Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023107082
Vol. 10, No. 5, Oktober 2023, hlm. 1095-1104 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

EASY DATA AUGMENTATION UNTUK DATA YANG *IMBALANCE* PADA KONSULTASI KESEHATAN DARING

Anisa Nur Azizah[1], Misbachul Falach Asy'ari[2], Ifnu Wisma Dwi Prastya[3], Diana Purwitasari[*4]

[1,2,3,4]Institut Teknologi Sepuluh Nopember, Surabaya
Email: [1]anisaazizah069@gmail.com, [2]misbachulfalach@gmail.com, [3]ifnuprastya@unugiri.ac.id, [4]diana@if.its.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 17 Februari 2023, diterima untuk diterbitkan: 03 Oktober 2023)

Abstrak

Pendekatan augmentasi teks sering digunakan untuk menangani *imbalance* data pada kasus klasifikasi teks, seperti teks Konsultasi Kesehatan Daring (KKD), yaitu alodokter.com. Teknik *oversampling* dapat mengatasi kondisi *skewed* terhadap kelas mayoritas. Namun, augmentasi teks dapat mengubah konten dan konteks teks karena kata-kata teks tambahan yang berlebihan. Penelitian kami menyelidiki algoritma *Easy Data Augmentation* (EDA), yang berbasis parafrase kalimat dalam teks KKD dengan menggunakan teknik *Synonym Replacement* (SR), *Random Insertion* (RI), *Random Swap* (RS), dan *Random Deletion* (RD). Kami menggunakan Tesaurus Bahasa Indonesia untuk mengubah sinonim di EDA dan melakukan percobaan pada parameter yang dibutuhkan oleh algoritma untuk mendapatkan hasil augmentasi teks yang optimal. Kemudian, percobaan menyelidiki proses augmentasi kami menggunakan pengklasifikasi *Random Forest, Naïve Bayes,* dan metode berbasis peningkatan seperti XGBoost dan ADABoost, yang menghasilkan peningkatan akurasi rata-rata sebesar 0,63. Hasil parameter EDA terbaik diperoleh dengan menambahkan nilai 0,1 pada semua teknik EDA mendapatkan 88,86% dan 88,44% untuk akurasi dan nilai *F1-score*. Kami juga memverifikasi hasil EDA dengan mengukur koherensi teks sebelum dan sesudah augmentasi menggunakan pemodelan topik *Latent Dirichlet Allocation* (LDA) untuk memastikan konsistensi topik. Proses EDA dengan RI memberikan koherensi yang lebih baik sebesar 0,55 dan dapat mendukung implementasi EDA untuk menangani *imbalance* data, yang pada akhirnya dapat meningkatkan kinerja klasifikasi.

Kata kunci: *Augmentasi Teks, EDA, Klasifikasi, Konsultasi Kesehatan Daring*

EASY DATA AUGMENTATION FOR IMBALANCED DATA OF ONLINE HEALTH CONSULTATION TEXTS

Abstract

The text augmentation approach is often utilized for handling imbalanced data of classifying text corpus, such as online health consultation (OHC) texts, i.e., alodokter.com. The oversampling technique can overcome the skewed condition towards majority classes. However, text augmentation could change text content and context because of excessive words of additional texts. Our work investigates the Easy Data Augmentation (EDA) algorithm, which is sentence paraphrase-based in the OHC texts that often in non-formal sentences by using techniques of synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD). We employ the Indonesian thesaurus for changing synonyms in the EDA and do empirical experiments on parameters required by the algorithm to obtain optimal results of text augmentation. Then, the experiments investigate our augmentation process using classifiers of Random Forest, Naïve Bayes, and boosting-based methods like XGBoost and ADABoost, which resulted in an average accuracy increase of 0.63. The best EDA parameter results were acquired by adding a value of 0.1 in all EDA techniques to get 88.86% and 88.44% for accuracy and F1-score values. We also verified the EDA results by measuring coherences of texts before and after augmentation using a topic modeling of Latent Dirichlet Allocation (LDA) to ensure topic consistency. The EDA process with RI gave better coherences of 0.55, and it could support the EDA application to handle imbalanced data, eventually improving the classification performance.

Keywords: *Text Augmentation, EDA, Classification, Online Health Consultation*

1. PENDAHULUAN

Pada saat ini, informasi tentang kesehatan menjadi kebutuhan utama tiap individu. Banyak berbagai *platform-platform website* kesehatan yang muncul sebagai pemberi informasi kesehatan yang mudah dan dapat diakses dimana saja atau dengan nama lain adalah Konsultasi Kesehatan Daring (KKD). KKD tersebut tidak hanya memberikan informasi seputar kesehatan, namun juga menyediakan forum diskusi tanya jawab bagi pengguna secara online (Jin et al., 2022). KKD tersebut sangat memudahkan pengguna dalam mendapatkan informasi yang lebih spesifik mengenai penyakit atau keluhan kesehatan yang sedang mereka alami. Adanya forum tersebut juga mempersempit ruang pencarian informasi bagi pengguna (Abdillah et al., 2022). Dalam domain dunia medis, KKD diharapkan mampu memberikan informasi yang tepat sesuai dengan pertanyaan pengguna. Agar dapat memberikan informasi yang efektif dan tepat sesuai permasalahan pengguna, perlu adanya pengenalan topik pembahasan.

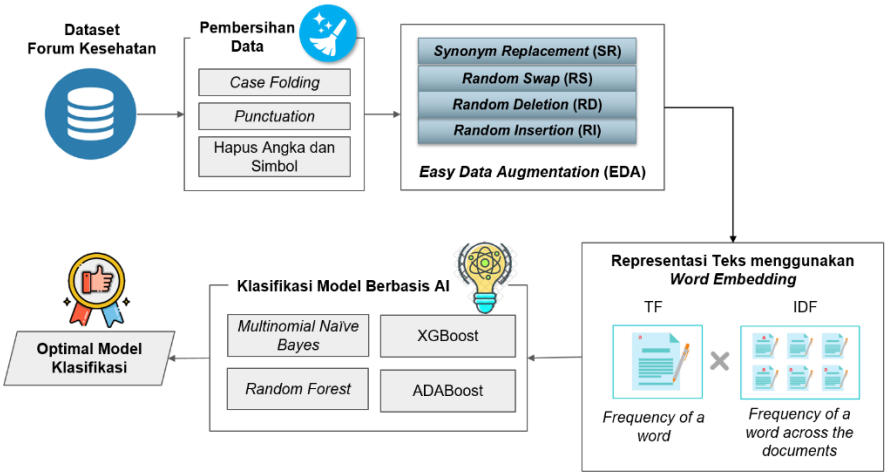
Kesalahan dalam memahami pertanyaan, menyebabkan kesalahan dalam memberikan jawaban yang tepat sehingga mengakibatkan kegagalan dalam diagnosis. Pemahaman yang baik antara suatu pertanyaan dan jawaban perlu adanya kesamaan topik suatu diagnosis penyakit maupun hal medis lainnya (Sarrouti and El Alaoui, 2020). Melalui pembacaan topik, memudahkan KKD dalam menentukan dokter ataupun tenaga ahli yang mampu menjawab dan berdiskusi pada forum tersebut. Masalahnya, tidak mudah dalam mengategorikan pertanyaan yang setiap hari mengalami penambahan. Oleh sebab itu, dibutuhkan sistem cerdas yang mampu mengategorikan pertanyaan dengan cepat, sesuai dengan label jawaban yang sesuai.

Pemilihan metode klasifikasi yang tepat juga sangat berpengaruh pada kinerja sistem klasifikasi. Saat ini metode-metode klasifikasi telah banyak dikembangkan dan beragam konsep pembelajaran. Kecerdasan buatan merupakan konsep pembelajaran dalam menyelesaikan permasalahan klasifikasi yang banyak digunakan (Sohail, Siddiqui and Ali, 2019). Kecerdasan buatan meniru konsep jaringan saraf yang dapat belajar secara terus-menerus hingga mendapatkan hasil yang optimal. Hal ini telah terbukti dapat meningkatkan hasil klasifikasi secara signifikan dibandingkan konsep peramalan statistika biasa (Thangaraj and Sivakami, 2018). Pada kecerdasan buatan memiliki beberapa metode klasifikasi seperti *Naïve Bayes*, *Decision Tree*, *Random Forest*, XGBoost, dan lain-lain. Selain menentukan metode terbaik dalam klasifikasi, pengetahuan tentang dataset juga penting. Salah satu penyebab kurang optimal sistem klasifikasi adalah ketersediaan jumlah kelas yang tidak seimbang atau *imbalance*.

Penanganan data *imbalance* pada suatu data teks dapat dilakukan dengan cara augmentasi data. Augmentasi merupakan suatu teknik *oversampling* atau memperbanyak jumlah data yang banyak digunakan untuk meningkatkan kinerja sistem klasifikasi teks. Terdapat 3 jenis augmentasi teks yaitu berbasis *noise*, parafrase, dan *sampling* (Li, Hou and Che, 2022). Augmentasi berbasis *sampling* merupakan pembuatan data teks baru yang didasarkan pada nilai distribusi data dan titik sampel pada data asli (Longpre et al., 2019). Augmentasi *noise* yaitu pembuatan data baru dengan menyelipkan kata atau simbol, *swapping* antar kata, penyisipan kata, dan penggantian kata pada sebuah data teks asli (Li, Hou and Che, 2022). Sedangkan augmentasi parafrase adalah pembuatan data serupa berdasarkan pengubahan kalimat namun tetap memperhatikan isi konteks pembicaraan pada suatu kalimat (Karimi, Rossi and Prati, 2021). Penelitian ini menggunakan data KKD yang sangat memperhatikan konteks bacaan sehingga melakukan augmentasi data berbasis parafrase.

Salah satu teknik augmentasi berbasis parafrase adalah *Easy Data Augmentation* (EDA). Terdapat 4 teknik pada penggunaan EDA yaitu *Synonym Replacement* (SR), *Random Insertion* (RI), *Random Swap* (RS), dan *Random Deletion* (RD) (Wei and Zou, 2019). Selain menangani *small dataset*, EDA juga dapat meningkatkan kinerja klasifikasi (He and Yang, 2021). Pada penelitian (Maslej-Kresnakova, Sarnovsky and Jackova, 2022) melakukan penerapan EDA dalam mendeteksi perilaku anti-sosial. Hasil EDA mendapatkan F1-*score* meningkat sebesar 2% dibandingkan data asli. Selanjutnya penelitian (Issifu and Ganiz, 2021) juga menghasilkan nilai evaluasi *F1-score* paling tinggi adalah 87,55%, dimana diperoleh dengan menerapkan augmentasi teks EDA dalam domain medis menggunakan *Named Entity Recognition* (NER). Penerapan EDA mampu meningkatkan kinerja model klasifikasi pada beberapa penelitian, namun belum ada yang mengukur bagaimana kecocokan hasil data EDA terhadap data asli. Pada penelitian ini menggunakan data KKD, dimana topik pembahasan sangat berpengaruh terhadap kedua pihak, pengguna maupun dokter yang secara online menjawab pertanyaan. Oleh sebab itu, penelitian ini menambahkan pengukuran secara koherensi untuk memvalidasi apakah data hasil augmentasi EDA tidak mengubah topik pembahasan dari data asli. Penelitian ini menggunakan pemodelan topik *Latent Dirichlet Allocation* (LDA) untuk mengukur tingkat koherensi tersebut.

Berdasarkan hasil studi literatur, penelitian ini melakukan klasifikasi data KKD, yaitu alodokter.com menggunakan proses augmentasi data untuk menangani data *imbalance.*



Gambar 1. Alur Penelitian Klasifikasi Data KKD

Pada penelitian ini bertujuan untuk mengklasifikasikan topik pembahasan pada domain kesehatan dengan melakukan augmentasi data yang bertujuan menyeimbangkan jumlah kelas data (topik kesehatan). Penelitian ini juga mengkomparasi beberapa metode kecerdasan buatan seperti *Multinomial Naïve Bayes*, *Random Forest*, AdaBoost, dan XGBoost untuk mendapatkan hasil klasifikasi topik yang optimal.

2. METODE PENELITIAN

Tahapan metode penelitian ini dimulai dari pemilihan data, pembersihan data, penerapan EDA, pra-pemrosesan data, *word embedding*, dan terakhir adalah proses klasifikasi. Langkah-langkah pengerjaan penelitian ini dapat dilihat pada Gambar 1.

2.1. Data Teks Konsultasi Kesehatan Daring

Data asli berupa teks Konsultasi Kesehatan Daring (KKD) di Indonesia yaitu dari alodokter.com pada tahun 2015 sampai 2021 (Abdillah et al., 2022). Penelitian ini menggunakan 10 kelas berupa topik penyakit atau keluhan yang paling banyak ditanyakan oleh pengguna. Total data berjumlah 89.569 data yang merupakan jawaban dari para dokter berdasarkan 10 topik teratas yang paling sering ditanyakan oleh pengguna. Tabel 1 menunjukkan contoh data tiap topik dan jumlah datanya.

2.2. Pembersihan Data

Pembersihan data merupakan tahapan yang penting sebelum dilakukan proses EDA. Langkah pertama adalah melakukan *case folding*. Proses ini menyamaratakan huruf kapital dalam kata menjadi huruf kecil (*lower case*). Proses ini penting untuk proses selanjutnya ketika mencari sinonim suatu kata. Selanjutnya dilakukan proses *punctuation* untuk menghilangkan semua tanda baca yang ada pada kalimat. Proses terakhir adalah menghapus karakter-karakter spesial yang tidak dibutuhkan, seperti angka dan simbol lainnya. Hal ini dilakukan karena tanda baca dan karakter spesial tersebut tidak diperlukan dalam proses selanjutnya.

2.3. *Easy Data Augmentation* (EDA)

EDA merupakan salah satu metode augmentasi teks yang dapat digunakan untuk menangani kelas data yang *imbalance*. Penelitian ini menggunakan empat teknik untuk menambahkan suatu kata dalam data. Teknik pertama adalah *Synonym Replacement* (SR) untuk mengganti suatu kata dengan kata sinonimnya. Kata sinonim yang digunakan berasal dari *library* Python Tesaurus dalam format json. Tesaurus Bahasa Indonesia berasal dari Pusat Bahasa karya Departemen Pendidikan Nasional pada tahun 2008 (Ibad, Soepriyanto and Husna, 2018). Teknik kedua merupakan *Random Insertion* (RI) yang digunakan untuk menyisipkan suatu kata yang ada dalam data secara acak pada kalimat. Teknik ketiga adalah *Random Swap* (RS) yang digunakan untuk menukar urutan suatu kata secara acak pada suatu kalimat. Teknik keempat merupakan *Random Deletion* (RD), yaitu teknik yang berfungsi untuk menghapus suatu kata secara acak pada suatu kalimat.

Untuk menentukan berapa jumlah kata yang disinonimkan, disisipkan, ditukar, atau dihapus dalam satu datum ditentukan oleh parameter α. Nilai Parameter α ini menentukan jumlah perubahan kata sebanyak n kali dalam kalimat l. Secara matematis, jumlah n dapat ditentukan sebagai berikut.

$$n = \alpha \times l \quad (1)$$

Dimana semakin tinggi nilai α, maka semakin banyak perubahan terjadi pada suatu kalimat. Penelitian ini menguj beberapa nilai parameter α pada keempat teknik EDA.

Sebelum melakukan keempat teknik EDA, data terlebih dahulu dibagi menjadi data latih dan data uji, seperti yang ditunjukkan pada Tabel 1. Teknik augmentasi EDA yang digunakan merupakan teknik *oversampling*, dimana kelas pada topik/kelas dengan data tertinggi akan menjadi acuan dalam augmentasi data.

Tabel 1. Contoh Dataset Pada Tiap Label Kelas

Topik	Jawaban	Data Latih	Data Uji	Jumlah Data
kehamilan	Terima kasih telah bertanya di Alodokter.com … (diperpendek) … kesehatan ibu dan janin. Semoga bermanfaat. dr. Previyanti	12073	3018	15091
menstruasi	Alo Hikmatus Sa'diyah, Perdarahan … (diperpendek) … hasil yang positif. Demikian penjelasan dari saya, semoga bermanfaat.	10530	2568	13098
obat	Alo, terimakasih sudah bertanya … (diperpendek) … konsumsi banyak air putih, yaitu lebih dari 2 L sehari. Demikian semoga bermanfaat.	10336	2615	12951
bayi	Terima kasih Catur Wie atas pertanyaannya … (diperpendek) … untuk memastikan kondisi yang dialami saat ini. Sekian dari saya, semoga dapat membantu dr. Christian Chandra	7214	1800	9014
intim-wanita	Hai. Labia merupakan bibir dari vagina. … (diperpendek) … anda baca mengenai Operasi vagina semoga bermanfaat. Terimakasih	6233	1589	7822
gangguan-pencernaan	Halo, Nyeri perut bawah bisa disebabkan akibat … (diperpendek) … segera konsultasikan kondisi Anda ke dokter. Semoga membantu. dr. Yusi	5668	1415	7083
kontrasepsi	Halo, Hali. Baiklah akan saya jelaskan. KB Suntikan 3 … (diperpendek) … setelah 3 bulan atau lebih penghentian KB. Semoga bermanfaat dr. Eni Yulvia S	5603	1390	6993
asam-lambung	Halo Akhoirul,Penyakit asam lambung atau GERD … (diperpendek) … Jangan minum alkohol dan soda Simak juga diskusi berikut ini.Semoga bermanfaat,Dr. Yosephine	4750	1141	5891
infeksi-saluran-kemih	Hai, Nyeri perut bawah bisa disebabkan oleh: Infeksi saluran kemih … (diperpendek) … banyak makan buah dan sayuran. Sekian, semoga bermanfaat. dr. Radius Kusuma	4731	1151	5882
intim-laki-laki	Selamat malam. Sebelum nya perlu ditegaskan … (diperpendek) … BPH Sekian informasi dari saya, semoga bermanfaat. terima kasih. dr.Anthony	4512	1232	5744

Sebagai contoh, topik 'menstruasi' akan diaugmentasi menjadi n=2 dan sisa augmentasinya akan dibuang hingga total data pada topik ini sama dengan total data pada topik tertinggi (kehamilan). Begitu juga dengan topik yang lainnya akan diaugmentasi dengan teknik yang sama hingga jumlahnya menjadi n=12073 data. Keempat teknik EDA dilakukan secara acak pada suatu kata di dalam suatu kalimat dengan mengecualikan kata *stop word* yang berasal dari *library* Python Sastrawi bahasa Indonesia.

2.4. *Word Embedding* Menggunakan TF-IDF

Data berupa teks harus direpresentasikan menjadi fitur numerik agar dapat diproses oleh sistem klasifikasi. Representasi teks kedalam fitur numerik ini biasanya dikenal sebagai *word embedding*. Terdapat beberapa metode *word embedding*, diantaranya adalah *Term Frequency - Inverse Document Frequency* (TF-IDF) seperti yang digunakan pada penelitian ini. TF-IDF adalah hasil kali dari frekuensi kemunculan *term* t pada dokumen d dibagi dengan total *term* pada dokumen d dan *Inverse Document Frequency* (IDF). bentuk persamaan dapat diitulis Persamaan (2) sebagai berikut.

$$tfIDF = \frac{f_d(t)}{max f_d(t)} \times \log\left(\frac{N}{df(t)}\right) \quad (2)$$

Dimana d adalah kumpulan kata atau kalimat, N adalah jumlah korpus, dimana korpus adalah total kumpulan kalimat, dan df adalah jumlah d dengan *term* t. Sehingga TF-IDF lebih terfokus pada kata-kata yang sering banyak muncul yang dinilai sebagai ciri dari dokumen tersebut (Lubis et al., 2021).

Eksperimen ini menggunakan fitur maksimum sebanyak 15.000. Artinya kosakata yang dibuat fhanya mempertimbangkan fitur teratas yang diurutkan berdasarkan frekuensi *term* di seluruh korpus. Selain itu, parameter lain yang digunakan adalah N_{gram} yang menentukan batas bawah dan atas rentang nilai N untuk N_{gram} berbeda yang akan diekstraksi. Unigram dan bigram digunakan pada penelitian ini. Dimana unigram adalah sebuah kata berurutan dalam sebuah kalimat, sedangkan bigram adalah dua kata berurutan dalam sebuah kalimat.

2.5. Klasifikasi Teks Konsultasi Kesehatan Daring

Penelitian ini melakukan uji coba metode klasifikasi untuk mendapatkan model terbaik. Keempat metode klasifikasi yang digunakan adalah sebagai berikut.

a. *Multinomial Naïve Bayes*

Metode ini merupakan perkembangan dari algoritma *Naïve Bayes* untuk data yang terdistribusi secara *multinomial*. Metode ini menerapkan *conditional probability* yang dilakukan tanpa memperhitungkan urutan kata secara umum. Metode ini memperhitungkan jumlah kata yang muncul.

$$C_{MAP} = \arg\max P(c) \prod_{k=1}^{m} P(t_k|c) \quad (3)$$

Parameter $P(c)$ adalah *prior* probabilitas, dimana dihitung dari peluang kemunculan kelas c pada seluruh pengamatan. Nilai probabilitas dihitung dengan rumus $P(c) = \frac{N_k}{N'}$, dimana N_k adalah jumlah kemunculan t_k dalam kelas c dan N adalah jumlah total datalatih. Sedangkan parameter $P(t_k|c)$

merupakan *conditional probability,* dimana menghitung estimasi probabilitas *likelihood* pada kejadian t_k disetiap kelas c menggunakan *Laplacean prior.*

$$P(t_k|c) = \frac{1 + T_k}{|V| + T'} \quad (4)$$

Dimana T_k adalah jumlah kemunculan t_k dalam data c, $|V|$ adalah total kata unik pada keseluruhan kelas, dan T adalah jumlah total muncul kata dalam c (Lubis et al., 2021).

b. *Random Forest*

Metode ini merupakan bagian dari banyak algoritma *Decision Tree.* Metode klasifikasi *Random Forest* dihasilkan dari proses *Bagging ensemble* untuk meningkatkan model klasifikasi. Setiap *Decision Tree akan* terdiri dari *decision node, leaf node*, dan *root node* (Khan et al., 2021). Hasil akhir model dihitung berdasarkan voting terbanyak menggunakan perhitungan *majority-voting system* (Sun et al., 2020).

c. XGBoost

XGBoost adalah kependekan dari "*eXtreme Gradient Boosting*". "*eXtreme*" mengacu pada peningkatan kecepatan dibandingkan *Gradient Boosting* tradisional. Selain itu, XGBoost menyertakan algoritma penemuan terpisah yang unik, untuk mengoptimalkan *tree*, bersama dengan regularisasi bawaan yang mengurangi *overfitting*. XGBoost mengoptimalkan fungsi objektif Persamaan (5).

$$L^{(t)} = \sum_{i=1}^{n} l\left(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(t-1)} + \mathrm{f_t}(\mathrm{x_i})\right) + \Omega(\mathrm{f}_t) \quad (5)$$

Dimana L adalah *loss function,* y adalah data aktual, $\hat{y}$ adalah data hasil prediksi, dan parameter Ω merupakan parameter regularisasi yang akan membuat model berusaha menghindari *overfitting* (Chen and Guestrin, 2016).

d. AdaBoost

Adaptive Boosting atau AdaBoost adalah model klasifikasi yang kuat dan sesuai untuk konstruksi adaptif. Hal ini menunjukkan bahwa semua pengklasifikasi lain diubah menjadi lebih lemah jika terjadi kesalahan klasifikasi dengan ambang batas pengklasifikasi yang digunakan sebelumnya. Pada setiap iterasi, contoh data pelatihan diberi bobot ulang sesuai dengan efisiensi klasifikasi. Untuk pengklasifikasi yang lemah, bobot dihitung berdasarkan keakuratan klasifikasi. Bobot yang diberikan digunakan untuk *voting* pengklasifikasi. Jika tingkat kesalahan rendah, bobot lebih diberikan pada pengklasifikasikasinya dan proses ini diulangi. Bobot pengklasifikasi yang memilih objek khusus ini disertakan. Dan kelas yang mendapatkan bobot total tertinggi itulah yang dijadikan kelas prediktif untuk objek spesifik tersebut. Kami menggunakan parameter $n_{estimator}$ =100. Parameter ini adalah jumlah maksimum estimator di mana peningkatan dihentikan. Jika cocok, prosedur *training* dihentikan lebih awal (Thiyagarajan and Shanthi, 2019).

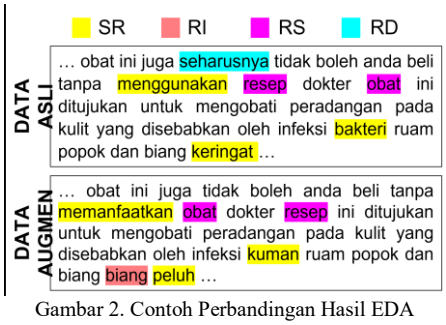
2.6. Koherensi

Latent Dirichlet Allocation (LDA) adalah suatu model yang bertujuan untuk mengungkap struktur tematik laten atau yang tersembunyi dari suatu korpus. Penelitian ini menggunakan LDA untuk membuat model data latih asli (tanpa augmentasi). Model tersebut digunakan untuk menghitung nilai koherensi hasil augmentasi. Data latih asli terlebih dahulu diubah menjadi korpus menggunakan TF-IDF kemudian dijadikan parameter model LDA dengan topik berjumlah 10. Hasil model LDA ini dijadikan acuan untuk menghitung nilai koherensi hasil augmentasi masing-masing teknik EDA.

Nilai koherensi dihitung berdasarkan 4 tahapan. (1) Segmentasi menggunakan metode *S-one-set*, yaitu, ukuran korpus akan dihitung atas pasangan kata yang sama yang ada dalam data asli dan data EDA. (2) Perhitungan probabilitas dihitung melalui jendela geser berukuran sesuai dimensi korpus (100) yang bergerak di atas teks. (3) Ukuran konfirmasi menggunakan ukuran konfirmasi tidak langsung. Kata-kata dari setiap elemen pasangan kata dibandingkan dengan semua kata lainnya dari himpunan korpus. Skor akhir adalah *cosinus* kesamaan antara dua vektor ukuran. (4) Agregasi koherensi akhir adalah rata-rata aritmatika dari langkah-langkah konfirmasi. Hasilnya akan didapatkan nilai koherensi tiap hasil augmentasi terhadap data aslinya (Syed and Spruit, 2017).

3. HASIL DAN PEMBAHASAN

Penelitian ini melakukan klasifikasi topik kesehatan yang bertujuan untuk mempermudah pengguna platform kesehehatan dalam memahami topik penyakit. Total data berjumlah 89.569 dengan 10 topik kesehatan. Tahap awal pemprosesan adalah pembersihan data dengan melakukan *case folding, punctuiation,* dan penghapusan angka dan simbol. Setelah itu, data dibagi menjadi dua bagian yaitu data *training* dan *testing* dengan jumlah data masing-masing adalah 71.655 dan 17.914 data.



Gambar 2. Contoh Perbandingan Hasil EDA

3.1 Penerapan EDA

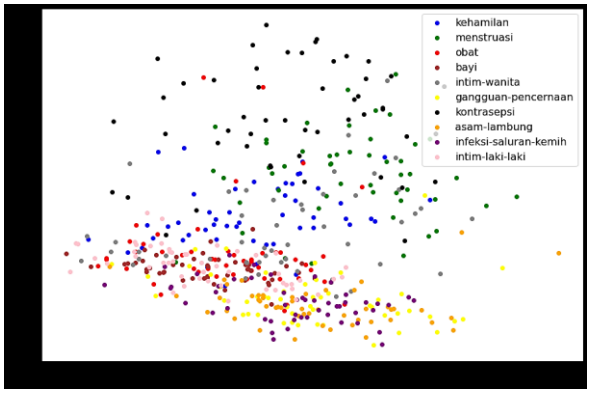
Penelitian ini mengusulkan penggunaan augmentasi EDA dalam menangani *imbalance class* data tanya jawab kesehatan. Proses augmentasi dilakukan setelah pembagian data training dan testing sebanyak 80:20. Augmentasi dilakukan pada data training dengan jumlah kelas acuan untuk dilakukan *over sampling* adalah 12.073. Pada metode EDA terdapat beberapa *hyperparameter* yaitu tingkat SR, RI, RS, dan RD. *hyperparameter* tersebut menentukan jumlah perubahaan pada tiap kalimat. Pada penelitian ini nilai a_SR, a_RI, a_RS, dan a_RD dilakukan uji coba untuk mendapatkan hasil augmentasi paling baik. Perubahan tiap kata dihitung menggunakan Persamaan (1) dengan parameter sebesar 0,4 pada tiap konsep dan 0,1 pada seluruh konsep perubahan. Selanjutnya melakukan inisialisasi jumlah augmentasi tiap kalimat data (n_{aug}) dengan nilai antara 1 atau 2 disesuaikan dengan jumlah data asli untuk mencapai *balance* data. Kelas acuan adalah kelas 0 atau kelas kehamilan dengan jumlah sampel data terbanyak.

Sampel pengubahan data augmentasi EDA dapat dilihat pada Gambar 2. Berdasarkan Gambar 2, SR ditunjukkan dengan mengubah kata "menggunakan" menjadi "memanfaatkan", mengubah kata "bakteri" menadi "kuman", dan kata "keringat" menjadi "peluh". Ketiga kata tersebut diganti dengan sinonim yang tepat sehingga tidak mengubah konteks kalimat. RI ditunjukkan dengan menambahkan kata "biang" sama dengan kata sebelumnya. RS ditunjukkan dengan mengubah urutan dari kata "resep" dan "obat". RD ditunjukkan dengan menghilangkan kata "seharusnya" pada kalimat.

3.2 Hasil Representasi Teks

Setiap beda uji coba dataset dilakukan representasi teks menggunakan hasil matriks TF-IDF. Hasil matriks TF-IDF menunjukkan frekuensi kemunculan *term-term* pada tiap dokumen. Dokumen-dokumen pada tiap topik kesehatan dihitung nilai TF-IDF dan hasil representasi *word embedding* tiap topik ditunjukkan pada Gambar 3 representasi tersebut merupakan hasil dari reduksi matriks TF-IDF menjadi 2 fitur (x,y) menggunakan *TruncatedSVD function*.

Berdasarkan Gambar 3 diatas menunjukkan bahwa *term-term* pada topik gangguan pencernaan, asam lambung, dan infeksi saluran kemih memiliki kelompok *term* yang sama atau mirip. Kemudian sebagian besar *term* pada topik obat banyak pada 2 topik intim laki-laki dan bayi. Kelompok *term* lain ada pada topik kehamilan, intim wanita, dan mentruasi. Selanjutnya kelompok yang memiliki sebaran paling luas adalah topik kontrasepsi dan ditambah beberapa *term* topik menstruasi. Hasil *scatter plot* data tiap topik kesehatan menunjukkan kemiripan data yang dibuktikan dari kelompok *term-term* TF-IDF.



Gambar 3. Contoh Perbandingan Hasil EDA

3.3 Hasil Komparasi Metode Klasifikasi

Hasil augmentasi data pelatihan digunakan dalam pelatihan model dengan mengkomparasi 4 metode yaitu *Multinomial Naïve Bayes, Random Forest,* XGBoost, dan AdaBoost untuk mendapatkan model klasifikasi paling optimal. Hasil keseluruhan uji coba dapat dilihat pada Tabel 2. Evaluasi model klasifikasi membandingkan nilai akurasi dan F1 score. Nilai akurasi model dapat menentukan seberapa besar model dapat memprediksi kebenaran label kelas pada seluruh data. Sedangkan F1-*score* berdasarkan presisi dan *recall* dari suatu model klasifikasi. Sedangkan *Negative Predictive Value* (NPV) menunjukkan kinerja model dalam memprediksi data yang tidak termasuk pada kelas tersebut.

Berdasarkan Tabel 2, hasil paling baik adalah model klasifikasi *Random Forest* dengan menerapkan uji coba nilai 0,1 diseluruh teknik EDA. Akurasi terbaik adalah 88,86%, presisi 88,55%, *recall* 88,86%, *F1 Score* 88,44%, spesifisitas 98,75%, dan NPV sebesar 98,75%. Selisih terbesar antara hasil data augmentasi dan data asli adalah pada model klasifikasi AdaBoost dengan selisih akurasi 2,85%. Selisih nilai tersebut lebih baik pada data augmentasi dengan 0,4 teknik RD.

Penelitian ini melakukan uji coba pada empat metode klasifikasi yaitu ADABoost, *Multinomial Naïve Bayes, Random Forest*, dan XGBoost. Hasil komparasi antar metode dapat dilihat pada Gambar 4. Hasil terbaik dalam mengklasifikasikan topik kesehatan adalah *Random Forest* dengan rata-rata akurasi sebesar 88,57% dan rata-rata *F1 Score* adalah 88,12%. Nilai akurasi menunjukkan tingkat model dalam memprediksi data yang benar, sedangkan *F1 Score* menunjukkan tingkat model dalam meminimalkan kesalahan dalam prediksi. Dengan kata lain metode *Random Forest* baik dalam mengklasifikasikan topik pada data tanya-jawab kesehatan. Sedangkan model paling rendah adalah metode ADABoost dengan rata-rata akurasi sebesar 69,28% dan rata-rata *F1 Score* adalah 68,89%.

Tabel 2. Hasil Uji Coba Klasifikasi Data Tanya Jawab Kesehatan

Metode Klasifikasi	Parameter Augmentasi EDA				Akurasi (%)	Presisi (%)	*Recall* (%)	F1-Score (%)	Spesifisitas (%)	NPV (%)
	SR	RI	RS	RD						
XGBoost	0,4	0	0	0	79,22	79.17	78.85	78,85	97.65	97.66
	0	0,4	0	0	79,25	79.25	78.74	78,79	97.65	97.67
	0	0	0,4	0	80,09	79.4	80.6	79,86	97.76	97.75
	0	0	0	0,4	79,99	79.39	80.45	79,8	97.75	97.74
	0,1	0,1	0,1	0,1	79,83	79.09	80.53	79,61	97.74	97.72
	Tanpa Augmentasi				**80,51**	**80.59**	**80.18**	**80,24**	**97.79**	**97.81**
ADABoost	0,4	0	0	0	70,62	70.19	70.19	69,42	96.68	96.71
	0	0,4	0	0	67,49	67.48	69.07	67,21	96.37	96.37
	0	0	0,4	0	69,34	68.45	70.8	68,95	96.56	96.56
	0	0	0	0,4	**71,37**	**70.92**	**73.03**	**71,07**	**96.8**	**96.79**
	0,1	0,1	0,1	0,1	68,34	68.48	71.02	68,3	96.48	96.47
	Tanpa Augmentasi				68,52	68.53	70.59	68,37	96.48	96.48
Random Forest	0,4	0	0	0	88,32	88.52	87.55	87,84	98.68	98.7
	0	0,4	0	0	88,53	88.64	87.95	88,12	98.7	98.72
	0	0	0,4	0	88,6	88.26	88.14	88,07	98.72	98.72
	0	0	0	0,4	88,74	88.3	88.54	88,27	98.73	98.74
	0,1	0,1	0,1	0,1	**88,86**	**88.55**	**88.68**	**88,44**	**98.75**	**98.75**
	Tanpa Augmentasi				88,39	88.77	87.55	87,96	98.68	98.7
Naïve Bayes	0,4	0	0	0	72,75	77.13	69.8	71,88	96.85	96.97
	0	0,4	0	0	75,77	76.16	75.25	75,41	97.25	97.28
	0	0	0,4	0	75,92	75.54	76.02	75,61	97.28	97.29
	0	0	0	0,4	75,96	75.2	76.59	75,72	97.3	97.29
	0,1	0,1	0,1	0,1	75,96	75.14	76.59	75,69	97.3	97.29
	Tanpa Augmentasi				**76,13**	**75.93**	**76.13**	**75,89**	**97.3**	**97.31**

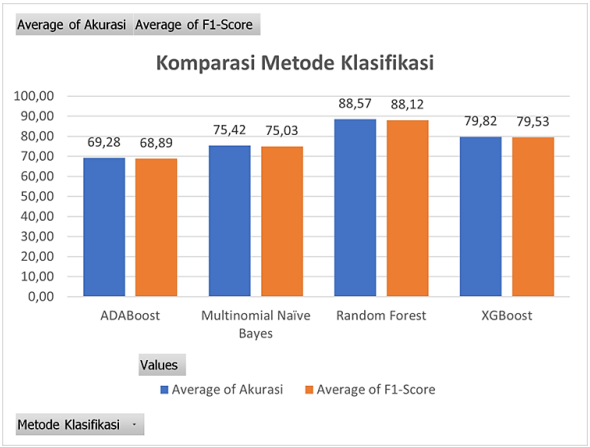
Tabel 3. Hasil Evaluasi Tiap Kelas Klasifikasi Data Tanya Jawab Kesehatan

Metode Klasifikasi	Uji Coba	Kelas 0	Kelas 1	Kelas 2	Kelas 3	Kelas 4	Kelas 5	Kelas 6	Kelas 7	Kelas 8	Kelas 9
AdaBoost	ORI	58,98	55,27	74,32	94,62	38,24	49,47	88,56	85,14	80,70	80,59
	EDA	62,76	62,93	71,31	94,17	40,03	49,26	86,48	81,71	77,07	82,47
Multinomial Naïve Bayes	ORI	74,78	73,74	81,58	93,51	53,96	58,36	78,06	79,63	81,63	86,07
	EDA	75,07	74,50	82,05	93,01	50,38	56,57	75,71	77,62	79,74	83,85
Random Forest	ORI	89,23	87,71	97,03	97,67	75,00	72,76	90,14	87,78	89,37	88,77
	EDA	87,27	87,05	96,68	97,68	77,72	73,18	92,04	89,64	91,67	88,77
XGBoost	ORI	82,24	76,95	84,98	95,62	62,28	64,15	85,70	80,48	86,14	83,29
	EDA	79,72	75,27	82,01	95,47	63,41	63,05	86,99	82,07	86,50	83,86
Rata-rata		76,26	74,18	83,75	95,22	57,63	60,85	85,46	83,01	84,10	84,71

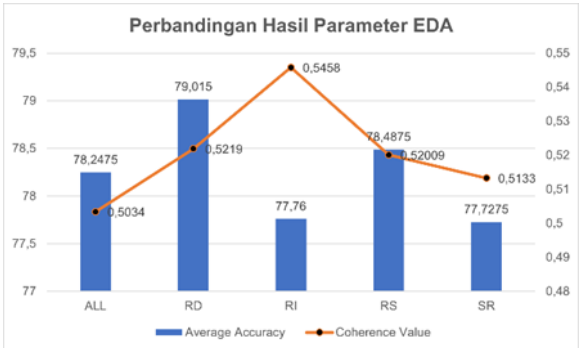
3.4 Penentuan Parameter EDA Terbaik

Penentuan besar parameter disetiap teknik EDA didasarkan uji coba untuk menunjukkan teknik EDA paling berpengaruh baik pada model klasifikasi. Setiap teknik EDA diuji dengan nilai 0,4 dengan menghiraukan teknik lain (diberi nilai 0). Sedangkan berdasarkan rekomendasi metode EDA dengan 0,1 diseluruh teknik EDA. Hasil augmentasi data diukur menggunakan koherensi dengan metode *Latent Dirichlet Allocation* (LDA). Pengukuran dilakukan dengan mempertimbangkan *score* ketepatan topik pada data asli dan data hasil augmentasi. Hasil analisis perbandingan evaluasi teknik EDA dengan nilai koherensi dapat dilihat pada Gambar 5.

Pada percobaan parameter terbaik adalah RI dengan nilai koherensi sebesar 0,546. Penyisipan atau memasukkan sinonim kata pada teks tidak merubah konteks topik pada suatu kalimat. Berdasarkan hasil tersebut memberikan kesimpulan bahwa perubahan RI



Gambar 4. Grafik Perbandingan Hasil Uji Coba Parameter EDA



Gambar 5. Perbandingan Nilai Koherensi dan Rata-rata Evaluasi Tiap Teknik EDA

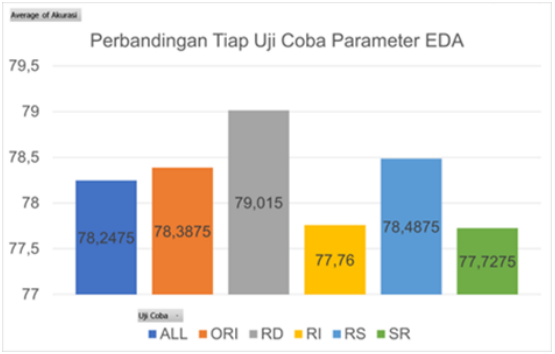
dapat memberikan hasil kualitas data yang baik dalam proses augmentasi data teks. Berdasarkan perbandingan antara hasil evaluasi dan nilai koherensi, hasil sangat mengejutkan bahwa teknik RI mendapatkan nilai tertinggi koherensi namun menghasilkan nilai akurasi paling kecil pada klasifikasi. Sedangkan hasil rata-rata akurasi aling baik adalah teknik RD dengan selisih 1,26% lebih baik dibandingkan penggunakan teknik RI. Hal in menunjukkan bahwa RI penyisipan kata sinonim pada kalimat belum baik dibandingkan RD. sedangkan teknik RD dapat dikatakan baik dalam merepresentasikan data hasil augmentasi. Penggunaan teknik RD atau penghapusan kata secara random dapat mengurangi kompleksitas data. Mengingat bahwa data tanya jawab kesehatan memiliki panjang kata yang banyak. Teknik RD dapat mengoptimalkan kinerja sistem klasifikasi dan meringankan kerja komputasi pada proses training

Tabel 4. Perbandingan Kelas Mayor dan Minor

Metode Klasifikasi	Uji Coba	ORI (%)	EDA (%)	Selisih (%)
AdaBoost	Kelas Mayor	58,98	62,76	3,78
	Kelas Minor	80,59	82,47	1,88
Multinomial Naïve Bayes	Kelas Mayor	74,78	75,07	0,28
	Kelas Minor	86,07	83,85	-2,23
Random Forest	Kelas Mayor	89,23	87,27	-1,96
	Kelas Minor	88,77	88,77	0,00
XGBoost	Kelas Mayor	82,24	79,72	-2,52
	Kelas Minor	83,29	83,86	0,57

model. Dapat disimpulkan bahwa augmentasi data teks dengan teknik RD mendapatkan hasil yang baik dalam klasifikasi data tanya jawab kesehatan.

Berdasarkan uji coba teknik EDA, hasil terbaik adalah dengan penggunaan teknik RD sebesar 0,4 dan teknik lain sama dengan 0. Nilai rata-rata akurasi teknik RD sebesar 79,02% dan selisih 0,63% dengan model data asli (tanpa EDA), dapat dilihat pada Gambar 6. Teknik RD dapat dengan baik bekerja pada model klasifikasi karena dapat mengurangi variasi kata pada setiap data. Pada penelitian ini data tanya jawab kesehatan memiliki paling banyak sekitar 1.900 jumlah kata pada data sehingga teknik RD dapat dengan baik mengurangi kompleksitas model. Sedangkan penerapan hanya teknik RI menghasilkan nilai paling kecil dengan rata-rata



Gambar 6. Perbandingan Rata-rata Evaluasi Tiap Uji Coba

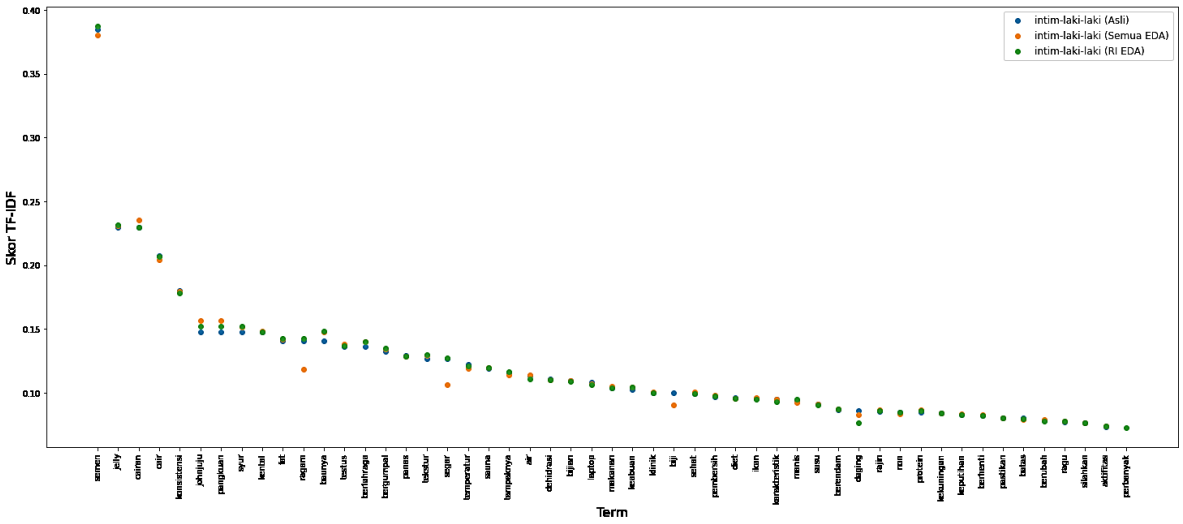
akurasi 77,76%. Teknik RI memiliki kinerja kurang baik disebabkan penyisipan kata sinonim mungkin dapat memberikan penekanan pada kata-kata tertentu yang sama.

Berdasarkan perbedaan nilai akurasi, hasil penggunaan semua parameter (all) mendapatkan nilai paling tinggi dibandingkan hanya menggunakan salah satu parameter EDA saja. Visualisasi antar skor TF-IDF dan *term* pada beda data (data asli, data dengan semua parameter EDA, dan data dengan parameter RI) ditunjukkan pada Gambar 7. Grafik tersebut mengambil 50 *term* tertinggi pada topik intim laki-laki, dimana topik tersebut adalah kelas minor yang menghasilkan paling banyak data baru. Berdasarkan grafik tersebut menunjukkan bahwa beberapa kata pada beda uji data memiliki skor TF-IDF yang berbeda. Sangat terlihat perbedaan tersebut antara hasil penggunaan semua parameter dan hanya menggunakan RI terlihat pada *term* "fat" dan "segar" bahwa skor TF-IDF semua parameter lebih kecil dibandingkan RI. Hal tersebut bisa terjadi karena adanya penggunaan RD, RS, dan SR yang menjadikan skor TI-IDF pada *term* tersebut menurun. Sedangkan pada *term* "biji" tidak termasuk 50 *term* tertinggi pada data dengan penggunaan parameter RI saja.

3.4 Analisa Hasil

Selanjutnya melakukan analisis tentang hasil kinerja klasifikasi data tanya jawab kesehatan pada tiap label kelas topik. Data ini memiliki 10 label kelas topik. Hasil perbandingan akurasi pada tiap label kelas dapat dilihat pada Tabel 3. diketahui bahwa label kelas 5 atau "gangguan pencernaan" memiliki akurasi yang paling rendah dibandingkan label kelas lainnya. Pada kelas gangguan pencernaan banyak terjadi kesalahan akibat topik pembahasan dan isi dari keluhan hampir sama dengan kelas lainnya seperti kelas Menstruasi, Asam-lambung, atau yang lainnya. Dapat dibuktikan pada sampel data Tabel 1, pada data Gangguan-pencernaan data sebagian besar membahas tentang nyeri perut sedangkan sakit perut juga dapat dialami oleh label kelas Menstruasi. Sedangkan nilai rata-rata hasil evaluasi paling baik adalah pada kelas Bayi dengan nilai 95,22%. Hal tersebut karena data kelas Bayi sangat spesifik dan berbeda dengan kelask-kelas lainnya.

Selanjutnya mengukur kinerja hasil augmentasi pada data mayor (kelas 0 atau Kehamilan) dan kelas minor (kelas 9 atau intim laki-laki), dapat dilihat pada Tabel 4. Berdasarkan tabel diatas menunjukkan bahwa hasil augmentasi data memiliki pengaruh yang cukup besar pada model klasifikasi. Pada keseluruhan metode klasifikasi, hasil data augmentasi mengalami kenaikan yang paling signifikan 3,78% pada kelas mayor. Sedangkan hampir seluruh kelas minor mendapatkan peningkatan setelah data augmentasi dengan nilai selisih paling signifikan adalah 1,88%. Dapat disimpulkan bahwa hasil augmentasi data memiliki pengaruh yang baik dalam klasifikasi topik



Gambar 7. Grafik 50 *Term* Skor TF-IDF Tertinggi Pada Topik Minoritas – Intim Laki-Laki

kesehatan, namun peningkatan yang kurang signifikan menjadi motivasi pada penlitian selanjutnya untuk memperbaiki atau memofikasi model pembelajaran untuk mendapatkan hasil lebih baik lagi. Salah satu permasalahan terbesar pada penelitian ini adalah data yang terlalu besar dan kompleks sehingga disarankan untuk mereduksi data teks agar model dapat mempelajari pola data dengan lebih baik lagi.

4. KESIMPULAN

Pada penelitian ini melakukan klasifikasi topik kesehatan berdasarkan data tanya jawab forum diskusi platform alodokter. Pada penelitian ini mengusulkan penerapan *Easy Data-Augmentasi* (EDA) untuk menangani *imbalance class* dan meningkatkan kinerja model klasifikasi. Tahap awal adalah pembersihan data, pengujian data augmentasi, representasi data teks menggunakan TF-IDF, sampai uji coba 4 metode klasifikasi.

Hasil klasifikasi paling baik adalah pada model *Random Forest* dengan nilai seluruh teknik EDA sebesar 0,1 mendapatkan nilai akurasi terbaik adalah 88,86% dan *F1 Score* adalah 88,44%. Nilai rata-rata akurasi teknik RD sebesar 79,02% dan selisih 0,63% dengan model data asli (tanpa augmentasi). Berdasarkan pengukuran kualitas data augmentasi koherensi model, teknik terbaik adalah RI dengan nilai koherensi sebesar 0,546. Hasil keseluruhan uji coba mendapatkan kesimpulan bahwa penerapan augmentasi data EDA berpengaruh baik pada model klasifikasi dan dapat meningkatkan kinerja model pada kelas minor.

DAFTAR PUSTAKA

ABDILLAH, A.F., PUTRA, C.B.P., APRIANTONI, A., JUANITA, S. dan PURWITASARI, D., 2022. Ensemble-based Methods for Multi-label Classification on Biomedical Question-Answer Data. Journal of Information Systems Engineering and Business Intelligence, 8(1), pp.42–50.

CHEN, T. dan GUESTRIN, C., 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. [online] New York, NY, USA: Association for Computing Machinery. pp.785–794.

HE, H. dan YANG, H., 2021. Deep visual semantic embedding with text data augmentation and word embedding initialization. Mathematical Problems in Engineering, 2021.

IBAD, A.Z., SOEPRIYANTO, Y. dan HUSNA, A., 2018. Thesaurus Termediasikan Augmented Reality Text Untuk Peningkatan Pemahaman Baca. JKTP: Jurnal Kajian Teknologi Pendidikan, 1(1), pp.1–6.

ISSIFU, A.M. dan GANIZ, M.C., 2021. A simple data augmentation method to improve the performance of named entity recognition models in medical domain. In: 2021 6th International Conference on Computer Science and Engineering (UBMK). IEEE. pp.763–768.

JIN, Q., YUAN, Z., XIONG, G., YU, Q., YING, H., TAN, C., CHEN, M., HUANG, S., LIU, X. dan YU, S., 2022. Biomedical question answering: A survey of approaches and challenges. ACM Computing Surveys (CSUR), 55(2), pp.1–36.

KARIMI, A., ROSSI, L. dan PRATI, A., 2021. Aeda: An easier data augmentation technique for text classification. arXiv preprint arXiv:2108.13230.

KHAN, M.Y., QAYOOM, A., NIZAMI, M.S., SIDDIQUI, M.S., WASI, S. dan RAAZI, S.M.K.-R., 2021. Automated Prediction of Good Dictionary EXamples (GDEX): A

Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. Complexity, 2021.

LI, B., HOU, Y. dan CHE, W., 2022. Data augmentation approaches in natural language processing: A survey. AI Open.

LONGPRE, S., LU, Y., TU, Z. dan DUBOIS, C., 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. arXiv preprint arXiv:1912.02145.

LUBIS, A.R., NASUTION, M.K.M., SITOMPUL, O.S. dan ZAMZAMI, E.M., 2021. The effect of the TF-IDF algorithm in times series in forecasting word on social media. Indones. J. Electr. Eng. Comput. Sci., 22(2), p.976.

MASLEJ-KRESNAKOVA, V., SARNOVSKY, M. dan JACKOVA, J., 2022. Use of Data Augmentation Techniques in Detection of Antisocial Behavior Using Deep Learning Methods. Future Internet, 14(9), p.260.

SARROUTI, M. dan EL ALAOUI, S.O., 2020. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artificial intelligence in medicine, 102, p.101767.

SOHAIL, S.S., SIDDIQUI, J. dan ALI, R., 2019. A comprehensive approach for the evaluation of recommender systems using implicit feedback. International Journal of Information Technology, 11(3), pp.549–567.

SUN, Y., LI, Y., ZENG, Q. dan BIAN, Y., 2020. Application research of text classification based on random forest algorithm. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). IEEE. pp.370–374.

SYED, S. dan SPRUIT, M., 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp.165–174.

THANGARAJ, M. dan SIVAKAMI, M., 2018. Text classification techniques: A literature review. Interdisciplinary Journal of Information, Knowledge, and Management, 13, p.117.

THIYAGARAJAN, D. dan SHANTHI, N., 2019. A modified multi objective heuristic for effective feature selection in text classification. Cluster Computing, [online] 22(5), pp.10625–10635.

WEI, J. dan ZOU, K., 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.