

## ASPECT BASED SENTIMENT ANALYSIS DATA KUESIONER DI RUMAH SAKIT MUHAMMADIYAH LAMONGAN MENGGUNAKAN ALGORITMA K-NN.

Mustain<sup>1</sup>, Hartarto Junaedi<sup>2</sup>, Endang Setiati<sup>2</sup>

<sup>1</sup>Teknik Informatika, fakultas teknik universitas islam lamongan, <sup>2</sup>Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, <sup>3</sup>Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, Jln Veteran 53 A Lamongan08113595996  
E-mail:mustain@unisla.ac.id

### ABSTRAKS

Kesulitan untuk mengorganisir data kuesioner yang bersifat konvensional melatarbelakangi penelitian ini. Oleh karena itu dibuat sistem yang memudahkan pengelompokan data kuesioner secara otomatis yang lengkap dengan sentimen yang terkandung didalamnya. Dataset yang digunakan dalam penelitian ini adalah data kuesioner rumah sakit Muhammadiyah lamongan. Penelitian ini hanya menangani kuesioner yang berbentuk teks. Data dengan fisik kertas direkap kemudian diinput ke database lengkap dengan kategori unit kerja dan sentiment. Selanjutnya dataset tersebut di dilakukan pre-prosesing yang meliputi penanganan negasi case folding, tokenizing, filtering dan stemming. Sebagai data uji komentar dari kuesioner akan dilakukan pre-prosesing selanjutnya dihitung tingkat kemiripan document dengan menggunakan metode K- Nearest Neighbor dan Vector Space Model. Jumlah data yang ditangani mempengaruhi performa system terutama dari akurasi dan kecepatan pada saat proses klasifikasi. Hasil dari sistem yang dibuat berupa ranking dokumen yang paling mirip dengan dataset berdasarkan urutan nilai cosine similarity. Ujicoba klasifikasi berdasarkan kelas kategori menghasilkan nilai akurasi 91 %. Ujicoba berdasarkan Kelas Sentimen sebesar 94 %. dari kombinasi keduanya system berhasil mendapat akurasi sebesar 86 %.

Kata Kunci: K-nn, Klasifikasi, VSM, Cosine Similarity.

### ABSTRACT

Difficulties to organize conventional questionnaire data are the background of this study. Therefore, a system was created to facilitate the grouping of questionnaire data automatically and completely with the sentiments contained. The dataset used in this study was questionnaire data from the Lamongan Muhammadiyah Hospital. This study only handles questionnaires in the form of text. Data with physical paper is recapped then inputted into a complete database with work unit and sentiment categories. Furthermore, the dataset is pre-processed which includes handling case negation, tokenizing, filtering and stemming. As a comment test data from the questionnaire will be pre-processed then the document similarity level is calculated using the TF IDF, K-Nearest Neighbor method and Vector Space Model. The amount of data handled affects the system performance especially from the accuracy and speed during the classification process. The results of the system are in the form of ranking documents that are most similar to the dataset based on the order of cosine similarity values. Classification tests based on class categories produce an accuracy value of 91%. Trials based on Sentiment Class were 94%. From the combination of the two systems, they managed to get an accuracy of 86%.

Keyword : K-nn, Klasifikasi, VSM, Cosine Similarity.

## 1. PENDAHULUAN

### 1.1 Latar Belakang

Perkembangan ilmu pengetahuan dan teknologi kian pesat. Hal tersebut dapat dilihat secara langsung maupun tidak langsung. Perkembangan tersebut telah berdampak pada segala aspek kehidupan manusia. Salah satu diantaranya adalah kemajuan Sistem Informasi di dunia yang berkembang sangat pesat termasuk di Indonesia hampir semua perusahaan, pendidikan, instansi pemerintah, termasuk juga rumah sakit sudah memanfaatkan sistem informasi sebagai penggerak manajemennya.

Rumah sakit Muhammadiyah Lamongan yang menjadi rumah sakit terbesar dengan luas lahan 39.428 M<sup>2</sup> dan luas lantai ± 16.959 M<sup>2</sup>, serta dilengkapi dengan 232 tempat tidur, Rumah Sakit Muhammadiyah juga sudah memanfaatkan

komputerisasi di semua lini pengelolaannya. Rumah sakit muhammadiyah juga sudah menggunakan jejaring sosial, website, sms gateway, memberikan informasi kepada masyarakat. Untuk menjadi rumah sakit dengan standart di atas rata rata maka tidak mungkin dapat meninggalkan sistem informasi sebagai penyambung hubungan dengan masyarakat.

Informasi yang di terima masyarakat dengan baik berimbas pada banyaknya pasien yang berobat di rumah sakit muhammadiyah Lamongan ± 10.000 orang per tahun. Untuk menjaga agar kepercayaan masyarakat kepada rumah sakit maka rumah sakit membuka layanan kritik dan saran melalui kuisoner dan kotak angket yang di sebar di beberapa titik di rumah sakit. Kritik dan saran yang masuk akan di rekap kemudian di jadikan bahan evaluasi pihak manajemen rumah sakit untuk dapat meningkatkan

kualitas dari pelayanan yang di berikan rumah sakit kepada masyarakat.

Dari banyaknya data yang masuk kerumah sakit, yang di kelola secara manual yang tentu sangat merepotkan. Dari paparan di atas maka dibuatlah sistem *Aspect Based Sentiment Analysis* data Kuesioner menggunakan algoritma K-NN untuk memudahkan pengolahan data dengan jumlah besar, dimana data itu dapat di manfaatkan sebagai referensi pihak manajemen untuk mengambil kebijakan dan menentukan strategi pengembangan jangka panjang.

Adapun dari latar belakang permasalahan yang di hadapi rumah sakit , maka penelitian ini memiliki tujuan dan manfaat sebagai berikut :

1. Menggali informasi yang terkandung didalam laporan masyarakat secara otomatis.
2. Menganalisa sentiment masyarakat kepada rumah sakit muhammadiyah Lamongan.
3. Mengkonversi data yang sifatnya manual kedalam database.

Sedangkan manfaat dari penelitian

1. Bahan evaluasi mutu kepuasan pasien
2. Sebagai bahan referensi untuk menentukan arah kebijakan.
3. Memberikan informasi yang akurat berdasarkan laporan masyarakat.

Bahan laporan tahunan ke direksi rumah sakit muhammadiyah agar dapat di gunakan untuk menentukan kebijakan di tahun mendatang sebagai langkah perbaikan pelayanan rumah sakit.

## 2. METODE PENELITIAN

Adapun metode penelitian yang akan dilakukan adalah supaya memberikan gambaran yang jelas untuk melakukan penelitian ini. Adapun beberapa metode yang dilakukan adalah sebagai berikut :

### a. Studi Literatur

Dengan mempelajari buku-buku referensi dan jurnal yang berkaitan dengan permasalahan penelitian yang diangkat serta mencari solusi yang terbaik. Topik bahasan utama yang dibutuhkan diantaranya klasifikasi dengan metode KNN (*K-Neares neighbor*) dan VSM (*Vector Space Model*).diantara yang menjadi rujukan.

**Tabel 1. Daftar judul referensi yang digunakan**

No	Judul	Penulis
1	sentiment analysis on twitter data using KNN and SVM	Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman

Dalam penelitian yang di lakukan oleh Mohammad Rezwanul Huq, Ahmad Ali, dan Anika Rahman yang berjudul sentiment analysis on twitter data using KNN and SVM peneliti mencoba

membandingkan metode yang paling memiliki tingkat keakuratan dalam perhitungan. Sebagai bahan penelitian Muhammad rezwanul huq serta timnya menjadikan data di twitter dengan panjang maksimal 140 karakter pembatasan ini di lakukan karena untuk mendapatkan twittan yang ringkas dan ekspresif, yang hebat dari penelitian ini peneliti tidak memanfaatkan emoticon sebagai penanda akan tetapi menganggap emoticon adalah kata yang tidak penting.

### b. Analisa

Melakukan uji coba secara teoritis terhadap masalah yang diangkat guna menganalisa apakah rancangan algoritma yang digunakan dapat menghasilkan solusi yang sesuai dengan tujuan penelitian.

### c. Implementasi

Membuat program dari hasil rancangan algoritma yang telah dibuat untuk mengimplementasikan serta membuktikan bahwa hasil analisa secara teoritis yang telah dilakukan benar-benar sesuai yang diharapkan.

### d. Pengujian

Pengujian dilakukan untuk melihat apakah data yang telah menjadi input akan diproses sesuai dengan output yang diharapkan. Hal ini juga dilakukan untuk mengevaluasi apakah metode yang diusulkan mampu menjawab tujuan yang telah usulkan.

### e. Dokumentasi

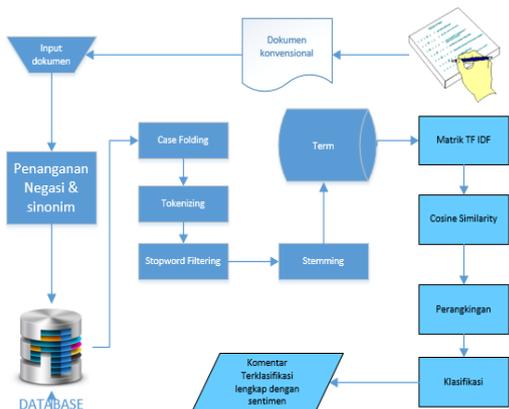
Merupakan langkah akhir, penyusunan laporan mulai dari latar belakang permasalahan sampai dengan pengambilan kesimpulan akan dijelaskan dalam tahap dokumentasi ini.

## 3. ARSITEKTUR SISTEM

Arsitektur sistem yang digunakan untuk menggambarkan sistem kerja yang digunakan dalam proses analisa dan implementasi. Dimulai dari pengumpulan data konvensional dari pasien kemudian dijadikan data file excel yang dapat di masukan kedalam database selanjutnya sebelum dilakukan *preprocessing* dimana dalam arsitektur disebut juga dengan istilah ekstraksi data maka terlebih dahulu dilakukan pelabelan secara manual terhadap data mentah sesuai dengan kelas yang sesuai dengan pesan yang ditulis, sekaligus pelabelan kelas sentimennya. Selanjutnya data disimpan menjadi data berlabel. Ekstraksi data dan diteruskan pada proses pengolahan data latih atau *training* ini nanti yang akan dijadikan penentu kelas klasifikasi terhadap data baru. Semua proses yang di jelaskan di atas itu termasuk proses *preprocessing*.

Setelah mendapatkan sebuah data *training* maka tahap selanjutnya adalah proses klasifikasi dimana untuk proses ini dilakukan dengan menggunakan data kuesioner dari pasien. Proses klasifikasi akan dilakukan dengan cara merubah pesan menjadi matrik dokumen agar dapat dihitung dengan menggunakan algoritma K-NN untuk mencari kelas

apa yang cocok. Ada beberapa kelas diantaranya apotik, kebersihan, jadwalpraktek, fasilitas, pelayanan dan lain-lain. Proses selanjutnya dilakukan proses pengujian klasifikasi analisis sentimentnya berdasarkan kelas positif, negative dan netral. Untuk mempermudah dalam memahami berikut peneliti sajikan desain sistem pada gambar 1.



Gambar 1. Desain Arsitektur Sistem

### 3.1 Analisa Kebutuhan Data

Data yang dipakai dalam penelitian ini adalah bersumber dari kuesioner yang di berikan oleh pihak rumah sakit untuk setiap pasien yang di rawat dan bersumber dari kotak angket yang disediakan disetiap lorong rumah sakit. data yang berasal dari kuesioner hanya akan di ambil pesan yang tertulis di kolom komentar.

Data yang akan dipakai sebagai data penelitian ini sebanyak 1700 record terdiri data latih atau data training yang berjumlah 1400 record yang terbagi menjadi enam kelas sedangkan untuk dari uji sebanyak 300 record. Hal ini dilakukan agar pengujian akan lebih mendapatkan hasil yang maksimal. Selain itu kebutuhan data yang dipakai pada penelitian ini disesuaikan dengan data yang dibutuhkan saat perencanaan awal dan mempercepat proses pelabelan secara manual, karena pada penelitian ini adalah model supervised learning. Artinya butuh data latih untuk menentukan klasifikasi yang baik dan akurat. Semakin banyak data latih semakin baik pulan hasil dari klasifikasi yang diperoleh, namun butuh waktu yang lama dalam proses pelabelannya. Berikut rincian dari data yang digunakan.

Tabel 3. Rincian Data Latih

No	Kategori	Jumlah	Positif	Negatif	Netral
1	Apotik	44	2	38	3
2	Kebersihan	56	10	43	3
3	Jadwal praktek	83	3	74	6
4	Fasilitas	225	40	169	14
5	Pelayanan	770	513	237	64
6	Lain-Lain	222	104	84	34
Total		1400	672	645	124

### 4. KLASIFIKASI DOKUMEN

Preprosesing data dilakukan setelah pasien atau pengunjung rumah sakit mengisi form yang telah disediakan oleh pihak rumah sakit. tahapan-tahapan dalam preprosesing meliputi case folding, tokenizing, filtering dan steaming. Contoh dokumen yang terkategori lengkap dengan sentiment dan query yang ditunjukkan pada tabel di bawah ini.

Tabel 4. Rincian Data Latih

D	Komentar	Sentimen	Kategori
D1	Semua bagus, hanya obat lama antrianya 9	Netral	Apotik
D2	Pendaftaran dan DEPO farmasi lama.	Negatif	Apotik
D3	Ruang perawatan dan kamar mandi kurang bersih.	Negatif	Kebersihan
D4	Pelayanan baik dan menyenangkan susteranya ramah. :	Positif	Pelayanan
D5	Layananpetugas pendaftaran lama, kurang cekatan.	Negatif	Pelayanan
query	Layanan obat sangat lama.	?	?

#### 4.1 Case Folding

Tidak semua dokumen teks konsisten dalam menggunakan huruf kapital. Oleh karena itu, peran case folding dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi bentuk standar "Lowercase" sebagai contoh user yang ingin mendapatkan informasi "pelayanan" dan mengetikkan "PELAYANAN" tetap di berikan kata pelayanan. Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai dengan huruf "z" yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

#### 4.2 Tokenizing

Tahap Tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan. Sebagai contoh karakter whitespace, seperti enter, tabulasi, spasi dianggap sebagai pemisah kata. Namun untuk karakter petik tunggal ('), titik (.), semikolon (;), titik dua (:) atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata.

#### 4.3 Filtering

Tahap Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma stoplist (membuang kata kurang penting)

atau wordlist (menyimpan kata penting). Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari”.

#### 4.4 Stemming

Stemming adalah proses pemetaan dan penguraian bentuk dari suatu kata menjadi bentuk kata dasarnya. Dalam proses Stemming, array kata diperiksa satu-persatu dengan kata dasar, jika terdapat imbuhan maka dilakukan proses penggantian dengan kata dasar. Stemming disini metodenya memakai Porter Stemming dengan kata dasar berupa array.

#### 4.5 Klasifikasi

Pembentukan matrik pada proses klasifikasi menggunakan K-NN melalui proses perhitungan jarak term dokumen latih dan dokumen uji dimana pada penelitian ini menggunakan cosine similarity dengan rumus pecahan 1

$$\text{Cosine } Q = \sin(Q, D_i) = \frac{q \cdot D_i}{|Q| \times |Q_i|} (1)$$

Dan dikarenakan jumlah term adalah continuous-valued attributes, maka normalisasi menggunakan nilai interval 0 hingga 1. Tahap awal yaitu pembentukan matrik TF (Term Frequency) serta normalisasi data seperti yang ditunjukkan pada tabel.

**Tabel 5. Tabel Matrik TF**

Term	Tf (term Frequency)					Tf
	D1	D2	D3	D4	D5	
antrian	1					
bagus	1					
baik				1		
bersih			1			
cepat					1	
daftar		1			1	
depo		1				
farmasi		1				
kamar			1			
kurang			1		1	
lama	1	1			1	1
layan				1	1	1
mandi			1			
obat	1					1
ramah				1		
perawat an			1			
ruang			1			
senang				1		
suster				1		
tugas					1	

Tahap kedua menghitung seberapa sering kata itu muncul dalam dokumen yang akan digunakan untuk menghitung besaran bobot kata pada dokumen.

**Tabel 5. Tabel Matrik Lengkap dengan TF IDF**

term	term frequency (tf)						df	idf log (n/df)
	d1	d2	d3	d4	d5	Q		
antrian	1						1	0,778
bagus	1						1	0,778
baik				1			1	0,778
bersih			1				1	0,778
cepat					1		1	0,778
daftar		1			1		2	0,477
depo		1					1	0,778
farmasi		1					1	0,778
kamar			1				1	0,778
kurang			1		1		2	0,477
lama	1	1			1	1	4	0,176
layan				1	1	1	3	0,301
mandi			1				1	0,778
obat	1					1	2	0,477
ramah				1			1	0,778
perawat an			1				1	0,778
ruang			1				1	0,778
senang				1			1	0,778
suster				1			1	0,778
tugas							1	0,778

Pada tabel di atas *df* yang merupakan jumlah term disemua dokumen, *df* dari term “lama” = 4 yang diperoleh dari jumlah dari tem “Lama” yang terdapat pada dokumen D1,D2,D5 dan terdapat di query atau data uji. Sedangkan perhitungan *idf* (*inverse document frequency*) diperoleh dari  $\log(n/df)$  dimana *n* adalah jumlah semua dokumen termasuk dokumen uji. *Idf* term “lama” diperoleh dari  $\log(6/4) = 0.176$ .

Begitu juga dengan term “layan” = 3 yang diperoleh dari jumlah dari tem “layan” yang terdapat pada dokume D4,D5, dan terdapat di query atau data uji. Sedangkan perhitungan *idf* (*inverse document frequency*) diperoleh dari  $\log(n/df)$  dimana *n* adalah jumlah semua dokumen termasuk dokumen uji. *Idf* term “lama” diperoleh dari  $\log(6/3) = 0.301$ .

Untuk term “obat” = 2 yang diperoleh dari jumlah dari tem “obat” yang terdapat pada dokume D1, dan terdapat di query atau data uji. Sedangkan perhitungan *idf* (*inverse document frequency*) diperoleh dari  $\log(n/df)$  dimana *n* adalah jumlah semua dokumen termasuk dokumen uji. *Idf* term “lama” diperoleh dari  $\log(6/2) = 0.477$ . Tahap ketiga menghitung besaran bobot perkata yang terdapat pada dokumen untuk dapat dibandingkan dengan query. Yang di jabarkan pada tabel 6.

**Tabel 6. Matrik Pembobotan Tiap Term**

term	wdt = tf*idf					Q
	d1	d2	d3	d4	d5	
antrian	0,778	0,000	0,000	0,000	0,000	0,000
bagus	0,778	0,000	0,000	0,000	0,000	0,000
baik	0,000	0,000	0,000	0,778	0,000	0,000
bersih	0,000	0,000	0,778	0,000	0,000	0,000
cekat	0,000	0,000	0,000	0,000	0,778	0,000
daftar	0,000	0,477	0,000	0,000	0,477	0,000
depo	0,000	0,778	0,000	0,000	0,000	0,000
farmasi	0,000	0,778	0,000	0,000	0,000	0,000
kamar	0,000	0,000	0,778	0,000	0,000	0,000
kurang	0,000	0,000	0,477	0,000	0,477	0,000
lama	0,176	0,176	0,000	0,000	0,176	0,176
layan	0,000	0,000	0,000	0,301	0,301	0,301
mandi	0,000	0,000	0,778	0,000	0,000	0,000
obat	0,477	0,000	0,000	0,000	0,000	0,477
ramah	0,000	0,000	0,000	0,778	0,000	0,000
perawatan	0,000	0,000	0,778	0,000	0,000	0,000
ruang	0,000	0,000	0,778	0,000	0,000	0,000
senang	0,000	0,000	0,000	0,778	0,000	0,000
suster	0,000	0,000	0,000	0,778	0,000	0,000
tugas	0,000	0,000	0,000	0,000	0,778	0,000

Pada tabel diatas bobot dari term “lama” diperoleh dari perkalian tf dan idf yaitu  $1 * 0.176 = 0.176$  kemudian untuk term “layan”  $1 * 0.301 = 0.301$  dan “obat” didapatkan nilai  $1 * 0.477 = 0.477$ . selanjutnya akan dihitung dengan menggunakan Euclidean distance seperti yang di jabarkan detail pada tabel

Tahap keempat pada Tabel 7 pembentukan matrik rentang antara dokumen uji dan dokumen latih serta panjang vector. Perhitungan ini juga disebut sebagai vector space model.

Nilai rentang pada term “lama”, “layan” dan term “obat” diperoleh dari perkalian w term dokumen uji dengan w term dari dokumen latih. Untuk term lama Nilai 0.031 pada d1.d2 dan d5 diperoleh dari  $0.176 * 0.176$ . untuk term “layan” mendapatkan nilai 0.091 pada dokumen d4 dan d5 di dapat dari  $0.301 * 0.301$ . kemudian untuk term “obat” mendapatkan nilai 0.228 pada d1 yang di dapat dari  $0.447 * 0.447$  dari nilai rentang ini kemudian di jumlahkan untuk masing masing dokumen. Sedangkan nilai panjang vector adalah nilai kuadrat dari rentang, kemudian dijumlahkan untuk masing-masing dokumen. Hasil dari penjumlahan panjang vector selanjutnya diakar kuadratkan menggunakan rumus SQRT di excel.

**Tabel 7. Matrik Pembobotan Tiap Term**

cosine = wdQ * wdt					panjang vektor = sqrt(wdt)					
d1	d2	d3	d4	d5	d1	d2	d3	d4	d5	Q
0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.228	0.000	0.000	0.228	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.228	0.000	0.228
0.031	0.031	0.000	0.000	0.031	0.031	0.031	0.000	0.000	0.031	0.031
0.000	0.000	0.000	0.091	0.091	0.000	0.000	0.000	0.091	0.091	0.091
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000
0.228	0.000	0.000	0.000	0.000	0.228	0.000	0.000	0.000	0.000	0.228
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.606	0.000
0.259	0.031	0.000	0.091	0.122	1.470	1.470	3.255	2.513	1.788	0.349
					1.212	1.212	1.804	1.585	1.337	0.591

Setelah mendapatkan nilai akar (sqrt) dari panjang vector masing-masing dokumen baik dokumen uji maupun dokumen latih, langkah selanjutnya yaitu menghitung nilai kemiripan dengan cara cosine similarity antara vector kata dengan kata kunci dengan tiap dokumen. Prosesnya menggunakan rumus sebagai berikut :

$$Cos(di) = \frac{\sum rentang di}{[\sqrt{\sum panjangvektor dq}] * \sqrt{\sum panjangvektor di}}$$

sehingga diperoleh perhitungan cosine untuk masing-masing sebagai berikut :

$$\begin{aligned} \cos(dQ,d1) &= 0.259/(0.591*1.212) \\ \cos(dQ,d2) &= 0.031/(0.591*1.212) \\ \cos(dQ,d3) &= 0/(0.591*1.804) \\ \cos(dQ,d4) &= 0.091/(0.591*1.585) \\ \cos(dQ,d5) &= 0.122/(0.591*1.337) \end{aligned}$$

Tahap kelima mengurutkan berdasarkan nilai kemiripan tertinggi untuk digunakan menentukan dokumen mana yang paling mirip dengan query, semakin besar nilainya maka dokumen itu semakin mirip dengan query seperti yang terlihat dalam tabel8.

**Tabel 8. Matrik Pembobotan Tiap Term**

Dokumen	Cosine	Kategori	Sentimen
d1	0,361	Apotek	Netral
d5	0,154	Pelayanan	negatif
d4	0,097	Apotek	negatif

Dengan pengambilan tetangga terdekat sebanyak 1 ( $k=1$ ) maka dokumen uji berlabuh ke kategori apotek dengan sentiment netral. Jika diambil  $k=2$  maka dokumen memiliki 2 kemungkinan yaitu apotek dengan sentiment netral dan pelayanan dengan sentiment negative. Sedangkan jika  $k=3$  maka dokumen memiliki kesamaan dengan dokumen yang terkategori apotek sebanyak 2 dan sentimen memiliki 2 kesamaan yaitu negatif. Maka dapat di ambil jumlah kemunculan terbanyak kategori dan sentiment. Dengan ini dokumen uji ditetapkan masuk kategori apotek dengan sentiment negatif.

## 5. DESAIN SISTEM DAN UJI COBA

### 5.1 Penanganan Negasi

Didalam ujicoba penulis akan menjabarkan proses penanganan negasi secara terpisah dengan proses secara keseluruhan karena negasi di tangani secara khusus didalam aplikasi komentar yang dipakai dalam ujicoba “pelayanan sangat baik, akan tetapi lantainya kurang bersih dan sampah menumpuk”.

Gambar 2. Input Form

Komentar akan di input didalam form komentar pada program seperti yang terlihat pada gambar 5.16. ketika di submit kedalam program akan dilakukan pengecekan apakah komentar yang di input mengandung kata negasi atau tidak. Jika ternyata ada kata negasi didalamnya maka sistem akan mencari kedalam database untuk diganti dengan arti kata sesungguhnya berikut daftar kata yang sudah di masukan kedalam database.

Tabel 9. Daftar Kata Negasi

No	Kata Negasi	Kata Pengganti
1	belum baik	buruk
2	tidak puas	kecewa
3	belum bersih	kotor
4	belum galak	ramah
5	kurang baik	jahat
6	tidak menyenangkan	menyedihkan
7	tidak ramah	galak
8	tidak terawat	terlantar
9	kurang Tepat	meleset
10	tidak cantik	jelek
11	tidak Bermanfaat	sia-sia

Tabel 9. Lanjutan

No	Kata Negasi	Kata Pengganti
12	tidak Besar	kecil
13	tidak baik	buruk
14	kurang ramah	galak
15	tidak menarik	membesankan
16	tidak bersih	kotor
17	kurang Menarik	mebosankan

Sebelum komentar di masukan ke database maka di replace terlebih dahulu dengan tujuan komentar yang di input dapat diterjemahkan oleh sistem dengan arti kata yang jelas.

Gambar 3. Hasil Input Dengan Negasi

### 5.2 Hasil Uji Coba

Dari 300 dokumen dalam proses uji coba, terdapat 257 dokument yang berhasil terklasifikasi dengan tepat. 43 tidak berhasil terklasifikasi dengan tepat klasifikasi dibagi menjadi dua bagian yaitu klasifikasi dengan kelas kategori dan kelas sentimen. Rincian jumlah document dapat di lihat pada table.

Tabel 10. Pembagian Data Uji

NO	Kategori	Jumlah Dokumen
1	Apotik	14
2	Fasilitas	48
3	JadwalPraktek	35
4	Kebersihan	13
5	Pelayanan	135
6	Lain-Lain	55
	Jumlah	300

Hasil dari klasifikasi dengan kelas kategori terdapat 300 data uji dan 275 yang berhasil terklasifikasi dengan benar 25 tidak berhasil terklasifikasi dengan tepat. Sedangkan klasifikasi berdasarkan kelas sentimen 282 yang berhasil terklasifikasi dengan tepat, 18 tidak berhasil terklasifikasi dengan tepat.berikut hasil rician dari ujicoba yang dilakukan.

**Tabel 10. Pembagian Data Uji**

Rician	Jumlah	Berhasil	Tidak	Prosentase
	data		berhasil	
Keseluruhan	300	257	43	86%
Sentimen	300	282	18	94%
Kategori	300	275	25	92%

hasil ujicoba menunjukkan bahwa penggunaan metode K-nearest neighbor telah memenuhi hipotesis yang di usulkan pada proposal yaitu 80 %. Dalam menghitung akurasi keberhasilan digunakan rumus.

$$\text{Akurasi} = \frac{\text{Hasil Klasifikasi Benar}}{\text{Jumlah Data Uji}}$$

$$\text{Akurasi} = \frac{257}{300}$$

$$\text{Akurasi} = 86 \%$$

Akurasi menunjukkan bahwa klasifikasi dengan menggunakan metode K-Nearest Neighbor Hasil yang baik, Karena proses klasifikasi melalui dua kali proses klasifikasi berdasarkan kategori dokumen dan juga sentimen yang terkandung di dalamnya.

## 6. KESIMPULAN

Berdasarkan hasil ujicoba pada sistem diambil kesimpulan antara lain:

1. Dengan menggunakan metode klasifikasi K-Nearest Neighbor terhadap dokumen kuesioner dapat disimpulkan hasilnya baik. Dengan tingkat prosentase 87 %.
2. Metode K-Nearest Neighbor tidak cocok di pakai untuk pemrosesan skala besar, dikarenakan proses training data akan di ulang-ulang secara keseluruhan setelah terdapat data training yang baru. Proses ini akan mengakibatkan sistem berjalan lambat.
3. Kendala yang dihadapi selama penelitian yaitu pada saat preprosesing dikarenakan data kuesioner yang terdapat pada rumah sakit tidak menggunakan kata baku dan juga banyak mengandung kata singkatan.
4. Sistem yang peneliti bangun ini dapat membantu mendistribusikan pesan yang disampaikan oleh pengguna rumah sakit tepat pada target yang di maksud.

Sistem yang dibangun juga dilengkapi dengan penangan negasi agar pesan yang dikirim dapat diartikan sesuai dengan maksud dan arti kata yang sesungguhnya.

## PUSTAKA

- Admin.2017. "Algoritma K-Nearest Neighbor (K-NN)" informatikalogi.com. <https://informatikalogi.com/algoritma-k-nn-k-nearest-neighbor> dia akses 03Juli 2017.
- Ilyantanto.2011. *Stemming Bahasa Indonesia dengan Algoritma Nazief dan Andriani*". 28-juni2011. (<https://liyantanto.wordpress.com/2011/06/28/stemming-bahasa-indonesia-dengan-algoritma-nazief-dan-andriani>). Diakses 03 juli 2017.
- Li Baoli, Yu Shiwen, and Lu Qin, *An Improved k-Nearest Neighbor Algorithm for Text Categorization*, Department of Computer Science and Technology Peking University, Beijing, P.R. China, 100871, 2003.
- Lu Qin, *An Improved k-Nearest Neighbor Algorithm for Text Categorization*, Department of Computer Science and Technology. Peking University, Beijing, P.R. China, 100871, 2003.
- Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman. 2017. "Sentiment Analysis on Twitter Data using KNN and SVM". Vol. 8, No. 6, 2017. (IJACSA) International Journal of Advanced Computer Science and Applications. Dhaka, Bangladesh.
- Mesariya.P, Madia.N, Kumar.A. 2016. "Document Ranking using Customizes Vector Method – A Review", International Journal of Computer Science and Mobile Computing (IJCSMC). pg. 287-290, Vol.5 Issue.3
- Taeho Jo. 2017. "Using K Nearest Neighbors for Text Segmentation with Feature Similarity". Department of Computer and Information Communication Engineering . Hongik University