Klasifikasi Teks Multilabel pada Forum Tanya Jawab Kesehatan Bayi (Studi Kasus: Alodokter)

Mery Yulinda Rahmi, Sulis Avandhy Putra, Lidiya Yuniarti

Abstrak

Adanya forum tanya jawab kesehatan bayi seperti pada situs Alodokter dapat menjadi salah satu media informasi yang kredibel bagi para ibu untuk meningkatkan pengetahuan terkait kesehatan bayinya. Namun, terlalu banyaknya informasi dalam forum dapat menyulitkan para ibu saat ingin menemukan topik tertentu yang dibutuhkan. Klasifikasi teks multilabel pada forum tanya jawab kesehatan bayi dapat menjadi langkah dasar dalam pengorganisasian informasi yang ada. Model klasifikasi teks multilabel pada penelitian ini dibangun dengan membandingkan algoritme *Naïve Bayes*, *Linear SVM*, *Logistic Regression*, KNN, *Decision Tree*, *Random Forest*, *XGBoost*, dan *Stacking*; mencoba tiga rasio pembagian data latih dan data uji (70:30, 80:20, 90:10); serta mencoba dua metode ekstraksi fitur kata (TF-IDF dan vektorisasi biner). Hasil penelitian menunjukkan bahwa model terbaik adalah model *Stacking* yang menerapkan rasio pembagian data 90:10 dan TF-IDF dengan nilai rata-rata F1-*score* pada seluruh label mencapai 73,3%, nilai rata-rata *accuracy* sebesar 89,5%, nilai rata-rata *precision* sebesar 79,2%, dan nilai rata-rata *recall* sebesar 68,6%. Dengan menggunakan model terbaik, hasil prediksi terhadap lima dari enam label teks menunjukkan nilai F1-*score* lebih dari 70%. Hasil implementasi model juga menunjukkan bahwa model mampu mengklasifikasikan teks dengan cukup baik.

1 Latar Belakang

Bayi merupakan anak manusia yang baru lahir, dimulai dari usia 0 bulan hingga 12 bulan. Masa bayi dibagi menjadi masa neonatal (usia 0 sampai 28 hari) dan pasca neonatal (usia 29 hari sampai 12 bulan) (Yulianeu & Rahmayati, 2015). Pada tahun 2020, sebanyak 20.266 dari jumlah 28.158 bayi dibawah lima tahun (balita) di Indonesia sebesar 71.97% meninggal dunia pada masa neonatal. Mayoritas kematian balita neonatal (usia 0 sampai 28 hari) sebanyak 35.2% diakibatkan oleh berat badan yang rendah, 27.4% diakibatkan oleh asfiksia, 11.4% dikarenakan kelainan kongenita, infeksi sebesar 3.4%, 0.03% tetanus neonatrium dan 22.5% disebabkan oleh masalah lainnya. Kematian pada bayi pasca neonatal (usia 29 sampai 12 bulan) disebabkan oleh penumonia 14.5% yang menjadi penyebab terbanyak, disebabkan diare sebanyak 9.8%, kelainan kongenital sebanyak 0.5%, penyakit syaraf 0.9% dan sebanyak 73.9% disebabkan oleh masalah lainnya (Kusnandar, 2022). Tingginya angka kematian pada bayi ini memiliki banyak faktor yang menjadi penyebabnya, salah satunya yaitu ketidakpahaman dan ketidaksiapan ibu dalam perawatan bayi baru lahir sehingga menyebabkan masalah kesehatan bahkan kematian pada bayi (Wasiah & Artamevia, 2021).

Forum tanya jawab kesehatan bayi seperti pada situs Alodokter dapat menjadi sumber informasi yang kredibel dalam hal ini. Pada forum tersebut, para ibu dapat berkonsultasi dengan dokter terkait kondisi atau kesehatan bayinya. Adanya media ini dapat meningkatkan pengetahuan para ibu dalam merawat bayinya dengan tepat. Akan tetapi, terlalu banyaknya informasi dalam forum dapat mengakibatkan para ibu selaku pengguna merasa kesulitan dalam mencari dan/atau memahami informasi yang ada.

Klasifikasi teks dapat menjadi tahapan dasar dalam mengorganisir informasi di dalam suatu teks. Penelitian ini melakukan klasifikasi teks multilabel terhadap informasi pada forum tanya jawab kesehatan bayi di Alodokter. Klasifikasi teks multilabel dilakukan dengan mendeteksi sejumlah label informatif di suatu teks, seperti mendeteksi label ‘Penyakit’ pada kalimat yang menjelaskan tentang suatu penyakit pada bayi.

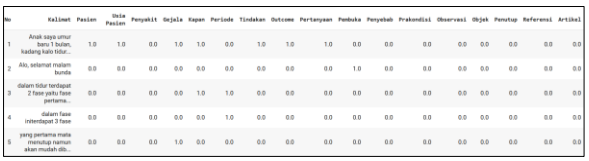
Pada akhirnya, klasifikasi teks multilabel membuat kategorisasi teks menjadi mudah dilakukan, contohnya mengategorisasikan informasi berdasarkan jenis penyakit bayi sehingga mempercepat pengguna dalam menemukan informasi yang dibutuhkan.

Pada penelitian sebelumnya yang dilakukan oleh Efrizoni, dkk., klasifikasi teks multilabel pada teks berita dilakukan dengan menggunakan algoritma *Naïve Bayes, SVM, Decision Tree, KNN, Random Forest,* dan *Logistic Regression* untuk membandingkan kinerja dari ekstraksi fitur BoW, TF-IDF, Doc2Vec, dan Word2vec pada 100 sampel data tribunnews dengn *split* data yang digunakan 50:50. 70:30, 80:20 dan 90:10. (Efrizoni, Defit, Tajuddin, & Anggrawan, 2022). Kemudian, penelitian oleh Jayapermana, dkk. melakukan klasifikasi teks multilabel pada *tweets* tentang vaksin COVID-19 di Twitter menggunakan *stacking ensemble* yang mengombinasikan *base learner* berupa *Logistic Regression, Random Forest,* dan SVM menggunakan *meta-learner* berupa *Logistic Regression* (Jayapermana, Aradea, & Kurniati, 2022).

Dengan demikian, penelitian ini bertujuan membangun model klasifikasi teks multilabel pada forum tanya jawab kesehatan bayi, hingga diperoleh hasil prediksi/klasifikasi yang akurat. Performa prediksi dari model yang berbasis algoritma *Naïve Bayes, Linear SVM, Logistic Regression,* KNN, *Decision Tree, Random Forest, XGBoost,* dan *Stacked Generalization* dengan beberapa skenario eksperimen terhadap rasio pembagian data latih dan data uji (70:30, 80:20, 90:10) serta terhadap metode ekstraksi fitur kata (TF-IDF, vektorisasi biner) akan dibandingkan untuk diketahui algoritma dan skenario mana yang menghasilkan model paling akurat dalam mengklasifikasikan teks.

2 Data

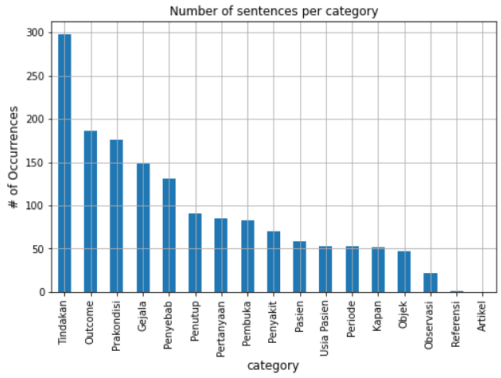
Data yang digunakan dalam penelitian ini bersumber dari salah satu website forum tanya jawab Alodokter. Topik forum tanya jawab yang dipilih sebagai sumber data adalah topik bayi dengan 60 forum tanya jawab. Data yang diambil merupakan data dengan periode dari Mei hingga November 2022. Data mentah selanjutnya melalui proses pelabelan secara manual terhadap setiap kalimat dalam teks. Dari keseluruhan kalimat yang telah dilabeli, terdapat 17 label unik, meliputi label ‘Pasien’, ‘Usia Pasien’, ‘Penyakit’, ‘Gejala’, ‘Kapan’, ‘Periode’, ‘Tindakan’, ‘*Outcome*’, ‘Pertanyaan’, ‘Pembuka’, ‘Penyebab’, ‘Prakondisi’, ‘Obyek’, ‘Penutup’, ‘Referensi’, ‘Observasi’, dan ‘Artikel’. Setelah data teks berhasil dilabeli, maka data mentah ditransformasi menjadi data tabular. Data tabular akan memiliki 18 kolom, yaitu kolom dari setiap label dengan dua kolom tambahan, yaitu nomor dan teks kalimat. Pratinjau dataset ditunjukkan oleh Gambar 1.



No	Kalimat	Pasien	Usia Pasien	Penyakit	Gejala	Kapan	Periode	Tindakan	Outcome	Pertanyaan	Pembuka	Penyebab	Prakondisi	Observasi	Objek	Penutup	Referensi	Artikel
1	Anak saya umur baru 1 bulan, kadang kalo tidur...	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Alo, selamat malam bunda	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	dalam tidur terdapat 2 fase yaitu fase pertama...	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	dalam fase interdapat 3 fase	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	yang pertama mata menutup namun akan mudah dib...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 1: Pratinjau Dataset

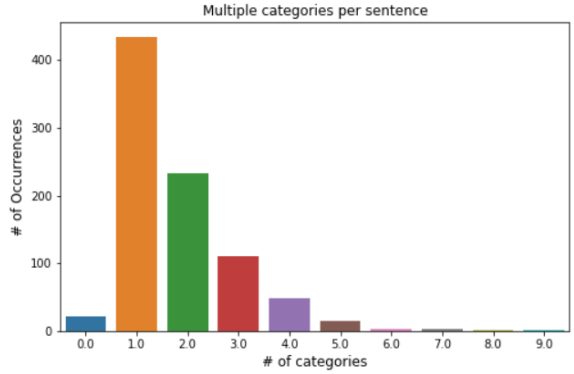
3 Analisis Data



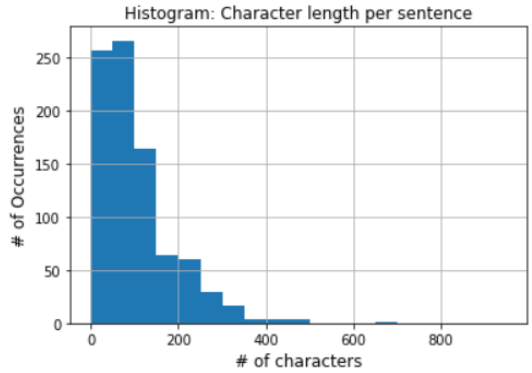
Gambar 2: Jumlah Kalimat Setiap Kategori

Gambar 2 merupakan grafik menunjukkan jumlah kalimat yang ada pada setiap label atau kategori. Dari gambar tersebut, dapat diketahui bahwa label ‘tindakan’ memiliki jumlah kalimat tertinggi dibandingkan dengan label lainnya dengan jumlah 298 kalimat. Hal ini dapat terjadi karena kalimat yang digunakan pada forum ini baik dari pemberi pertanyaan yaitu orang tua pasien maupun penjawab pertanyaan yaitu para dokter lebih sering untuk memberikan pernyataan berupa tindakan yang telah dilakukan atau sugesti untuk melakukan sesuatu dalam menangani permasalahan pada bayi. Selain itu juga, mereka juga biasanya memberikan pernyataan lebih lanjut terkait akibat dari tindakan yang telah

dilakukan. Hal ini juga membuat kalimat yang berlabel ‘outcome’ menjadi label terbanyak kedua setelah ‘tindakan’. Sementara itu, label dengan jumlah terendah adalah label ‘artikel’. Hal ini terjadi karena pemberi dan penjawab pertanyaan sama sekali tidak pernah mengutip suatu tautan untuk dijadikan panduan. Pemberi pertanyaan biasanya langsung menyampaikan secara jelas berdasarkan pengalaman yang dimilikinya dan penjawab pertanyaan biasanya langsung memberikan jawaban secara singkat dan padat sesuai dengan kebutuhan pemberi pertanyaan.



Gambar 3: Kalimat dengan Multi Label



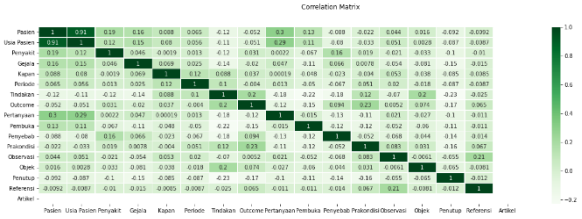
Gambar 4: Panjang Karakter per Kalimat

Gambar 3 merupakan grafik yang menunjukkan jumlah kalimat berdasarkan banyak label yang dimilikinya. Gambar tersebut menunjukkan bahwa sebagian besar kalimat hanya mendapat satu label saja. Selain itu, jumlah kalimat dengan dua label adalah 232 kalimat, dengan tiga label 111 kalimat dengan empat label 48 kalimat, dengan lima label 15 kalimat, dengan enam dan tujuh label masing-masing tiga kalimat dan dengan delapan dan sembilan label masing-masing satu kalimat. Ada pula 21 kalimat yang tidak terlabeli sama sekali. Adanya kalimat yang memiliki beberapa label sekaligus merupakan konsekuensi dari adanya beberapa pernyataan yang diberikan mengandung beberapa informasi seperti informasi pasien dengan usianya atau informasi tindakan yang dilakukan dengan luaran yang diharapkan.

Gambar 4 menunjukkan terkait distribusi jumlah karakter pada setiap kalimatnya. Dari gambar tersebut, dapat diketahui bahwa sebagian besar kalimat memiliki jumlah karakter kurang dari 200 karakter. Selain itu juga, diketahui ada beberapa kalimat yang berperilaku sebagai pencilan dengan jumlah karakter berada di antara 600 hingga 800 karakter.

Gambar 5: *Wordcloud* pada Lima Label

Gambar 5 merupakan visualisasi dari *wordcloud* pada lima label teratas dalam segi jumlah kalimat. Kelima gambar tersebut memilki kesamaan yaitu kalimat yang paling sering muncul adalah bayi. Hal ini wajar terjadi mengingat topik forum yang digunakan sebagai sumber data. Pada *wordcloud* label ‘gejala’, dapat terlihat beberapa kata yang menonjol yaitu ‘rewel’, ‘muntah’, dan ‘tidur’. Ketiga kata ini mengindikasikan bahwa gejala yang sering terjadi pada pasien bayi adalah bayi menjadi rewel, bayi sering muntah, atau bayi yang terganggu dalam tidurnya. Selain itu, *worcloud* label ‘penyebab’ memberikan informasi bahwa penyakit yang sering diderita bayi lebih sering terjadi karena infeksi baik virus, bakteri, maupun patogen lainnya yang dapat menyerang berbagai sistem tubuh milik bayi.



Gambar 6: Matriks Korelasi antar Label

Gambar 6 menunjukkan matriks korelasi antar label pada dataset. Sebagian besar hasil yang didapatkan tidak menunjukkan adanya korelasi yang signifikan antar labelnya. Hanya ada satu pasangan label yang berkorelasi tinggi yaitu label 'pasien' dengan 'usia pasien'. Hal ini menunjukkan bahwa kemunculan label 'pasien' pada suatu kalimat dibarengi dengan kemunculan label 'usia pasien'.

4 Metodologi Eksperimen

Eksperimen yang dilakukan terbagi menjadi tujuh tahapan, yaitu pembersihan teks, seleksi label, pembagian data latih dan data uji, transformasi data, pembangunan model, serta uji coba dan evaluasi model. Pada tahap pertama, pembersihan teks dilakukan dengan mengubah seluruh *string* menjadi *lowercase*, memperbaiki kata-kata yang masih menggunakan bahasa gaul dengan cara membuat *dictionary* yang berisikan bahasa gaul dan padanan kata bakunya sesuai KBBI, menghapus *special character*, serta mengatur semua *whitespace* menjadi satu spasi.

Label	Label yang Digabungkan
Usia Pasien	'Pasien', 'Usia Pasien'
Penyakit dan Kondisi Pasien	'Penyakit', 'Gejala', 'Penyebab', 'Prakondisi', 'Observasi'
Keterangan Waktu	'Kapan', 'Periode'
Tindakan dan Hasil Tindakan	'Tindakan', '*Outcome*', 'Objek'
Pertanyaan	'Pertanyaan'
Salam	'Pembuka', 'Penutup

Tabel 1: Hasil Gabungan Label

Kemudian, tahap seleksi label mempertimbangkan label 'Artikel' (0 kalimat) dan label 'Referensi' (1 kalimat) untuk tidak digunakan dalam membangun model karena kemunculannya yang sangat jarang. Maka, total label yang digunakan adalah sebanyak 15 dari 17 label. Kelimabelas label lalu digabungkan menjadi hanya enam label. Hal ini dilakukan guna mempermudah model dalam mengklasifikasikan label-label pada teks. Di samping itu, alasan penggabungan label ini dikarenakan terdapat beberapa label yang redundan seperti pada label 'Usia Pasien' dan 'Pasien' serta label-label yang kurang bermakna jika berdiri sendiri, seperti label 'Prakondisi' dan 'Observasi. Enam label baru yang merupakan hasil penggabungan kelimabelas label ditunjukkan oleh Tabel 1.

Dataset lalu dibagi menjadi data latih dan data uji dengan mencoba beberapa skenario, yakni membagi dengan rasio data latih terhadap data uji sebesar **70:30, 80:20, dan 90:10** dari keseluruhan dataset. Data latih merupakan data yang digunakan untuk membangun model dan data uji sebagai data untuk validasi model.

Sebelum membangun model, data teks yang masih dalam bentuk kalimat perlu ditransformasi. Transformasi data meliputi *stemming* dan ekstrasi fitur kata. *Stemming* yaitu mengubah kata ke dalam bentuk kata dasarnya (*stem*) guna mereduksi tingginya dimensionalitas dari fitur kata (Efrizoni, Defit, Tajuddin, & Anggrawan, 2022). Dalam eksperimen ini, *stemming* dilakukan dengan menggunakan *tool* berupa StemmerFactory dari *library* Sastrawi. Sedangkan, ekstraksi fitur kata adalah memetakan kata ke vektor numerik sehingga dapat diproses untuk komputasi (Efrizoni, Defit, Tajuddin, & Anggrawan, 2022). Dalam mengekstraksi fitur kata, *stop words* dalam corpus diabaikan. Kamus *stop words* bahasa Indonesia dari *library* NLTK digunakan dalam hal ini. Pendekatan ekstraksi fitur kata pada penelitian ini juga menggunakan pendekatan yang sering digunakan untuk mengekstraksi fitur kata, antara lain representasi biner yang menunjukkan ada atau tidaknya kata dalam sebuah kalimat dengan 0 dan 1 serta representasi berupa bobot berdasarkan rumus TF-IDF. Pada eksperimen ini, transformasi data akan mencoba dua skenario, yaitu **ekstraksi fitur kata dengan representasi biner atau dengan rumus TF-IDF**. Rumus TF-IDF ditunjukkan oleh persamaan (1) (Pedregosa, et al., 2011).

$$tf.idf(t,d) = tf(t,d) * [\log(\frac{n}{df(t)}) + 1] \quad (1)$$

Setelah itu, sejumlah algoritme dipilih untuk membangun model klasifikasi multilabel. Algoritme dalam eksperimen ini meliputi *Naive Bayes*, *Linear SVM*, *Logistic Regression*, KNN, *Decision Tree*, serta teknik *ensemble*, yaitu *Random Forest (bagging)*, *XGBoost (boosting)*, dan *Stacked Generalization (stacking)* yang mana *stacking* ini mengombinasikan ketujuh algoritme sebelumnya (Jayapermana, Aradea, & Kurniati, 2022; Efrizoni, Defit, Tajuddin, & Anggrawan, 2022; Murtopo, Pratiwi, & Fadilah, 2022). Pada

Algoritme	***Library***	***Hyperparame-ter tuning***
Naive Bayes	sklearn. naive_bayes. MultinomialNB	fit_prior= False, class_prior = None
Linear SVM	sklearn. svm. LinearSVC	C= 0.1, dual= False, class_weight= 'balanced', random_state= 42
Logistic Regression	sklearn. linear_model. LogisticRegression	C= 0.1, class_weight= 'balanced', dual= False
KNN	sklearn. neighbors. KNeighborsClassifier	n_neighbors= 5, weights='distance'
Decision Tree	sklearn. tree. DecisionTreeClassifier	criterion='entropy, random_state=42
Random Forest	sklearn. ensemble. RandomForestClassifier	max_features= None, random_state= 42, min_impurity_decrease= 0.001, min_weight_fraction_leaf= 0.001
XGBoost	xgboost. XGBClassifier	base_score= 0.2, booster= 'gbtree', gamma= 0, learning_rate= 0.1, n_estimators= 500, reg_alpha= 0, reg_lambda= 1, random_state= 0
Stacking	sklearn. ensemble. StackingClassifier	final_estimator= LogisticRegression(), cv= 5

Tabel 2: *Hyperparameter tuning.*

saat membangun model, perlu juga dilakukan *hyperparameter tuning*, yaitu proses pengaturan *hyperparameter* dari suatu algoritme guna mengontrol algoritme tersebut agar dapat bekerja secara optimal (Naury, Fudholi, & Hidayatullah, 2021). *Hyperparameter tuning* terhadap setiap algoritme tersebut ditunjukkan pada Tabel 2.

Setelah *hyperparameter tuning*, algoritme kemudian dilatih terhadap data latih menggunakan strategi klasifikasi multilabel yang disebut sebagai *one-vs-the-rest. One-vs-the-rest* akan melatih model sebanyak *n* kali sesuai dengan banyaknya label *(n)* (Pedregosa, et al., 2011). Dengan begitu, performa model dalam memprediksi masing-masing label secara terpisah dapat lebih mudah diketahui.

Pada akhirnya, model perlu memprediksi data uji sehingga dapat dievaluasi keakuratan prediksinya. Evaluasi model didasarkan pada metrik *accuracy,* F1-*score*, *precision,* dan *recall* yang telah umum digunakan dalam permasalahan klasifikasi teks multilabel. Rumus *accuracy*, F1-*score*, *precision*, dan *recall* ditunjukkan pada persamaan (2), (3), (4), dan (5) secara berurutan (Efrizoni, Defit, Tajuddin, & Anggrawan, 2022).

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (2)$$

$$F1-score = \frac{2*precision*recall}{precision+recall} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

5 Hasil dan Analisis Eksperimen

Eksperimen dilakukan terhadap kedelapan jenis algoritme yang akan menjadi model klasifikasi. Eksperimen ini terdiri dari enam skenario yang memvariasikan rasio data latih dengan data uji (70:30, 80:20, 90:10) dan memvariasikan metode ekstraksi fitur kata (vektorisasi biner dan TF-IDF).

Rank	**Algoritme**	**Rasio Data Latih dan Data Uji**	**Metode Ekstraksi Fitur**	**Rata-rata F1-*score***	**Rata-rata *Acc.***	**Rata-rata *Prec.***	**Rata-rata *Recall***
1	*Stacking*	90:10	TF-IDF	73,3%	89,5%	79,2%	68,6%
2	*Linear SVM*	70:30	Vektorisasi Biner	71,5%	88,5%	71,6%	74,5%
3	*Random Forest*	90:10	TF-IDF	71%	89,3%	79,5%	65,3%
4	*Decision Tree*	70:30	Vektorisasi Biner	70,8%	89%	73,6%	69,2%
5	*Logistic Regression*	80:20	Vektorisasi Biner	70,3%	88%	66,3%	77,9%
6	*XGBoost*	70:30	Vektorisasi Biner	67,2%	89%	81%	60,7%
7	*Naïve Bayes*	70:30	Vektorisasi Biner	58,8%	84,3%	52,1%	72,5%
8	KNN	90:10	TF-IDF	54,1%	87%	86,1%	42,2%

Tabel 3: Skenario Terbaik untuk Setiap Jenis Algoritme (diurutkan berdasarkan F1-*score).*

Pada Tabel 3, ditunjukkan bahwa model berbasis algoritme *Stacking* dengan rasio pembagian data latih dan data uji sebesar 90:10 dan ekstraksi fitur menggunakan TF-IDF mempunyai nilai rata-rata F1-*score* yang paling unggul sebesar 73,3%. Model ini juga menghasilkan nilai *accuracy* sebesar 89,5%, nilai *precision* sebesar 79,2%, dan nilai *recall* sebesar 68,6%. Nilai-nilai tesrsebut merupakan nilai rata-rata keakuratan prediksi terhadap keenam label.

Lima model teratas telah menunjukkan nilai rata-rata F1-*score* lebih dari 70% dan nilai *accuracy* mendekati 90%. Selain itu, ditemukan bahwa model yang lebih cocok terhadap metode ekstraksi fitur berupa TF-IDF, ternyata menggunakan lebih banyak data latih. Hal ini ditunjukkan oleh model berbasis algoritme *Stacking*, *Random Forest*, dan KNN yang mana skenario terbaiknya yakni menggunakan rasio pembagian data sebesar 90:10 dan TF-IDF

sebagai metode ekstraksi fiturnya. Sementara itu, pada algoritme lainnya yang berperforma lebih baik jika menerapkan vektorisasi biner, model tersebut cenderung membutuhkan data latih yang relatif lebih sedikit. Contohnya, model *Linear SVM* menggunakan rasio pembagian data sebesar 70:30 dan *Logistic Regression* menggunakan rasio 80:20, dimana keduanya sama-sama lebih cocok dengan metode vektorisasi biner.

Rank	Label	F1-*score*	*Accuracy*	*Precision*	*Recall*
1	Salam	73,3%	89,5%	79,2%	68,6%
2	Usia Pasien	71,5%	88,5%	71,6%	74,5%
3	Tindakan dan Hasil Tindakan	71%	89,3%	79,5%	65,3%
4	Pertanyaan	70,8%	89%	73,6%	69,2%
5	Penyakit dan Kondisi Pasien	70,3%	88%	66,3%	77,9%
6	Keterangan Waktu	67,2%	89%	81%	60,7%

Tabel 5: Hasil Prediksi terhadap Setiap Label menggunakan Model Terbaik (*Stacking*, 90:10, TF-IDF) dan diurutkan berdasarkan F1-*score*

Dengan menggunakan model terbaik, yakni model berbasis *Stacking* (rasio 90:10, TF-IDF), dihasilkan keakuratan prediksi untuk setiap labelnya seperti pada Tabel 4. Dari enam label yang ada, terdapat lima label dengan nilai F1-*score* yang mencapai lebih dari 70%. Label 'Salam' dapat diprediksi dengan baik dikarenakan kalimat yang berlabel 'Salam' mempunyai pola yang sangat jelas, dimana sebagian besar kalimat tersebut adalah "Selamat pagi/siang/sore/malam" dan "Semoga membantu". Selanjutnya, label 'Usia Pasien' terdapat pada kalimat dengan kata-kata, seperti "bayi" atau "usia *x* bulan" sehingga label ini dapat diprediksi dengan baik. Kemudian, label 'Tindakan dan Hasil Tindakan' juga diprediksi dengan baik dikarenakan saran tindakan yang diberikan oleh dokter umumnya hampir sama. Contoh saran tersebut, seperti "jika terdapat tanda-tanda seperti ini, maka segera larikan anak ke UGD terdekat", "untuk memastikannya, bunda bisa konsultasikan atau tanyakan langsung ke dokter anak", dan lain sebagainya. Kalimat dengan label 'Pertanyaan' memiliki kata-kata penciri, seperti "bagaimana" dan "apa". Akibatnya, label tersebut dapat diprediksi dengan cukup baik. Label 'Penyakit dan Kondisi Pasien' berada pada kalimat, seperti "gangguan tersebut dapat disebabkan oleh infeksi bakteri", "apabila anak masih mengalami gejala, seperti rewel, muntah, demam…", "kondisi tersebut wajar terjadi pada bayi", dan lain sebagainya. Adanya pola kata yang cukup kentara pada label ini membuat model dapat memprediksinya dengan cukup baik pula. Label pada peringkat terakhir yaitu 'Keterangan Waktu'. Hal ini disebabkan oleh sedikitnya data teks yang berlabel 'Kapan' ataupun berlabel 'Periode' sehingga model belum terlalu memahami ciri-ciri kalimat yang memiliki label tersebut. Walaupun demikian, nilai F1-*score*-nya tidak terlalu jauh dari 70%. Beberapa kata penciri dari label ini yaitu "fase", "jam", "menit", "hari", "minggu", dan "kemarin.

No.	Teks	(Prediksi) Label Teks
1.	Saya baru saja melahirkan dan saat ini usia bayi saya 2 bulan, Kalo dilihat2 kok ukuran kepala bayi saya lebih kecil dari bayi lainnya ya dan bentuk lehernya pendek..	1. 'Usia Pasien' 2. 'Penyakit dan Kondisi Pasien'
2.	Beberapa jenis kelainan fisik yang bisa diamati pada penderita Down syndrome ialah ukuran kepala yang lebih kecil, bagian belakang kepala datar, hidung dan mulut kecil, sudut mata luar naik ke atas, muncul bintik-bintik putih di iris mata, telinga kecil atau abnormal bentuknya, leher pendek, kulit leher belakang kendur, telapak tangan lebar dan hanya terdapat 1 garis tangan, tungkai kecil dan pendek jarinya, serta otot lemah dan sangat lentur.	'Penyakit dan Kondisi Pasien'
3.	Selamat siang alodokter, dok saya ibu dari bayi usia 3 bulan, bayi saya sekitar 2 jam sekali nyusu, dia masih ASI dok, saya mau bertanya berapa sering ya bayi usia 3 bulan pipis selama sehari semalam ya?	1. 'Salam' 2. 'Usia Pasien' 3. 'Pertanyaan' (gagal memprediksi label 'Penyakit dan Kondisi Pasien')

Tabel 4: Hasil Implementasi Model Klasifikasi

Model yang telah dibangun perlu diimplementasikan untuk memastikan apakah model dapat bekerja dengan baik. Implementasi dilakukan dengan menggunakan model terbaik untuk mengklasifikasikan teks yang belum pernah dilihat oleh model sebelumnya. Hasil implementasi ditunjukkan pada Tabel 5. Berdasarkan hasil tersebut, dapat dilihat bahwa model yang dibangun mampu mengklasifikasikan teks multilabel dengan cukup baik, dimana teks diambil dari forum tanya jawab kesehatan bayi pada situs Alodokter.

6 Kesimpulan

Model klasifikasi teks multilabel pada forum tanya jawab Bayi dibangun dengan membandingkan performa prediksi dari sejumlah model yang berbasis pada delapan jenis algoritme, yaitu *Stacking*, *Linear SVM*, *Random Forest*, *Decision Tree*, *Logistic Regression*, *XGBoost*, *Naïve Bayes*, dan KNN. Selain itu, skenario eksperimen yang dilakukan adalah dengan mencoba sejumlah rasio pembagian data latih dan data uji (70:30, 80:20, 90:10) serta metode ekstraksi fitur kata (vektorisasi biner dan TF-IDF). Hasil eksperimen menunjukkan bahwa model terbaik adalah model berbasis *Stacking* dengan rasio pembagian data 90:10 dan ekstraksi fitur kata menggunakan metode TF-IDF sehingga menghasilkan nilai rata-rata F1-*score* pada seluruh label mencapai 73,3%, nilai rata-rata *accuracy* sebesar 89,5%, nilai rata-rata *precision* sebesar 79,2%, dan nilai rata-rata *recall* sebesar 68,6%.

Berdasarkan hasil prediksi dari model terbaik, diperoleh lima dari enam label (hasil penggabungan 15 label) yang telah mencapai nilai F1-*score* lebih dari 70%, yakni meliputi label ‘Salam’, ‘Usia Pasien’, ‘Tindakan dan Hasil Tindakan’, ‘Pertanyaan’, serta ‘Penyakit dan Kondisi Pasien’. Sementara itu, label ‘Keterangan Waktu’ belum diprediksi dengan cukup baik, yang mana perolehan nilai F1-*score*-nya sebesar 67,2% dikarenakan sedikitnya jumlah data untuk label tersebut. Walaupun demikian, implementasi dari model terbaik menunjukkan bahwa model cukup baik dalam mengklasifikasikan label-label pada teks yang belum pernah dilihat oleh model sebelumnya.

Penelitian kedepannya dapat membangun model klasifikasi teks multilabel yang tidak hanya menangkap pola kata-kata dalam teks, namun juga mempertimbangkan makna serta konteks kata di dalam kalimat. Model klasifikasi tersebut dapat menggunakan model bahasa berbasis *transfer learning*, seperti BERT, GPT-3, dan XLNet.

7 Referensi

Efrizoni, L., Defit, S., Tajuddin, M., & Anggrawan, A. (2022). Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel menggunakan Algoritma Machine Learning. *Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, *21*(3), 653-666.

Jayapermana, R., Aradea, A., & Kurniati, N. I. (2022). Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topics on Twitter. *Scientific Journal of Informatics*, *9*(1), 8-15.

Kusnandar, V. B. (2022). *Demografi*. Diambil kembali dari Kematian Balita di Indonesia Capai 28,2 Ribu pada 2020: https://databoks.katadata.co.id/datapublish/2021/10/22/kematian-balita-di-indonesia-capai-282-ribu-pada-2020#:~:text=Mayoritas%20atau%2035%2C2%25%20kematian,%2C%20yakni%2014%2C5%25.

Murtopo, A. A., Pratiwi, A., & Fadilah, N. (2022). TINJAUAN PUSTAKA SISTEMATIS: KLASIFIKASI UJARAN KEBENCIAN PADA SOSIAL MEDIA DENGAN ALGORITMA NEURAL NETWORK. *Indonesian Journal of Informatics and Research*, *3*(1), 49-57.

Naury, C., Fudholi, D. H., & Hidayatullah, A. F. (2021). Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM. *Jurnal Media Informatika Budidarma*, 24-33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Wasiah, A., & Artamevia, S. (2021). Pelatihan Perawatan Bayi Baru Lahir. *Journal of Community Engagement in Health*.

Yulianeu, A., & Rahmayati, N. M. (2015). SISTEM PAKAR PENENTU MAKANAN PENDAMPING AIR SUSU IBU PADA BAYI USIA 6 BULAN SAMPAI 12 BULAN MENGGUNAKAN METODE FORWARD CHAINING. *JUTEKIN (Jurnal eknik Informatika)*.