FINAL PROJECT

KLASIFIKASI TEKS MULTI-LABEL PADA FORUM TANYA JAWAB KESEHATAN BAYI (STUDI KASUS: ALODOKTER)

PENGOLAHAN BAHASA ALAMI (A)
KELOMPOK 7

Mery Yulinda Rahmi 05211940000003
Sulis Avandhy Putra 05211940000084
Lidiya Yuniarti 05211940007001

LATAR BELAKANG

- **Bayi** merupakan anak manusia yang baru lahir, usia bayi dimulai dari **0 bulan-12 bulan. Massa bayi** terbagi menjadi **neonatal (usia 0-28 hari)** dan **pasca neonatal (usia 29-12 bulan).**
- Pada **tahun 2020,** sebanyak **71.97% balita di Indonesia meninggal dunia** pada masa neonatal. Tingginya angka kematian pada bayi ini **disebabkan** oleh beberapa **faktor** diantaranya **virus atau kuman** yang terjadi **selama proses persalinan** atau **setelah persalinan** dan **menyebabkan bayi rentan terhadap penyakit** dan **faktor lain** seperti **kurangnya pengetahuan ibu dalam marawat bayi dan ketidak siapan ibu dalam merawat bayi**. **Sehingga bayi membutuhkan perawatan yang baik agar kesehatan bayi terjaga**.
- Pada penelitian **klasifikasi teks multi-label** dilakukan untuk **memprediksi hasil klasifikasi dari forum tanya jawab Alodokter (bayi)** yang menggunakan beberapa algoritma **untuk menghasilkan informasi terkait bayi dan dari informasi ini diharapkan dapat menjadi langkah preventif untuk lebih memperhatikan kesehatan bayinya**.

DATA

Sumber Data :
Forum tanya jawab ALODOKTER dengan **topik "Bayi"** dengan jumlah forum yang digunakan sebagai data adalah 60 forum tanya jawab

Periode Data :
Mei 2022 - November 2022

Link Data :
https://www.alodokter.com/komunitas/topic-tag/bayi/page/1

DATA

Label Teks :

1. Pasien
2. Usia Pasien
3. Penyakit
4. Gejala
5. Kapan
6. Periode
7. Tindakan
8. Outcome
9. Pertanyaan
10. Pembuka
11. Penyebab
12. Prakondisi
13. Objek
14. Penutup
15. Referensi
16. Observasi

*17. Artikel

Profil Dataset :

- Berukuran 869 baris x 19 kolom
- Kolom berisi : "No", "Kalimat", dan Kolom Label Teks

*17. Artikel = tidak ada label "Artikel" pada semua teks yang digunakan

ANALISIS STATISTIK

Didapatkan hasil analisis jumlah kalimat di setiap kategori:

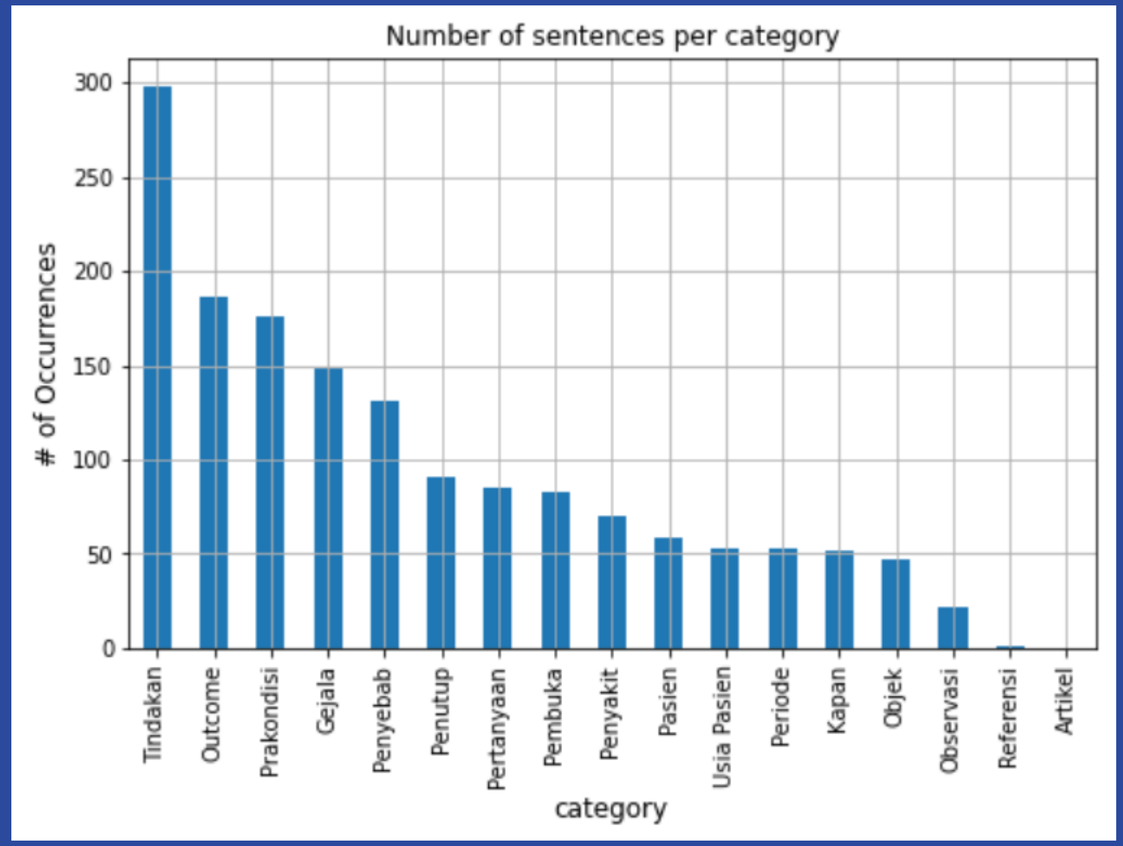
jumlah **label terbanyak:**

1. **Tindakan** sebanyak 298 kalimat
2. **Outcome** sebanyak 186 kalimat
3. **Prakondisi** sebanyak 176 kalimat

dan jumlah **label tersedikit:**

1. **Artikel** sebanyak 0 kalimat
2. **Referensi** sebanyak 1 kalimat
3. **Observasi** sebanyak 22 kalimat

	category	number_of_sentences
6	Tindakan	298.0
7	Outcome	186.0
11	Prakondisi	176.0
3	Gejala	148.0
10	Penyebab	131.0
14	Penutup	91.0
8	Pertanyaan	85.0
9	Pembuka	83.0
2	Penyakit	70.0
0	Pasien	59.0
1	Usia Pasien	53.0
5	Periode	53.0
4	Kapan	51.0
13	Objek	47.0
12	Observasi	22.0
15	Referensi	1.0
16	Artikel	0.0

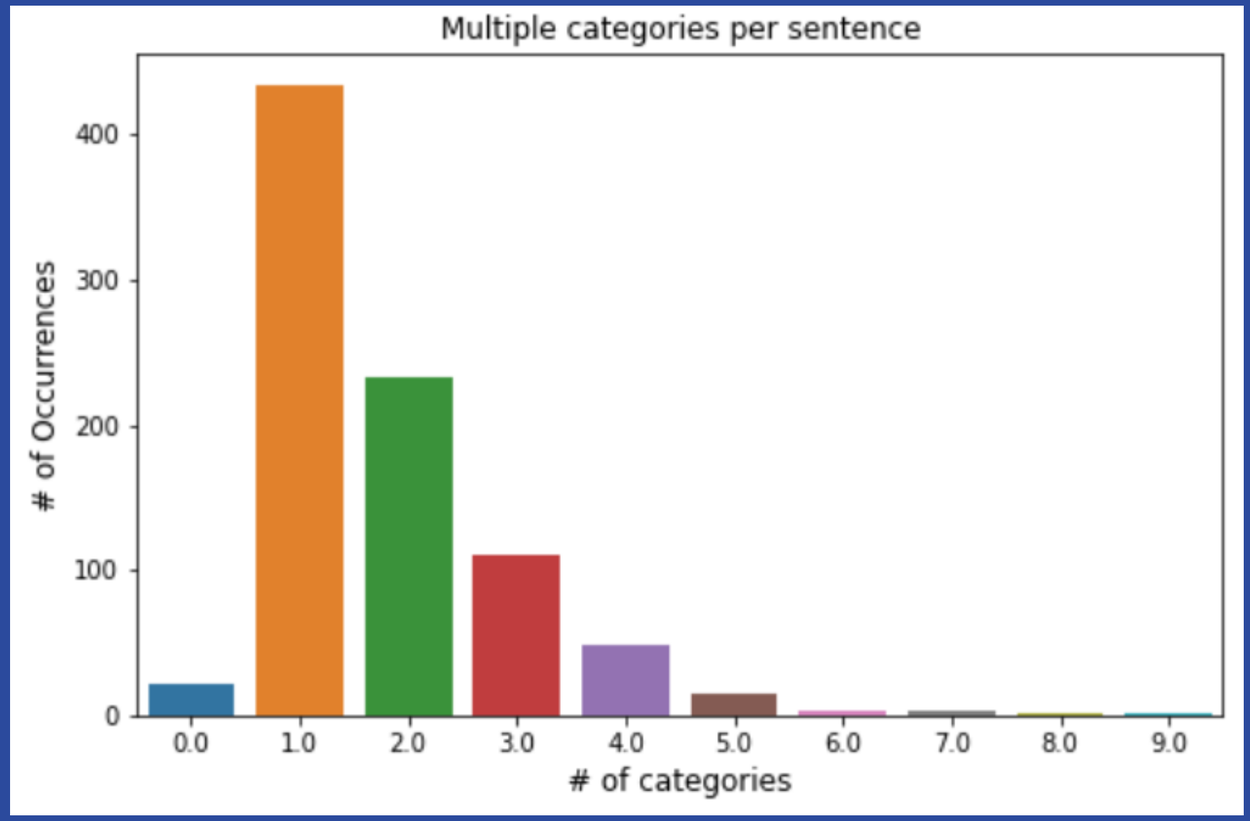


ANALISIS STATISTIK

Berdasarkan grafik tersebut, dapat diketahui bahwa **pada umumnya, kalimat hanya memiliki 1 label.**

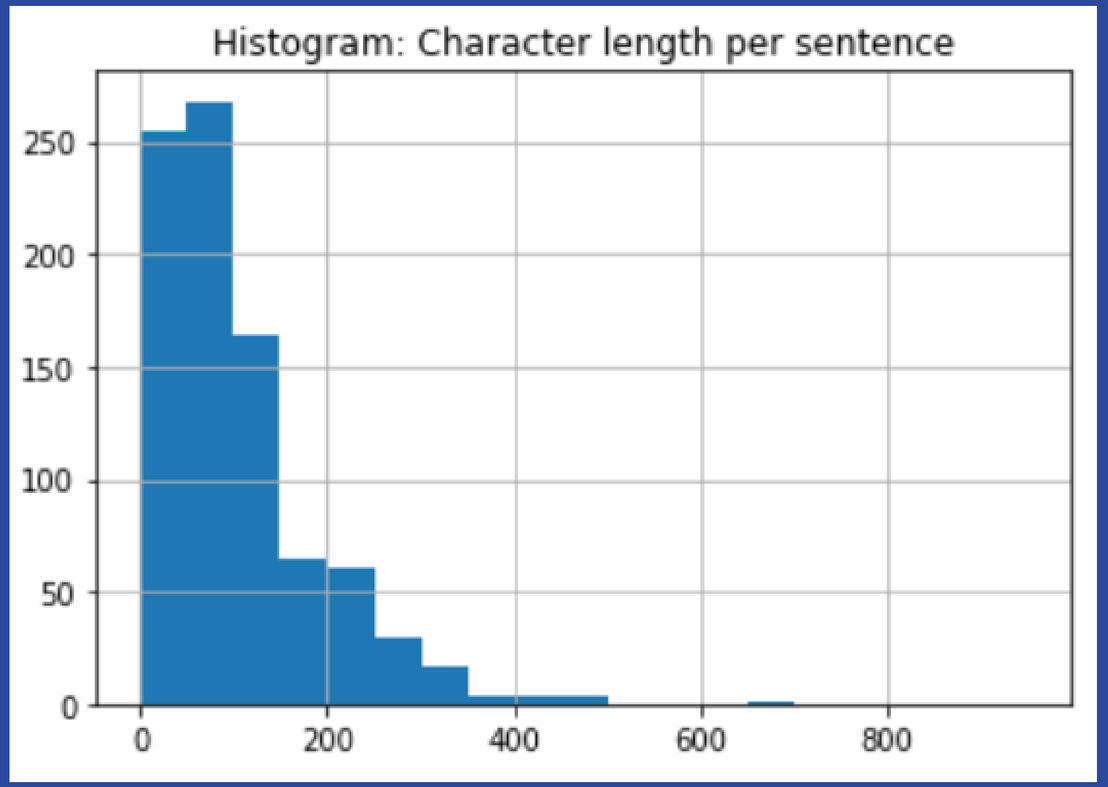
Selain itu, jumlah kalimat dengan dua label adalah 232 kalimat, dengan tiga label 111 kalimat dengan 4 label 48 kalimat, dengan 5 label 15 kalimat, dengan 6 dan 7 label masing-masing 3 kalimat dan dengan 8 dan 9 label masing-masing 1 kalimat

Sementara itu, **kalimat yang tidak terlabeli** sama sekali berjumlah **21 kalimat**



ANALISIS STATISTIK

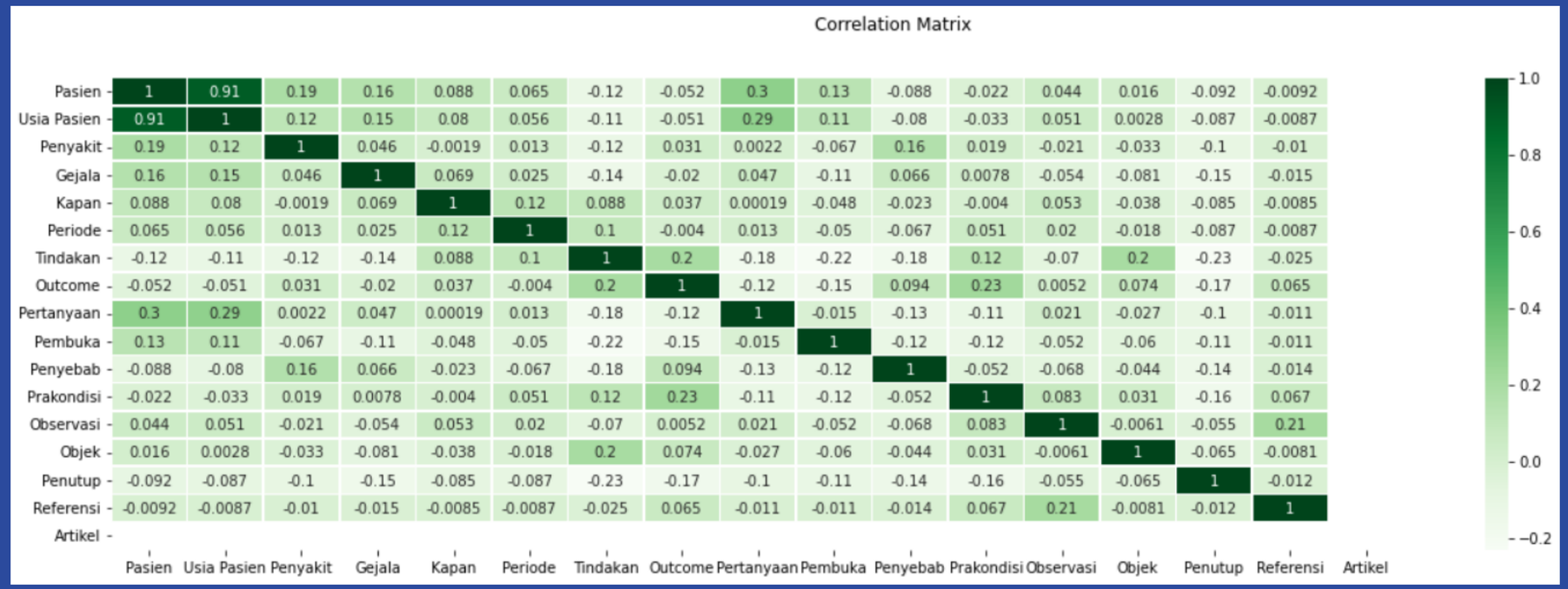
Grafik di samping merupakan disribusi dari panjang kalimat pada korpus. Berdasarkan grafik tersebut, diketahui bahwa kebanyakan kalimat memiliki **panjang kurang dari 200 karakter**



ANALISIS STATISTIK

Gambar tersebut menunjukkan wordcloud dari lima label yang memiliki kalimat terbanyak.

ANALISIS STATISTIK



Gambar tesebut menunjukkan matriks korelasi anatr 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’.

METODOLOGI

Task yang akan dilakukan

Multilabelled text classification, yaitu mengklasifikasikan label-label di dalam sebuah teks

1

Pembersihan teks

1. Mengubah semua string menjadi lowercase
2. Mengubah semua kata menjadi kata yang sesuai dengan KBBI
3. Menghapus special character, termasuk tanda baca (e.g. titik, koma, tanda petik dan lain-lain)
4. Mengatur semua whitespace menjadi 1 spasi

2

Seleksi Label

Tahapan ini akan mempertimbangkan **label ‘Artikel’ (0 kalimat) dan label ‘Referensi’ (1 kalimat) untuk tidak digunakan** dalam membangun model karena kemunculannya yang sangat jarang.

METODOLOGI

3

Pembagian Data Latih dan Data Tes

Dataset akan 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.

4

Transformasi Data

Tahapan ini akan mentransformasi dataset melalui proses *stemming* dan ekstrasi fitur kata. *Stemming* dilakukan dengan StemmerFactory dari *library* Sastrawi. Sedangkan, ekstraksi fitur kata dilakukan dengan dua skenario, yaitu **vektorisasi biner dan TF-IDF.**

METODOLOGI

5

Pembangunan Model

Pada tahapan ini, model klasifikasi multilabel akan dikembangkan dengan menggunakan 8 algoritma.

Pada saat membangun model, dilakukan pula, *hyperparameter tuning* , yaitu proses pengaturan *hyperparameter* dari suatu algoritme guna mengontrol algoritme tersebut agar dapat bekerja secara optimal.

Algoritme dilatih terhadap data latih menggunakan strategi klasifikasi multilabel yang disebut sebagai *one-vs-the-rest*.

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

METODOLOGI

Metrik Evaluasi Model

- **Akurasi** untuk melihat akurasi kelas secara keseluruhan
- **F1-score** untuk melihat akurasi berdasarkan *precision* maupun *recall*
- **Precision** untuk melihat ketepatan prediksi terhadap kelas positif
- **Recall** untuk melihat banyaknya kelas positif yang berhasil diprediksi

HASIL EKSPERIMEN

MODEL TERBAIK

Algorithm with Maximum F1-Score across all scenarios: LINEAR SVM

	Algorithm	Test size	Vectorization	F1-Score	Accuracy	Precision	Recall
9	LinSVM	0.2	Binary	0.621350	0.919923	0.665164	0.643141

Algorithm with Maximum Accuracy across all scenarios: STACKING

	Algorithm	Test size	Vectorization	F1-Score	Accuracy	Precision	Recall
15	Stacking	0.2	Binary	0.597915	0.937548	0.776410	0.517948

Berdasarkan uji coba eksperimen yang paling unggul adalah **Linear SVM** maka didapatkan nilai **rata-rata F1-Score yang unggul sebesar 62.1%,** sedangkan **Stacking unggul** dari segi **accuracy sebesar 93.76%.**

HASIL EKSPERIMEN

SKENARIO TERBAIK UNTUK SETIAP JENIS MODEL

	Algorithm	Test size	Vectorization	F1-Score	Accuracy	Precision	Recall
9	LinSVM	0.2	Binary	0.621350	0.919923	0.665164	0.643141
10	LogReg	0.2	Binary	0.604719	0.905747	0.582342	0.682646
13	RF	0.2	Binary	0.599037	0.934483	0.790844	0.525419
12	DT	0.2	Binary	0.597940	0.920690	0.681636	0.577239
15	Stacking	0.2	Binary	0.597915	0.937548	0.776410	0.517948
22	XGB	0.1	TF-IDF	0.541917	0.927969	0.723016	0.470422
0	NB	0.3	Binary	0.463488	0.856450	0.373791	0.642515
19	KNN	0.1	TF-IDF	0.301190	0.912644	0.647222	0.238313

Tabel tersebut menunjukkan skenario terbaik berdasar **evaluasi f1-score** untuk setiap jenis model yang telah dibangun. **Lima model teratas** sama-sama menerapkan rasio pembagian data latih **80:20** dengan **vektorisasi biner**

HASIL EKSPERIMEN

HASIL PREDIKSI UNTUK TIAP LABEL MENGGUNAKAN MODEL TERBAIK

Tabel tersebut menunjukkan rerata skor yang didapat pada setiap label pada **model terbaik** yaitu algoritma linear SVM dengan rasio data latih dengan data uji sebesar 80:20 dan vektorisasi biner. Dari hasil tersebut, diperoleh bahwa **7 label dari 15 label telah mencapai nilai F1-score > 70%.** Meskipun, **label yang sering muncul, yaitu '*Outcome*' dan 'Prakondisi' masih belum diprediksi dengan maksimal** karena pola kata-katanya yang terlalu beragam.

Label	F1-Score	Accuracy	Precision	Recall
Penutup	1	1	1	1
Tindakan	0.852459	0.896552	0.838710	0.866667
Pasien	0.833333	0.977011	0.769231	0.909091
Usia Pasien	0.818182	0.977011	0.750000	0.900000
Gejala	0.734694	0.925287	0.642857	0.857143
Pembuka	0.702703	0.936782	0.650000	0.764706
Kapan	0.700000	0.965517	0.777778	0.636364
Observasi	0.666667	0.988506	1	0.500000
Penyebab	0.553191	0.879310	0.541667	0.565217
Pertanyaan	0.536585	0.890805	0.423077	0.733333
Outcome	0.531646	0.787356	0.567568	0.500000
Prakondisi	0.419355	0.793103	0.419355	0.419355
Penyakit	0.400000	0.896552	0.375000	0.428571
Objek	0.285714	0.942529	1	0.166667
Periode	0.285714	0.942529	0.222222	0.400000

KESIMPULAN

1

Model terbaik adalah model berbasis Linear SVM dengan rasio pembagian data 80:20 dan vektorisasi biner sehingga menghasilkan nilai rata-rata F1- score pada seluruh label mencapai 62,1%.

2

Berdasarkan prediksi dari model terbaik, diperoleh 7 label dari 15 label yang telah mencapai nilai F1-score lebih dari 70%, yakni meliputi label ‘Penutup’, ‘Tindakan’, ‘Pasien’, ‘Usia Pasien’, ‘Gejala’, ‘Pembuka’, dan ‘Kapan’.

3

Penelitian selanjutnya dapat membangun model klasifikasi teks multilabel yang juga mempertimbangkan makna serta konteks kata di dalam kalimat dengan menggunakan model bahasa berbasis *transfer learning*

REFERENSI

1. Aneu Yulianeu, 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)

2. Asyaul Wasiah, S. A. (2021). Pelatihan Perawatan Bayi Baru Lahir. Journal of Community Engagement in Health

3. 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.

4. 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.

5. 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.

REFERENSI

6 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.

7 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.

8 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.

9 Quafafou, A. B. (2015). PENGGABUNGAN KEPUTUSAN PADA KLASIFIKASI MULTI-LABEL. JUTI: Jurnal Ilmiah Teknologi Informasi.

10 Susanti, N. (2011). PERAN IBU MENYUSUI YANG BEKERJA DALAM PEMBERIAN ASI EKSKLUSIF BAGI BAYINYA. Jurnal Kesehatan dan Keadilan Gender.

11 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.

THANKS!

DO YOU HAVE ANY QUESTIONS?