

ABSTRACT

Health services are an important indicator in assessing the quality of a health system. Bangkalan District Hospital faces challenges in maintaining the quality and efficiency of services, along with the increasing number of patients. Customer reviews are a valuable source of data in understanding public perceptions of hospital services. Therefore, sentiment analysis techniques are needed to classify reviews automatically. This research was conducted to analyze sentiment reviews at RSUD SYAMRABU Bangkalan using random forest. The advantages of the random forest method are that it can manage large amounts of data, less overfitting, resistant to noise and missing value data. The purpose of this research is to find out how the influence of the diversity of the number of trees on the performance of the model. The data used were 987 reviews obtained through the scraping process from Google Maps SYAMRABU Bangkalan Hospital. The data is labeled manually, namely with six labels consisting of one aspect and one sentiment including, positive service, negative service, positive human resources, negative human resources, and positive infrastructure and negative facilities and infrastructure. The research stages include preprocessing, weighting using TF-IDF, and classification using random forest with a binary relevance approach. Model evaluation is carried out using the 10-fold cross validation method. The test results showed a hamming loss value of 0.0765, micro-f1 of 0.9011, micro precision of 0.9140, and micro recall of 0.8889. These values indicate that the developed model is able to perform classification with fairly good performance.

Keywords: *TF-IDF, 10-Fold Cross Validation, SYAMRABU Bangkalan Hospital, Binary Relevance, Random Forest.*

ABSTRAK

Pelayanan kesehatan merupakan indikator penting dalam menilai kualitas suatu sistem kesehatan. RSUD Kabupaten Bangkalan menghadapi tantangan dalam menjaga mutu dan efisiensi pelayanan seiring dengan meningkatnya jumlah pasien. Ulasan pelanggan menjadi sumber data yang bernilai dalam memahami persepsi masyarakat terhadap layanan rumah sakit. Maka dari itu diperlukan teknik analisis sentiment untuk mengklasifikasikan ulasan secara otomatis. Penelitian ini dilakukan untuk menganalisis sentiment ulasan pada RSUD SYAMRABU Bangkalan dengan menggunakan *random forest*. Kelebihan dari metode *random forest* adalah dapat mengelola data dalam jumlah besar, lebih kecil terjadinya *overfitting*, tahan terhadap data *noise* dan *missing value*. Tujuan dari penelitian ini adalah untuk mengetahui seberapa pengaruh keragaman jumlah *tree* terhadap performa model. Data yang digunakan sebanyak 987 ulasan diperoleh melalui proses *scraping* dari *google maps* RSUD SYAMRABU Bangkalan. Data tersebut dilabeli secara manual yaitu dengan enam label terdiri dari satu aspek dan satu sentimen diantaranya, pelayanan positif, pelayanan negatif, sdm positif, sdm negatif, dan sarana prasarana positif dan sarana dan prasarana negatif. Tahapan penelitian meliputi *preprocessing*, pembobotan menggunakan TF-IDF, serta klasifikasi menggunakan *random forest* dengan pendekatan *binary relevance*. Evaluasi model dilakukan dengan metode *10-fold cross validation*. Hasil pengujian menunjukkan nilai *hamming loss* sebesar 0,0765, *micro-f1* sebesar 0,9011, *micro precision* sebesar 0,9140, dan *micro recall* sebesar 0,8889. Nilai-nilai tersebut mengindikasikan bahwa model yang dikembangkan mampu melakukan klasifikasi dengan performa yang cukup baik.

Kata Kunci: TF-IDF, *10-Fold Cross Validation*, RSUD SYAMRABU Bangkalan, *Binary Relevance*, *Random Forest*.

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
LEMBAR PERSEMBAHAN	iv
<i>ABSTRACT</i>	v
ABSTRAK	vi
KATA PENGANTAR	vii
DAFTAR ISI	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiii
DAFTAR SCRIPT	xv
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.2.1 Permasalahan	3
1.2.2 Metode Usulan	3
1.2.3 Pertanyaan Penelitian	3
1.3 Tujuan dan Manfaat Penelitian	4
1.3.1 Tujuan	4
1.3.2 Manfaat	4
1.4 Batasan Masalah	5
1.5 Sistematika Penulisan	6
BAB II TINJAUAN PUSTAKA	7
2.1 <i>Machine Learning</i>	7

2.2	<i>Python</i>	8
2.3	Klasifikasi	9
2.4	Text Mining	9
2.5	<i>K-Fold Cross Validation</i>	10
2.6	Pengukuran Kinerja Klasifikasi	12
2.6.1	<i>Micro Precision</i>	12
2.6.2	<i>Micro Recall</i>	12
2.6.3	<i>Micro F1-Score</i>	13
2.6.4	<i>Hamming Loss</i>	13
2.7	<i>Problem Transformation</i>	14
2.8	<i>Random Forest</i>	17
2.7	<i>Terms Frequency-Inverse Document Frequency (TF-IDF)</i>	19
2.8	<i>Association Rule</i>	20
2.9	Streamlit	21
2.10	Penelitian Terkait	21
BAB III METODE USULAN		29
3.1	Alur Penelitian	29
3.2	Pengumplan Data	30
3.3	<i>Labeling</i>	31
3.4	Analisis Data	35
3.5	<i>Preprocessing</i>	35
3.6	TF-IDF	44
3.7	<i>Binary Relevance</i>	46
3.8	Tahapan Random Forest	47
3.9	Evaluasi	51
3.10	<i>Support, Confidence, Lift</i>	52

3.11	Deployment	52
3.12	Skenario Uji Coba	53
BAB IV HASIL DAN PEMBAHASAN.....		55
4.1	Pembuatan Sistem	55
4.1.1	Pengumpulan Data.....	55
4.1.2	Pelabelan Data.....	57
4.1.3	<i>Data Preparation</i>	57
4.1.4	<i>Preprocessing Data</i>	58
4.1.5	Pembobotan TF-IDF	62
4.1.6	Klasifikasi <i>Random Forest</i> Dengan Pendekatan <i>Binary Relevance</i>	64
4.2	Hasil Skenario Uji Coba	70
4.3	Hasil Akhir	75
4.4	<i>Deployment</i>	79
4.5	<i>Support, Confidence, Lift</i>	81
BAB V KESIMPULAN.....		87
5.1	Kesimpulan	87
5.2	Saran.....	87
DAFTAR PUSTAKA.....		88
LAMPIRAN.....		93
DAFTAR RIWAYAT HIDUP		105

DAFTAR GAMBAR

Gambar 2. 1 <i>K-Fold Cross Validation</i>	11
Gambar 2. 2 Ilustrasi Proses BR.....	15
Gambar 2. 3 Ilustrasi Proses CC.....	16
Gambar 2. 4 Ilustrasi Proses LP	17
Gambar 2. 5 <i>Random Forest</i>	19
Gambar 3. 1 Diagram Alur Sistem	29
Gambar 3. 2 Tahapan Preprocessing	35
Gambar 3. 3 Implementasi Random Forest.....	48
Gambar 3. 4 Use Case Diagram Interface	52
Gambar 3. 5 Activity Diagram Sistem.....	53
Gambar 4. 1 Data hasil scraping.....	57
Gambar 4. 2 Hasil labeling data manual.....	57
Gambar 4. 3 Hasil TF-IDF.....	64
Gambar 4. 4 Decision Tree Label Pelayanan Positif.....	67
Gambar 4. 5 Decision tree Label Pelayanan Negatif.....	67
Gambar 4. 6 DecisionTree Label SDM Positif.....	68
Gambar 4. 7 Decision Tree Label SDM Negatif	68
Gambar 4. 8 Decision Tree Label Sarana dan Prasarana Positif	69
Gambar 4. 9 Decision Tree Label Sarana dan Prasarana Negatif.....	69
Gambar 4. 10 Potongan Visualisasi	70
Gambar 4. 11 Pengujian Jumlah Tree	76
Gambar 4. 12 Pengujian Kedalaman Tree	78
Gambar 4. 13 Menyimpan Model Terbaik	79
Gambar 4. 14 Tampilan Utama Streamlit	80
Gambar 4. 15 Tampilan Streamlit Ketika Memprediksi.....	80

DAFTAR TABEL

Table 2. 1 Contoh Perhitungan Hamming Loss.....	14
Table 2. 2 Penelitian Terkait	21
Tabel 3. 1 Contoh Dataset.....	30
Tabel 3. 2 Dataset Terlabeli	32
Tabel 3. 3 Cleansing	36
Tabel 3. 4 Case Folding	37
Tabel 3. 5 Tokenizing	38
Tabel 3. 6 Stopword Removal	40
Tabel 3. 7 Slangword.....	42
Tabel 3. 8 Steaming	43
Tabel 3. 9 Hasil TF-IDF	44
Tabel 3. 10 Sampel Bootstrap.....	49
Tabel 3. 11 Skenario Pengujian	53
Tabel 4. 13 Hamming Loss Jumlah Pohon.....	71
Tabel 4. 14 Micro-f1 Jumlah Pohon.....	71
Tabel 4. 15 Micro Precision Jumlah Pohon.....	72
Tabel 4. 16 Micro Recall Jumlah Pohon.....	72
Tabel 4. 17 Hamming Loss Kedalaman Pohon	73
Tabel 4. 18 Micro-f1 Kedalaman Pohon	74
Tabel 4. 19 Micro Precision Kedalaman Pohon	74
Tabel 4. 20 Micro Recall Kedalaman Pohon.....	75
Tabel 4. 21 Rata-Rata Evaluasi Jumlah Tree.....	76
Tabel 4. 22 Pegujian Waktu Komputasi Jumlah Tree.....	77
Tabel 4. 23 Rata-Rata Evaluasi Kedalaman Tree	77
Tabel 4. 24 Pegujian Waktu Komputasi Kedalaman Tree	78
Tabel 4. 25 Support Confidence Pelayanan Positif	81
Tabel 4. 26 Support Confidence Pelayanan Negatif.....	82
Tabel 4. 27 Support Confidence SDM Positif.....	82
Tabel 4. 28 Support Confidence SDM Negatif.....	83
Tabel 4. 29 Support Confidence Sarana dan Prasarana Positif.....	83
Tabel 4. 30 Support Confidence Sarana dan Prasarana Negatif.....	84

Tabel 4. 31 Rata-Rata Lift	84
---	----

DAFTAR SCRIPT

Script 4. 1 scraping google maps	56
Script 4. 2 Cleansing	58
Script 4. 3 Case folding	59
Script 4. 4 Tokenizing	60
Script 4. 5 Stopword Removal	60
Script 4. 6 Slangword.....	61
Script 4. 7 Stemming.....	62
Script 4. 8 Pembobotan TF-IDF	63
Script 4. 9 Menampilkan TF-IDF	64
Script 4. 10 Parameter 10-fold cross validation	64
Script 4. 11 Parameter Modeling Random Forest	65
Script 4. 12 Contoh Visualisasi Binary Relevance.....	66