

### **Prinsip Maximum Entropy**

Pada metode *Maximum Entropy* dinyatakan bahwa untuk memperoleh nilai *entropy* maksimal maka seluruh distribusi akan diusahakan untuk *inform*, apabila tidak terdapat informasi data yang lengkap. Pada klasifikasi teks dengan *Maximum Entropy* dilakukan pengestimasi distribusi fitur kemunculan kata. Perhitungan fitur didasari oleh penggunaan  $f_i \in \{0,1\}$  dalam pencarian informasi kemunculan suatu fitur dalam suatu dokumen. Sehingga dasarnya metode algoritma *Maximum Entropy* digunakan untuk mencari distribusi probabilitas yang paling seragam.

### **Algoritma Klasifikasi dengan Maximum Entropy**

Adapun proses algoritma klasifikasi teks menggunakan metode *Maximum Entropy* sebagai berikut;

1. Mengidentifikasi kata-kata spesifik yang ada di dalam dokumen (kalimat)

2. Membentuk matriks yang berisi nilai kemunculan kata-kata spesifik tersebut dengan indeks berikut:

$$\begin{cases} 1 : \text{jika } f_j \text{ muncul di dokumen } b \text{ pada kelas } a \\ 0 : \text{jika } f_j \text{ tidak muncul di dokumen } b \text{ pada kelas } a \end{cases}$$

3. Membangun data latih untuk membuat algoritma *Maximum Entropy* dengan menghitung nilai untuk setiap kelas  $a_j \cdot a_j^{(0)} = 1$

$$a_j^{(n+1)} = a_j^{(n)} \left( \frac{E_p f_j}{E^{(n)} f_j} \right)^{\frac{1}{c}}$$

Dimana

$$E_p f_j = \sum_{x \in \epsilon} p(x) f_j(x)$$

$$E^{(n)}(x) = \sum_{x \in \epsilon} p^{(n)}(x) f_j(x)$$

$$\forall x \in \sum_{j=1}^k f_j(x) = c$$

4. Menghitung *joint probability*  $p(a,b)$  untuk perhitungan data uji

$$a = \{\text{positif, negatif}\}$$

$$p^*(a,b) = \pi \prod_{j=1}^k a_j^{f_j(a,b)}$$

5. Menentukan kelas dokumen data uji dengan melihat nilai  $a^*$  tertinggi dari masing-masing kelas.

$$a^* = \text{argmax } p(a,b)$$

dengan  $a \in \{\text{positif, negatif}\}$

### **WORDCLOUD**

*Wordcloud* merupakan salah satu metode untuk menampilkan kata-kata populer yang berkaitan dengan kata kunci internet dari data teks, khususnya pada analisis *text mining*. *Wordcloud* dapat digunakan dalam menyoroti *trend* ataupun istilah populer dikalangan pengguna. *Wordcloud* dibentuk berdasarkan frekuensi kemunculan kata, dimana kata yang paling sering muncul di dalam teks akan memiliki ukuran paling besar, begitu pula sebaliknya. Pendekatan menggunakan *wordcloud* dapat memberikan penjelasan terhadap pertanyaan penelitian dengan sangat cepat dan mudah serta dapat pula dilakukan analisis yang komprehensif.

### **ASOSIASI TEKS**

Asosiasi teks diperoleh dengan melakukan pendekatan pada perhitungan nilai korelasi. Pada umumnya, nilai korelasi digunakan dalam menyatakan hubungan dua atau lebih variabel kuantitatif, namun pada asosiasi teks nilai korelasi dimaknai sebagai keeratan

hubungan antar dua atau lebih variabel kualitatif. Korelasi bertujuan untuk menemukan tingkat hubungan antara variabel (X) dan variabel (Y), dalam ketentuan data memiliki syarat-syarat tertentu.

$$r = \frac{1}{\sqrt{\{n\sum xi^2(\sum xi)^2\}\{n\sum yi^2(\sum yi)^2\}}}$$

dengan

r = nilai korelasi antara variabel x dan variabel y

n = banyaknya pasangan data x dan y

$\sum xi$  = jumlah nilai pada variabel x  $i = 1,2,3, \dots, n$

$\sum yi$  = jumlah nilai variabel y

$\sum xi^2$  = kuadrat dari total nilai variabel x

$\sum yi^2$  = kuadrat dari total nilai variabel y

$\sum xi \sum yi$  = jumlah dari hasil perkalian antara nilai variabel x dan variabel y

## METODE PENELITIAN

### Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data kualitatif. Dimana data yang menjadi pokok penelitian ini berasal dari teks yang terdapat dalam ulasan yang diteliti. Sumber data yang digunakan dalam penelitian ini adalah data sekunder. Data tersebut diperoleh dengan menggunakan teknik *scraping* dengan bantuan aplikasi bawaan *scraper* dari *Google Chrome* pada halaman situs aplikasi kesehatan *online Halodoc*

(<https://play.google.com/store/apps/details?id=com.linkdokter.halodoc.android>),

### Variabel Penelitian

Menurut Sugiyono (2011), variabel penelitian segala sesuatu yang ditentukan dan dipelajari oleh peneliti untuk memperoleh informasi yang digunakan dalam penarikan kesimpulan. Adapun

variabel yang digunakan dalam penelitian ini adalah sebagai berikut:

1. *Date*, yakni tanggal dipublikasikan ulasan.
2. *Rating*, yakni tingkat kepuasan pengguna terhadap aplikasi.
3. Ulasan atau *review*, yakni pendapat pengguna terhadap aplikasi.

## Langkah Penelitian

Langkah-langkah dalam penelitian ini yaitu :

1. Mengambil data ulasan dengan menggunakan *tools scraper* pada *Google Chrome*.
2. Mengumpulkan data ulasan sesuai dengan rate.
3. Praproses teks
  - a. *Translating teks* atau menerjemahkan semua ulasan berbahasa asing (Inggris) kedalam Bahasa Indonesia.
  - b. Melakukan *spelling normalization* atau perbaikan pada kata-kata singkatan yang masih salah ejaan atau kata-kata sinonim.
  - c. Melakukan *case folding* atau mengubah semua teks dengan menggunakan huruf kecil.
  - d. Melakukan *tokenizing* yakni memilih dokumen menjadi kata per kata.
  - e. Melakukan *filtering* atau pemilihan kata dalam dokumen yang kurang informatif.
4. Lalu dilakukan seleksi dengan TF-IDF yaitu merubah menjadi vektor dan pembobotan data.
5. Klasifikasi data *training* dan data *testing* menggunakan *5-Fold Cross Validation* dengan pembagian 80% : 20%
6. Klasifikasi data menggunakan *Maximum Entropy* untuk mengklasifikasikan ulasan yang berbentuk positif dan negatif.
7. Melakukan evaluasi dari hasil klasifikasi.
8. Melakukan visualisasi ulasan dengan *wordcloud*.
9. Interpretasi dan mengambil keputusan.

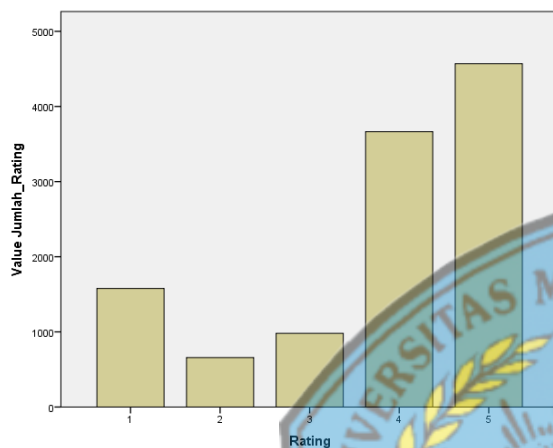
## HASIL PENELITIAN dan PEMBAHASAN

### Statistika Deskriptif

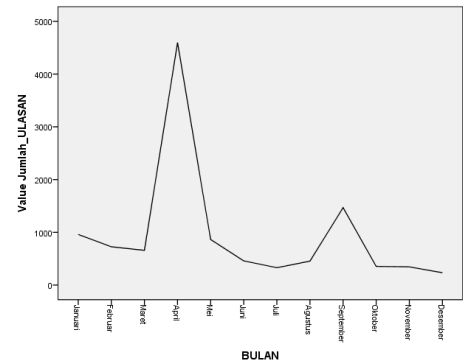
Analisis statistik deskriptif dalam penelitian ini digunakan untuk melihat gambaran



umum mengenai informasi dari aplikasi Halodoc berdasarkan data ulasan pengguna dari situs *Google Play*, yang dilihat dari beberapa aspek diantaranya jumlah ulasan yang masuk berdasarkan urutan waktu, *rating* aplikasi yang diberikan pengguna. Jumlah ulasan dari pengguna dikategorikan menjadi dua kategori yaitu ulasan positif dan ulasan negatif. Berikut ini merupakan *rating* yang menggambarkan penilaian yang diberikan pengguna terhadap aplikasi Halodoc pada situs *Google Play*.



Berdasarkan Gambar dapat dilihat jumlah ulasan dari *rating* pengguna. *Rating* pada situs *Google Play* mempunyai skala 1-5 dengan skor 1 untuk kategori “Sangat Tidak Suka”, skor 2 untuk kategori “Sangat Suka”, skor 3 untuk kategori “Cukup”, skor 4 untuk kategori “Suka”, dan skor 5 untuk kategori “Sangat Suka”. Dari penilaian pengguna diketahui pengguna Halodoc mempunyai penilaian yang baik terhadap aplikasi. Hal ini terlihat dari mayoritas pengguna memberikan *rating* 5 atau sangat suka terhadap aplikasi dengan jumlah 4568 ulasan dan diikuti oleh *rating* 4 atau suka terhadap aplikasi dengan jumlah 3664 ulasan. Adapun pengguna memberikan *rating* 3 atau cukup terhadap aplikasi sebanyak 980 ulasan, *rating* 2 atau tidak suka terhadap aplikasi sebanyak 658 ulasan, dan untuk *rating* 1 atau sangat tidak suka terhadap aplikasi sebanyak 1577 ulasan.



Berdasarkan grafik pada Gambar 4.3 dilihat jumlah ulasan dari para pengguna pada setiap bulannya tahun 2020. Jumlah ulasan terbanyak terdapat pada bulan April dengan 4596 ulasan. Jumlah ulasan terkecil pada bulan juli dengan 329 ulasan dan bulan desember 230 ulasan. Pada bulan April dan September terjadi kenaikan jumlah ulasan yang tinggi dibandingkan bulan-bulan lainnya. Hal ini disebabkan pada bulan April banyak masyarakat yang berkonsultasi tentang virus covid 19, dan pada bulan September jumlah yang terkonfirmasi terkena virus covid 19 tinggi, banyak masyarakat yang melakukan tes kesehatan.

### Preprocessing Data

#### 1. Translating

*Translating* adalah proses penyeragaman Bahasa, karena analisis dilakukan dalam Bahasa Indonesia maka semua ulasan berbahasa asing (Inggris) akan diterjemahkan ke dalam Bahasa Indonesia.

Data Input	Data hasil <i>translating</i>
I need an easy and fast application about health.	Saya membutuhkan aplikasi yang mudah dan cepat tentang kesehatan.

## 2. Spelling Normalization

*Spelling Normalization* adalah proses penyeragaman ejaan kata. Perbaikan dilakukan pada kata-kata singkatan, kata-kata yang masih salah ejaan atau kata-kata sinonim dengan makna serupa.

Data Input	Data hasil <i>Spelling Normalization</i>
Pelayanannya <u>sgt</u> baik tanpa <u>hrus</u> menunggu lama saya suka sekali dan chat respon dokter <u>sgt</u> baik <u>skli</u> .	Pelayanan sangat baik tanpa harus menunggu lama saya suka sekali dan chat respon dokter sangat baik sekali.

## 3. Case Folding

*Case folding* adalah proses penyeragaman bentuk huruf. Karakter lain selain huruf seperti angka, tanda baca, dan emoticon akan dianggap sebagai *delimiter* sehingga karakter tersebut akan dihapus.

Data Input	Data hasil <i>case folding</i>
Pelayanan sangat baik tanpa harus menunggu lama saya suka sekali dan chat respon	pelayanan sangat baik tanpa harus menunggu lama saya suka sekali dan chat respon

dokter sangat baik sekali.	dokter sangat baik sekali.
-------------------------------	-------------------------------

## 4. Tokenizing

*Tokenizing* adalah proses memisahkan kata per kata pada sebuah dokumen menjadi kata – kata yang paling independen.

Data Input	Data hasil <i>tokenizing</i>
pelayanan sangat baik tanpa harus menunggu lama saya suka sekali dan chat respon dokter sangat baik sekali.	pelayanan sangat baik tanpa harus menunggu lama saya suka sekali dan chat respon dokter sangat baik sekali.

## 5. Filtering

*Filtering* adalah proses penyaringan atau pemilihan kata dalam dokumen. tahap ini bertujuan untuk mengurangi dimensi kata di dalam *corpus* dengan menggunakan *stopwords*. *Stopwords* terdiri dari kata-kata yang kurang informatif atau kurang berpengaruh secara keseluruhan yang sering kali muncul dalam dokumen.

Data Input	Data hasil <i>tokenizing</i>
pelayanan <b>sangat</b> baik <b>tanpa harus</b>	pelayanan baik menunggu

menunggu	lama	suka	sekali
saya	suka	chat	
sekali	dan	respon	dokter
chat	respon	baik	
dokter	sangat		
baik sekali..			

Negatif	-1	Waktu pandemic takut rumah sakit
---------	----	--

### PELABELAN KELAS SENTIMEN

Sentimen	Jumlah Ulasan
Positif	2950
Negatif	1213

Berdasarkan Tabel hasil pelabelan kelas sentiment menunjukkan bahwa jumlah ulasan positif memiliki frekuensi yang lebih tinggi dibandingkan dengan jumlah ulasan negatif. Jumlah ulasan positif sebanyak 2950 ulasan, dan ulasan negatif sebanyak 1213 ulasan.

### PEMBUATAN DATA LATIH DAN UJI

Klasifikasi	Jumlah	Data Latih (80%)	Data Uji (20%)
Positif	2950	2360	590
Negatif	1213	971	242
Total	4163	3331	834

Dengan perbandingan data latih dan data uji sebesar 80%:20% dari total 4163 ulasan berbahasa Indonesia. Digunakan sebanyak 3331 ulasan sebagai data latih dan 834 ulasan sebagai data uji.

### HASIL EVALUASI

Kelas Sentimen	Skor	Ulasan
Positif	1	Layanannya cepat
Positif	2	Bermanfaat respon dokter cepat
Negatif	-1	Jelek chat dokter aja bayar

		Aktual			
		Class	Positif	Negatif	Jumlah
Prediksi	i	Positif	227	FN = 15	242
		Negatif	FP = 15	575	590
Jumlah			242	590	832

Pada metode *maximum entropy* diperoleh hasil prediksi pada kelas positif, dari 242 ulasan positif yang diuji, terdapat 227 ulasan yang sudah terklasifikasi dengan benar dan terdapat kesalahan prediksi sebesar 15 ulasan yang masuk ke dalam ulasan negatif. Sedangkan pada ulasan negatif yang diuji, dari total 590 ulasan terdapat 575 ulasan yang sudah terklasifikasi dengan benar sebagai ulasan negative dan terdapat kesalahan prediksi sebanyak 15 ulasan yang masuk ke dalam ulasan positif.

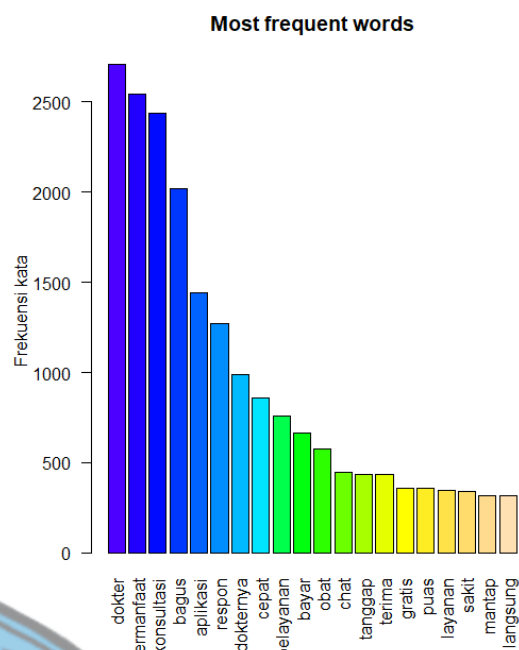
Hasil	Ulasana Halodoc	
	Negatif	Positif
Precision	94%	94%
Recall	98%	98%
F Measure	94,02 %	94,02 %
Accuracy	96,39%	

Berdasarkan hasil klasifikasi menggunakan *maximum entropy* pada tabel diatas untuk ulasan negatif dan positif diperoleh nilai *accuracy* sebesar 96,39% yang nilainya tinggi, hal ini dapat dikatakan bahwa ketepatan dalam memprediksi suatu kata sesuai dengan jawab suatu sistem. Sedangkan untuk *precision* pada kelas negatif sebesar 94% dan kelas positif sebesar 94% hal ini berarti proporsi jumlah dokumen teks yang relevan yang dikenali semua dokumen teks sesuai dengan sistem yakni kelas positif yang nilainya lebih tinggi dari kelas negatif.

Pada nilai *recaal* setiap kelas, kelas positif sebesar 98% dan kelas negatif sebesar 98%. Hal ini dapat dikatakan bahwa tingkat keberhasilan sistem data opini positif dalam menentukan kembali hasil sebuah informasi yaitu hanya memiliki sedikit kesalahan saat proses klasifikasi.

*f-measssure* adalah proporsi pengukur dari kinerja sistem masing-masing kelas sentiment. Kinerja sistem dalam pengklasifikasian kelas positif sebesar 94,02% dan kelas negatif 94,02%.

## Visualisasi Data Sentimen



Berdasarkan gambar dapat dilihat *wordcloud* untuk kata yang paling banyak muncul secara keseluruhan. Dengan menggunakan *wordcloud* dapat dilihat dengan lebih jelas kata-kata yang sering muncul, semakin besar ukuran teks maka semakin besar frekuensi yang dimiliki oleh kata tersebut. Berdasarkan *wordcloud* tersebut dapat dilihat bahwa kata dokter, bermanfaat, konsultasi, bagus dan aplikasi merupakan kata yang mendominasi artinya



Halodoc merasa puas terhadap pelayanan yang diberikan.

## 2. Sentiment negatif

Dokter	Kecewa	Bayar	Obat
- Pemberitahuan 0.27	- Antigen 0.27	- Biaya 0.23	- Penyedia 0.25
- Konsultasi 0.30	- Bicara 0.24	- Servis 0.26	- Apotek 0.29
- Request 0.23		- Kenaikan 0.26	- Buruh 0.27
- Sibuk 0.20			

Kata “dokter” menghasilkan beberapa asosiasi kata yang menginformasikan tentang banyaknya keluhan pengguna karena sibuknya dokter menyebabkan pelayanan kurang maksimal.

Kata “kecewa” menghasilkan beberapa asosiasi kata yang menginformasikan tentang keluhan pengguna terkait konsultasi dengan dokter terlalu cepat.

Kata “bayar” menghasilkan beberapa asosiasi kata yang menginformasikan tentang mahalanya biaya yang dikeluarkan pelanggan untuk membayar jasa dan membeli obat.

Kata “obat” menghasilkan beberapa asosiasi kata yang menginformasikan tentang keluhan pengguna terkait ketersediaan obat yang berada di apotek tidak sesuai yang direkomendasikan dokter, sehingga pelanggan perlu mencari obat yang sesuai dengan resep.

### Simpulan

1. Selama tahun 2020 terdapat 11447 ulasan aplikasi kesehatan *online* Halodoc *website* *Google Play*. Jumlah ulasan tertinggi pada bulan April dan September dipengaruhi oleh munculnya virus covid 19 dan banyaknya orang yang terkonfirmasi positif. Secara umum pengguna merasa puas dengan

kinerja Halodoc dilihat dengan banyaknya kata positif pada ulasan.

2. Pengklasifikasian data menjadi kelas sentiment positif sebanyak 2950 ulasan dan sentiment negatif sebanyak 1213 ulasan. Hasil dari penerapan algoritma *Maximum Entropy* dalam mengklasifikasikan data ulasan pengguna Halodoc menjadi kelas positif dan negatif dengan perbandingan data latih dan uji sebesar 80%:20% diperoleh hasil klasifikasi dengan tingkat akurasi sebesar 96,39%.
3. Pembangunan model dalam pembentukan data latih dan uji dapat meningkatkan ketepatan *machine learning* dari klasifikasi. Berdasarkan *cross validation*, kinerja sistem dalam pengklasifikasian kelas positif 95,47% dan kelas sentiment 95,47%. Dari nilai-nilai tersebut dapat dikatakan tingkat

### Saran

1. Bagi pihak Halodoc, hasil ekstraksi informasi yang diperoleh dari ulasan-ulasan pengguna khususnya ulasan yang berbentuk negatif dapat dijadikan bahan evaluasi dalam meningkatkan kepuasan pengguna dan memberikan pelayanan semaksimal mungkin.
2. Data yang digunakan pada penelitian ini hanya satu periode aplikasi tersebut berjalan pada system *Google Play*, sehingga perlu ditambahkan data agar hasil klasifikasi sentiment lebih baik.
3. Bagi peneliti selanjutnya, dapat menggunakan pendekatan *machine learning* lain sebagai pembandingan performa *Maximum Entropy* untuk mengklasifikasi ulasan aplikasi Halodoc ataupun pada situs lainnya.

### Daftar Pustaka

- APJII. (2018). Penetrasi & Profil Perilaku Pengguna Internet

- Indonesia. *Apjii*. Retrieved from [www.apjii.or.id](http://www.apjii.or.id)
- Anggreini, Dyta. (2008). Klasifikasi Topik Menggunakan Metode *Navie Bayes* dan *Maximum Entropy* pada artikel Media Massa dan Abstrak Tulisan. Skripsi. Universitas Indonesia.
- Alroy, B.A., Adikara, P.P., & Adinugroho, S. (2019). Klasifikasi Hoaks Menggunakan Metode *Maximum Entropy* Dengan Seleksi Fitur *Information Gain* Vol. 3(9): 9291-9298.
- Berry, M.W., & Kogan, J. (2010). *Text Mining Application and Theory*. United Kingdom: WILEY.
- Davies & Beyhon, P. (2004). *Database Systems Third Edition*. New York: Palgrave Macmillan.
- Fadlisyah, B.D.A. (2014). Statistika; Terapannya di Informatika, Edisi Pertama. Yogyakarta: Graha Ilmu.
- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured data*. New York: Cambridge University Press.
- Fritz, G. (2016). Analisa *Hike* pada Kran Lavatory Tipe S11234R di PT Surya Toto Indonesia Tbk. Skripsi. Program Diploma Teknik Mesin Sekolah Vokasi UGM Yogyakarta.
- Indriyarti, E.R., & Wibowo, S. (2020). Bisnis Kesehatan Berbasis Digital Intensi Pengguna Aplikasi Halodoc. Jurnal Pengabdian dan Kewirausahaan Vol. 4(2): 112-121.
- Gemilang, Z.A.N. (2018). Implementasi *Navie Bayes Classifier* dan Asosiasi untuk Analisis Sentimen Data Ulasa Aplikasi *E-Commerce Shoppe* pada situs *Google Play*. Skripsi. Program Studi Statistika FMIPA UII Yogyakarta.
- Han, K.J. (2001). *Data Mining: Concepts and Technique*. San Fransisco: John Wiley & Sons Inc.
- Han, J., & Kember, M. (2006). *Data Mining: Concepts and Techinique Second Edition*. San Frisco: Morgan Kauffman.
- Jamil, H.N. (2017). Analisis Sentimen pada *Online Review* Menggunakan Kombinasi Metode *Lexutcon Based* dan *Navie Bayes Classifer*. Skripsi Program Studi Statistika FMIPA UII Yogyakarta.
- Larose, T. (2005). *Discovering Knowledge in Data: an Introduction To Data Mining*. San Fransisco: Jhon Wiley & Sons Inc.
- Lee, L., & Pang, B. (2008). *Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Reterival*, 2 (1-2): 1-135.
- Liu, B. (2012). *Sentiment Analysis and Subjectivity. Synthesis Lectures on Human Language Technologies*. USA: Morgan & Claypool Publishers.
- Lovelock, C., & Writz, J. (2005). Manajemen Pemasaran Jasa. Indonesia: Kelompok Gramedia Indeks.
- MacIENNAN, J., Tang, Z., & Crival, B. (2009). *Data Mining with Microsoft SQL server 2008*. USA: Wiley Publishing Inc.
- Marres, R., Joan, V.F., & Wilson, P.G. (2013). *Document-level Sentiment Clasification: An Emprical Comparison between SVM and ANN. Export Systems with Application* 40(2): 621-633.

