Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023107325
Vol. 10, No. 6, Desember 2023, hlm. 1329-1340 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENERAPAN *TEXT AUGMENTATION* UNTUK MENGATASI DATA YANG TIDAK SEIMBANG PADA KLASIFIKASI TEKS BERBAHASA INDONESIA STUDI KASUS: DETEKSI JUDUL *CLICKBAIT* DAN KOMENTAR *HATE SPEECH* PADA BERITA *ONLINE*

Iftitah Athiyyah Rahma[1], Lya Hulliyyatus Suadaa[*2]

[1,2]Politeknik Statistika STIS, Jakarta Timur
Email: [1]221910989@stis.ac.id, [2]lya@stis.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 30 Mei 2023, diterima untuk diterbitkan: 27 November 2023)

Abstrak

Klasifikasi teks merupakan salah satu tugas yang fundamental dalam *natural language processing* (NLP). Dalam dunia nyata, data dan sumber daya yang tersedia untuk pengklasifikasian teks terbatas. Salah satu kendala pada data berlabel yang digunakan adalah *imbalanced data* atau data yang tidak seimbang. Permasalahan data yang tidak seimbang memengaruhi kinerja dan keakuratan model karena model hanya terfokus pada data dengan label mayoritas. Sementara itu, data berlabel minoritas cenderung diklasifikasikan tidak tepat oleh model, padahal untuk beberapa kasus kemampuan model untuk memprediksi data dengan label minoritas lebih penting. Untuk mengatasinya, penelitian ini melakukan pendekatan *oversampling* yaitu menambah data untuk menyeimbangkan *dataset*. Penerapan *oversampling* pada data teks dikenal dengan *text augmentation*. Pada penelitian ini dilakukan dua teknik *text augmentation* yaitu *synonym replacement* dan *back translation* pada beberapa kondisi ketidakseimbangan dan skenario augmentasi terhadap dua *dataset*. Berdasarkan hasil eksperimen, augmentasi mampu meningkatkan skor F1 label minoritas sekitar 0,1% hingga 8%. Pada seluruh kasus, augmentasi lebih signifikan pada *dataset* kecil. Hasil dari teknik *back translation* mampu meningkatkan skor F1 hingga 8% sehingga lebih baik dibandingkan dengan teknik *synonym replacement* yang meningkat hanya sampai 5%. Selain itu, hasil penelitian menunjukkan bahwa skenario jumlah augmentasi juga berpengaruh terhadap kenaikan skor F1. Semakin banyak jumlah data augmentasi belum tentu memberikan hasil yang semakin baik karena terindikasi *overfitting* pada data latih. Kata-kata yang tidak normal atau tidak baku pada *dataset* teks informal memengaruhi proses augmentasi sehingga hasil teks sintetis yang diperoleh tidak sebaik pada *dataset* teks formal.

Kata kunci: *data yang tidak seimbang, klasifikasi teks, text augmentation, synonym replacement, back translation*

APPLICATION OF TEXT AUGMENTATION TO HANDLING IMBALANCED DATA PROBLEM IN INDONESIAN TEXT CLASSIFICATION CASE STUDY: CLICKBAIT HEADLINE AND HATE SPEECH COMMENT DETECTION IN ONLINE NEWS

Abstract

Text classification is one of the fundamental tasks in natural language processing (NLP). However, data and resources for text classification are limited in actual application. One of the constraints on the dataset for text classification is imbalanced data, or the condition when one label has more data than the others. Imbalanced data affects the performance and accuracy of the model because the model only focuses on the majority label data. Meanwhile, the minority label data tends to be classified incorrectly by the model, even though, in some cases, the model's ability to predict data with minority labels is more important. To solve this problem, this research uses an oversampling approach to augment data and balance the dataset. The application of oversampling text data is known as text augmentation. This research uses two text augmentation techniques, synonym replacement and back translation, applied to several imbalance conditions and augmentation scenarios for two datasets. Based on experimental results, augmentation can increase the F1 score of the minority class about 0,1% to 8%. In all cases, augmentation is more significant in small datasets. The results of back translation were able to increase F1 score by up to 8%, making it better than synonym replacement which increased only up to 5%. In addition, this study's results show that the number of augmentation affects an increase in F1 score. Increasing the augmentation data

cannot ensure the results are getting better. Furthermore, words that are not normal in informal text datasets affect the augmentation process, so the results of synthetic text are better than the formal text dataset.

Keywords: *imbalanced data, text classification, text augmentation, synonym replacement, back translation*

1. PENDAHULUAN

Aktivitas manusia yang terdigitalisasi menghasilkan jejak digital dan juga data (Kurnia et al., 2021). Data yang dihasilkan tersebut merupakan *big data* yang memiliki karakteristik *volume*, *velocity*, *variety*, *veracity*, dan *value* (Gudivada et al., 2015). Potensi *big data* sebagai salah satu sumber informasi membuat para ahli dan peneliti semakin gencar melakukan penelitian untuk meningkatkan kualitas *big data*. Salah satu jenis *big data* yang memiliki jumlah sangat besar dan tersedia di mana saja adalah data teks. Pengolahan data teks dilakukan dengan *natural language processing* (NLP) yang memproses teks secara otomatis untuk menemukan pengetahuan di dalamnya seperti yang dilakukan manusia (Chowdhary et al., 2020).

Salah satu *task* yang fundamental dalam NLP adalah *text classification* atau klasifikasi teks yang implementasinya umum digunakan untuk mangkategorikan dokumen. Beberapa aplikasi klasifikasi teks diantaranya yaitu penyaringan *email* spam, analisis sentimen, dan deteksi *hate speech* (Razno, 2019). Tugas klasifikasi teks dilakukan untuk mengkategorikan teks ke dalam dua atau lebih kategori atau label. Untuk melakukan klasifikasi, model perlu dilatih terlebih dahulu sehingga dibutuhkan *dataset* yang mendukung proses pelatihan. Namun, data yang tersedia di dunia nyata sering kali tidak sempurna sehingga membutuhkan berbagai pemrosesan tambahan sebelum dielaborasi. Salah satu keterbatasan dan permasalahan terkait *dataset* dalam klasifikasi teks adalah data yang tidak seimbang atau *imbalanced data*. Pada data yang tidak seimbang, terjadi situasi di mana banyak *instance* dengan label tertentu jauh lebih sedikit dibandingkan dengan banyak *instance* berlabel lainnya (Ali et al., 2015; Sutoyo & Fadlurrahman, 2020; Verdikha et al., 2018).

Kondisi data yang tidak seimbang membuat distribusi data sampel untuk pelatihan menjadi tidak seimbang sehingga klasifikasi akan cenderung mengabaikan kelas/label yang jumlahnya sedikit dan fokus pada kelas/label dengan *instance* besar. Hal tersebut akan memengaruhi kinerja model klasifikasi. Nilai akurasi yang tinggi pun tidak dapat dijadikan gambaran performa model karena model sebagian besar hanya mengklasifikasi benar untuk kelas mayoritas (Verdikha et al., 2018; Gu et al., 2016). Padahal, untuk beberapa kasus, kemampuan untuk memprediksi kelas minoritas jauh lebih penting. Contohnya yaitu pada deteksi spam, *fraud*, dan *hate speech*, di mana teks dengan label tersebut jumlahnya lebih sedikit pada data yang tersedia.

Untuk mengatasi permasalahan data yang tidak seimbang tersebut terdapat dua solusi pendekatan, yaitu solusi pendekatan internal dan solusi pendekatan eksternal (Estabrooks et al., 2004). Pendekatan internal dilakukan dengan membuat suatu algoritma baru atau memodifikasi algoritma yang sudah ada untuk menurunkan tingkat kesalahan (*error rate*). Pendekatan internal bergantung pada algoritma klasifikasi yang digunakan sehingga untuk beberapa kasus, pendekatan internal efektif dalam menangani data yang tidak seimbang tetapi menjadikan algoritma tersebut lebih spesifik sehingga menimbulkan masalah jika diterapkan pada *dataset* yang berbeda. Sebaliknya, solusi pendekatan eksternal tidak bergantung pada algoritma klasifikasi. Oleh karena itu, pendekatan eksternal akan lebih serba guna dibandingkan dengan pendekatan internal.

Pendekatan eksternal dilakukan dengan melakukan *resampling* atau mengubah jumlah sampel pada *dataset* untuk kemudian diproses dalam algoritma atau model klasifikasi. Solusi pendekatan eksternal memiliki dua pendekatan *resampling* yang berbeda, yaitu *undersampling* dan *oversampling*. Pendekatan *oversampling* berarti menambah sampel kelas minoritas untuk membuatnya mendekati ukuran kelas mayoritas, sedangkan pendekatan *undersampling* berarti mengurangi sampel dari kelas mayoritas untuk membuatnya mendekati ukuran kelas minoritas (Estabrooks et al., 2004). Teknik *oversampling* yang umum digunakan seperti *random oversampling* dan SMOTE mengabaikan distribusi data dan menyebabkan *overfitting* (Wang et al., 2012). SMOTE banyak digunakan untuk menangani data yang tidak seimbang pada kasus klasifikasi teks. Hal tersebut dibuktikan melalui penelitian yang menyebutkan *oversampling* dengan SMOTE lebih baik dibandingkan *random oversampling* (Sanya & Suadaa, 2022).

Namun, dalam beberapa skenario aplikasi pada dunia nyata, SMOTE dapat memperoleh hasil yang kurang baik atau kontraproduktif pada banyak kasus (Saez et al., 2015; Barua et al., 2014). SMOTE memiliki beberapa kelemahan yaitu menghasilkan sampel yang tidak informatif, menghasilkan *noisy sample* atau sampel yang tidak berarti sehingga menyebabkan masalah *overgeneralization*, dan meningkatkan *overlapping* antarlabel di sekitar batas label karena *oversampling* dilakukan secara membabi buta (Soltanzadeh & Hashemzadeh, 2021). Pemilihan tetangga yang acak menimbulkan kebutaan yang kuat serta sulit menentukan jumlah tetangga yang dipilih dengan tepat menjadi keterbatasan penggunaan SMOTE (Jiang et al., 2021). Pada data teks, implementasinya dilakukan pada fitur teks (Rupapara

et al., 2021). Oleh karena itu, selain berbagai kelemahan yang disebutkan sebelumnya, data teks tambahan hasil SMOTE kurang baik secara bahasa dan konteks atau label data aslinya (Lu et al., 2021) padahal konteks kalimat penting dalam pemrosesan teks bahasa alami

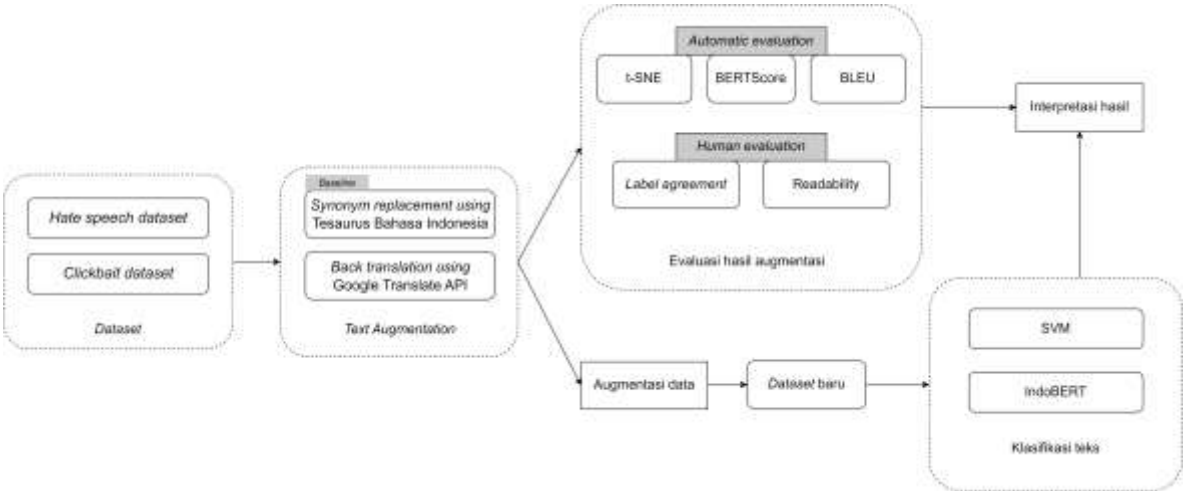
Oleh karena itu, diperlukan alternatif lain untuk penanganan data yang tidak seimbang. *Dataset* dari dunia nyata yang belum sempurna untuk pemodelan membutuhkan pemrosesan lebih lanjut untuk meningkatkan kualitasnya. Namun, saat ini terdapat keterbatasan pada sumber daya dan sistem NLP dalam melakukan hal tersebut. Sumber daya dan sistem NLP banyak tersedia untuk bahasa sumber daya tinggi (*high-resources languages*) seperti bahasa Inggris, Prancis, Spanyol, Jerman, dan Cina. Sementara itu, banyak bahasa yang memiliki sumber daya rendah (*low-resources languages*) seperti bahasa Bengali, Indonesia, Punjabi, Cebuano, dan Swahili. Padahal, bahasa dengan sumber daya rendah tersebut diucapkan dan ditulis oleh jutaan orang (Bock & Garnsey, 2008). Menurut Internet World Stats, Indonesia merupakan bahasa yang paling banyak digunakan keempat di internet dengan total pengguna sekitar 171 juta di seluruh dunia. Meskipun terdapat sejumlah besar data berbahasa Indonesia di internet, kemajuan penelitian NLP di Indonesia berjalan lambat (Cahyawijaya et al., 2021). Keterbatasan tersebut menimbulkan masalah karena ketersediaan *dataset* yang berkualitas memegang peranan penting dalam menentukan performa dan akurasi model. Pengembangan *dataset* sangat diperlukan, tetapi hal tersebut membutuhkan sumber daya yang besar. Salah satu alternatif untuk meningkatkan kualitas *dataset* berbentuk teks adalah dengan *text augmentation*. *Text augmentation* atau augmentasi data teks merupakan proses sekumpulan algoritma untuk menyusun teks sintetis dari sekumpulan data teks yang tersedia (Harywanto et al., 2022). Khususnya pada klasifikasi teks, tantangan utamanya adalah bagaimana menghasilkan teks baru tanpa memengaruhi label aslinya (Abdurrahman & Purwarianti, 2019).

Penerapan *text augmentation* masih sedikit dilakukan untuk kasus teks berbahasa Indonesia. *Text augmentation* dapat dilakukan dengan teknik *synonym replacement* menggunakan Tesaurus (Zhang & LeCun, 2015). Tesaurus yang digunakan dalam penelitian tersebut telah diurutkan sinonimnya berdasarkan kejadian yang umum/sering dalam bahasa. Pemilihan sinonim ditentukan pada probabilitas geometris. Dengan begitu, sinonim pertama lebih mungkin untuk dipilih. Jumlah kata yang diganti ditentukan juga oleh probabilitas geometris yang lebih cenderung mengganti sejumlah satu kata pada suatu kalimat. Penerapan penggantian sinonim juga dilakukan untuk meningkatkan performa klasifikasi pada *dataset* berbahasa Indonesia (Abdurrahman & Purwarianti, 2019). Metode yang digunakan adalah *synonym-based text augmentation*, yaitu mengganti satu atau beberapa kata dalam kalimat dengan sinonimnya. Penentuan jumlah kata yang diganti dihitung dengan mempertimbangkan panjang kalimat dan konstanta (*augmentation degree*). Kata pengganti terbaik dipilih menggunakan language model (LM) dengan *pre-trained word embedding* sebagai input dari *neural model*. Pengklasifikasian kalimat dilakukan menggunakan model *deep learning*. Hasil penelitian ini menunjukkan *5-gram neural* LM sebagai model terbaik. Selain itu, diperoleh peningkatan skor F1 dan akurasi masing-masing sebesar 2,91% dan 3,35%.

Penggantian sinonim untuk menghasilkan data teks baru dilakukan dengan penambahan *task* POS *tagging* (*part of speech tagging*) untuk mengetahui kelas kata sehingga sinonim yang diperoleh merupakan padanan yang tepat pada teks tersebut (Xiang et al., 2021; Jungiewicz & Smywinski-Pohl, 2019). Modifikasi POS *tagging* untuk mengidentifikasi kata yang akan diganti ditambah strategi tambahan untuk menemukan substitusi yang terkait secara semantik saat membuat teks baru dilakukan menggunakan POS *focused lexical substitution for data augmentation* (PLSDA) (Xiang et al., 2021). PLSDA signifikan meningkatkan kinerja algoritma analisis sentimen. Pada penelitian lain, POS *tagging* digunakan untuk membantu menemukan sinonim kata yang akan diganti (Jungiewicz & Smywinski-Pohl, 2019). Augmentasi dilakukan menggunakan Thesaurus serta WordNet, sedangkan klasifikasi teks dilakukan dengan pelatihan CNN. Diperoleh bahwa metode yang digunakan berhasil meningkatkan akurasi di sebagian besar kasus sebesar 1,2% lebih baik dibandingkan dengan model *baseline*.

Text augmentation telah banyak diterapkan pada kasus berbahasa Inggris karena didukung dengan sumber daya yang memadai. Contohnya penelitian oleh C. Coulombe menggunakan teknik augmentasi yang lebih *advance* dan praktis dengan memanfaatkan Cloud API NLP (Coulombe, 2018). Pada penelitian tersebut, digunakan SyntaxNet, API Google Translate, dan berbagai teknik augmentasi. Berdasarkan hasil eksperimen, akurasi meningkat sekitar 4,3% hingga 21,6% dari klasifikasi standar yang sederhana.

Berdasarkan latar belakang permasalahan dan kajian penelitian terdahulu, penelitian ini dilakukan untuk menjawab tiga tujuan, yaitu (1) melakukan augmentasi teks berbahasa Indonesia untuk *task* klasifikasi teks, (2) membandingkan hasil klasifikasi pada beberapa teknik augmentasi serta skenario augmentasi yang diterapkan untuk melihat pengaruh augmentasi, dan (3) mengevaluasi kualitas teks sintetis hasil augmentasi dari teknik yang diterapkan dan merekomendasikan teknik terbaik. Berdasarkan studi literatur dan sumber daya yang tersedia untuk bahasa Indonesia, penelitian ini menggunakan dua teknik augmentasi yaitu *synonym replacement* dan *back translation.* Hasil dari penelitian ini diharapkan



Gambar 1. Tahapan Penelitian

dapat menjadi masukan mengenai alternatif untuk penanganan data yang tidak seimbang terutama pada klasifikasi teks berbahasa Indonesia.

2. METODE PENELITIAN

2.1. Sumber Data

Penelitian ini menggunakan dua *dataset* dengan kriteria yang berbeda yaitu teks formal dan informal.

1) Data teks formal yang digunakan dalam penelitian ini adalah data judul berita *online* berbahasa Indonesia CLICK-ID *Dataset* (William & Sari, 2020). Data tersebut memiliki dua label yaitu *clickbait* dan *non-clickbait*. Total label *clickbait* sejumlah 3316 dan label *non-clickbait* sejumlah 5297.
2) Data teks informal yang digunakan merupakan komentar berita *online* berbahasa Indonesia (Sanya & Suadaa, 2022). Terdapat dua label pada teks yaitu *hate* dan *no hate.* Total 263 data berlabel *hate* dan 1.307 data berlabel *no hate.* Pada *dataset* ini telah tersedia kondisi tidak seimbang *slightly imbalanced* dan *imbalanced*

Untuk *dataset clickbait* dibangun kondisi tidak seimbang seperti *dataset hate speech*, yaitu *slightly imbalanced* dan *imbalanced*. Kondisi *slightly imbalanced* memiliki perbandingan label 1:2. Sementara *imbalanced* memiliki perbandingan 1:5.

2.2. Tahapan Penelitian

Secara umum, tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1. Penelitian dimulai dengan melakukan studi literatur untuk memperoleh *dataset* yang akan digunakan. Selanjutnya, *dataset* diaugmentasi dengan dua teknik yang berbeda. Teks hasil augmentasi dievaluasi menggunakan dua pendekatan yaitu *automatic evaluation* dan *human evaluation. Dataset* yang telah ditambahkan teks hasil augmentasi diklasifikasikan menggunakan dua pendekatan model klasifikasi. Hasil dari klasifikasi dan evaluasi diinterpretasikan untuk dapat menjawab tujuan penelitian.

2.2.1. *Text Augmentation*

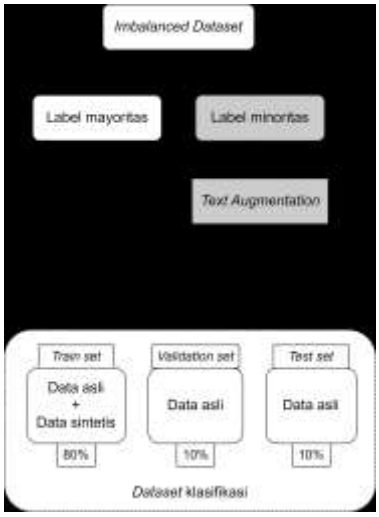
Proses augmentasi mengikuti skema pada Gambar 2. *Generate* teks dilakukan pada label minoritas saja pada *dataset* utama agar seimbang. Augmentasi dilakukan pada *train set* (Madukwe et al., 2022; Tesfagergish et al., 2021) di tiap kondisi yang menyesuaikan skenario augmentasi yaitu satu kali, dua kali, tiga kali, dan *balance.*

Teknik augmentasi yang digunakan yaitu *synonym replacement* dan *back translation*. Untuk kondisi *slightly imbalanced* yang memiliki perbandingan tidak seimbang 1:2, hanya dilakukan augmentasi sampai kondisinya seimbang yaitu kurang lebih satu kali augmentasi.

a. *Synonym Replacement*

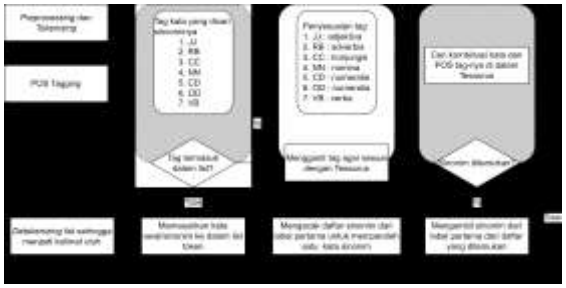
Teknik *synonym replacement* dilakukan dengan mengganti sebanyak n kata dengan sinonimnya. Sinonim yang digunakan diperoleh dari *website* Tesaurus Tematis Bahasa Indonesia oleh Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia (http://tesaurus.kemdikbud.go.id/tematis/). Akses ke *website* tersebut dibantu dengan sebuah modul yang dapat diimpor melalui Python.

Proses penggantian kata pada teknik *synonym replacement* dilakukan mengacu pada penelitian sebelumnya yang dimodifikasi untuk bahasa Indonesia (Jungiewicz & Smywinski-Pohl, 2019). Untuk mencari sinonim pada Tesaurus diperlukan informasi POS *tag* sehingga diketahui kedudukan kata tersebut dalam konteks kalimat. POS *tagging* dilakukan dengan model tagging CRF (Dinakaramani et al., 2014). Selain itu, dalam suatu kalimat, tidak semua kata diganti dengan sinonimnya. Mekanisme alur penggantian sinonim dapat dilihat pada Gambar 3. Terdapat dua kondisi agar kata tersebut dapat diganti. Kondisi pertama yaitu *tag* kata harus memiliki padanan di Tesaurus. Jika kondisi 1 terpenuhi, maka dilakukan pengecekan lanjutan ke kondisi 2 yaitu kombinasi *tag* dan kata ditemukan sinonimnya di Tesaurus.



Gambar 2. Skema Augmentasi Teks

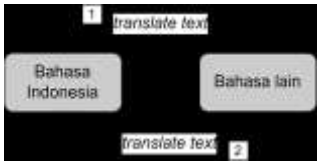
Jika salah satunya tidak terpenuhi maka kata tidak dapat diganti dengan sinonimnya. Jika sinonim yang ditemukan memiliki lebih dari satu label konteks, maka label konteks pertama akan dipilih. Hal tersebut dilakukan karena sebagian besar label konteks pertama mengandung konteks yang paling sesuai atau paling banyak digunakan. Kemudian, jika dalam satu konteks memiliki lebih dari satu sinonim maka dilakukan pengacakan. Pengacakan dilakukan dengan pertimbangan dalam satu label konteks, lema atau kata di dalamnya telah memiliki kemiripan atau *similarity* yang dekat. Pengacakan juga bertujuan untuk menghasilkan teks yang berbeda-beda ketika dilakukan lebih dari satu kali augmentasi.



Gambar 3. Mekanisme Alur Penggantian Sinonim Teknik *Synonym Replacement*

b. *Back Translation*

Back translation menghasilkan teks baru dengan menerjemahkan teks asli ke bahasa lain yang kemudian diterjemahkan kembali ke bahasa Indonesia. Ilustrasi konsep *back translation* dapat dilihat pada Gambar 4.



Gambar 4. Konsep Back Translation

Teknik augmentasi *back translation* dilakukan dengan API Google Translate yang diakses melalui Python. Proses augmentasi dilakukan sebanyak lima kali dengan bahasa yang berbeda yaitu Inggris, Cina, Melayu, Jawa, dan Tagalog. Pemilihan bahasa didasarkan pada intensitas penggunaan bahasa selain bahasa Indonesia oleh masyarakat Indonesia dan kedekatan bahasa. Selain itu, pemilihan bahasa Cina atau Mandarin didasarkan pada penemuan penelitian sebelumnya yang menyatakan hasil *back translation* bahasa Cina lebih beragam dan banyak perubahan dari kalimat aslinya (Natasya & Girsang, 2023). Oleh karena itu, penggunaan bahasa Cina akan mampu memperoleh teks baru yang berbeda dari teks asli sehingga diharapkan akan meningkatkan performa model dalam generalisasi.

2.2.2. Klasifikasi Teks

Klasifikasi teks dilakukan menggunakan SVM dan *pretrained transformer* IndoBERT-large (Cahyawijaya et al., 2021). Pengklasifikasian dilakukan untuk tiap kondisi ketidakseimbangan dan skenario augmentasi pada dua teknik augmentasi. Klasifikasi teks yang dilakukan juga menerapkan *5-fold cross validation* dan *grid search* untuk pencarian parameter terbaik.

2.2.3. Evaluasi Kualitas Teks Hasil Augmentasi

a. Plot t-SNE

Hasil t-SNE akan menghasilkan plot dua dimensi hasil ekstrak *output* yang berupa vektor yang merepresentasi kedekatan semantik dalam ruang laten antara teks asli dengan teks sintesis (Wei & Zou, 2019).

b. BLEU

BLEU mengukur aspek keragaman atau *diversiy* (Papineni et al., 2002). Skor BLEU yang lebih rendah menunjukkan augmentasi yang lebih baik karena mampu membuat teks yang lain dari teks asli (Feng *et al.*, 2020; Okimura *et al.*, 2022).

c. BERTScore

BERTScore menghitung aspek evaluasi *semantic content preservation* atau kelestarian semantik konten pada teks hasil augmentasi (Zhang et al., 2019). Semakin tinggi BERTScore, maka semantik konten antara keduanya semakin sesuai (Feng *et al.*, 2020; Okimura *et al.*, 2022).

d. *Human Evaluation*

Evaluasi teks augmentasi dengan *human evaluation* digunakan untuk menilai aspek kualitas teks yang tidak dapat diukur secara otomatis melalui komputasi. *Human evaluation* dilakukan pada sampel yang terpilih secara acak. Aspek yang dilihat yaitu *readability* atau keterbacaan dan *label agreement* atau kesesuaian label (Liu *et al.*, 2020; Pandey et al., 2021). Aspek keterbacaan menilai struktur bahasa atau apakah teks hasil augmentasi yang merupakan hasil dari sebuah program atau sebuah kecerdasan buatan mampu menyerupai teks dari manusia. Selain itu, aspek kesesuaian label menilai apakah

augmentasi yang dilakukan memengaruhi label teks yang diturunkan dari teks aslinya. Kesesuaian label sangat penting karena dapat memengaruhi kinerja model. Penilaian aspek keragaman dilakukan dengan menggunakan BWS (Best-Worst Scaling) dan dianalisis dengan rumus Maxdiff (Logit & Orme, 2009). Sementara itu, aspek kesesuaian label dilakukan dengan analisis proporsi yang menyatakan jumlah label teks sintetis yang sesuai dengan label teks aslinya dibandingkan keseluruhan sampel.

3. HASIL DAN PEMBAHASAN

3.1. Ringkasan *Dataset*

Kriteria *dataset* yang digunakan tertera pada Tabel 1. *Dataset hate comment* memiliki *instance* yang jauh lebih sedikit. Sementara itu, teks pada *dataset hate speech* lebih panjang dan bervariasi. Kedua *dataset* memiliki dua label/*binary.*

Tabel 1. Ringkasan Data

Dataset	*Instance*	Label	Kata Unik	Kata per Instance
Hate comment	1.570	*Binary*	6.973	20
Clickbait	8.613	*Binary*	13.398	9

3.2. *Text preprocessing*

Tahap *preprocessing* yang dilakukan untuk melakukan augmentasi disesuaikan dengan teknik augmentasi yang diterapkan.

3.2.1. *Proprocessing* teknik *synonym replacement*

Augmentasi dengan *synonym replacement* memerlukan beberapa tahapan *preprocessing* agar dapat diidentifikasi POS *tag* setiap katanya. *Preprocessing* yang dilakukan adalah *lower casing, removing punctuation, removing white space,* dan *tokenizing.*

3.2.2. *Preprocessing* teknik *back translation*

Augmentasi dengan *back translation* tidak memerlukan *text preprocessing* untuk dapat dijalankan. Oleh karena itu, teks pada *dataset* asli akan langsung diproses augmentasinya.

Teks kemudian dilakukan *preprocessing* kembali sebelum diklasifikasi. Tahap *preprocessing* yang dilakukan adalah *lower casing, removing punctuation,* dan *removing white space.* Untuk klasifikasi dengan model SVM, juga dilakukan *stemming* dan *stopwords removing.* Model IndoBERT telah dilatih pada korpus berbahasa Indonesia yang sangat besar sehingga dapat menyesuaikan konteks (Wilie *et al.*, 2020). Oleh karena itu, *stemming* dan *stopwords removal* tidak dilakukan karena dapat mengganggu konteks kalimat.

3.3. Klasifikasi Teks

Klasifikasi teks dilakukan menggunakan SVM dan IndoBERT. Klasifikasi tersebut diterapkan untuk semua kombinasi kondisi ketidakseimbangan dan skenario jumlah augmentasi pada kedua *dataset*. Untuk klasifikasi menggunakan SVM, dilakukan *hyperparameter tuning* dengan *grid search cross validation*, serta pemilihan kernel *linear* untuk menghindari *overfitting* pada data pelatihan asli (Han & Jiang, 2014) serta menyesuaikan kondisi *imbalanced* yang lebih sesuai menggunakan kernel *linear* (Sun et al., 2009). Untuk IndoBERT dilakukan pencarian *epoch* terbaik.

3.3.1. Kondisi *Slightly Imbalance*

Hasil klasifikasi tiap label tanpa augmentasi (WA), dengan augmentasi *synonym replacement* (SR), dan dengan augmentasi *back translation* (BT) dapat dilihat pada Tabel 2. Berdasarkan hasil eksperimen, skor F1 label minoritas pada kedua *dataset* meningkat dengan teknik *back translation*. Sementara augmentasi dengan teknik *synonym replacement* hanya mampu meningkatkan skor F1 pada *dataset hate speech*. *Dataset clickbait* memiliki ukuran *dataset* yang besar dan kriteria teks yang formal. Teks formal memiliki kata-kata yang baku dan struktur teks yang baik. Teks sintetis hasil *synonym replacement* yang belum cukup baik dapat berbahaya untuk pelatihan model (Kobayashi, 2018). Hal tersebut dapat menjadi penyebab penurunan skor F1 pada *dataset clickbait* teknik *synonym replacement*.

Tabel 2. Hasil Klasifikasi Kondisi *Slightly Imbalance*

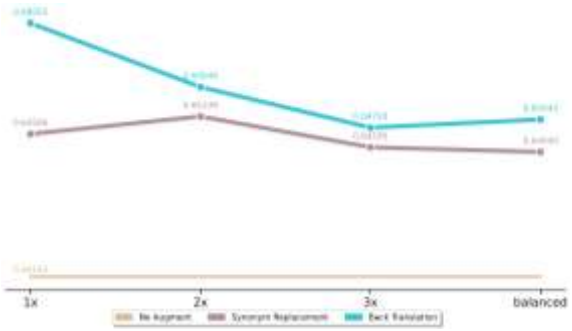
Model	*Dataset*	Skor F1 Label Minoritas		
		WA	SR	BT
SVM	*Clickbait*	0,61973	0,61881	0,62032
	Hate speech	0,60057	0,63288	0,63772
Indo-BERT	*Clickbait*	0,83515	0,83230	0,84044
	Hate speech	0,61789	0,64500	0,66974

Back translation mampu meningkatkan skor F1 sebesar 0,06% pada *dataset clickbait* dan 3,72% pada *dataset hate speech* dengan model SVM. Sementara itu, pada model IndoBERT, *back translation* mampu meningkatkan skor F1 sebesar 0,53% pada *dataset clickbait* dan 5,19% pada *dataset hate speech.* Selain itu, *synonym replacement* mampu meningkatkan skor F1 *dataset hate speech* sebesar 3,23% dengan SVM dan sebesar 2,71% dengan IndoBERT. Hasil eksperimen juga menunjukkan kenaikan yang signifikan terjadi pada *dataset hate speech* yaitu sekitar 3-5% dibanding *dataset clickbait* yang meningkat sekitar <1%. *Dataset hate speech* memiliki *instance* yang jauh lebih sedikit dibandingkan dengan *dataset clickbait*. Hal tersebut dapat menjadi penyebab augmentasi teks lebih efektif pada *dataset hate speech*. Selain itu, augmentasi terlihat lebih berpengaruh pada model klasifikasi IndoBERT.

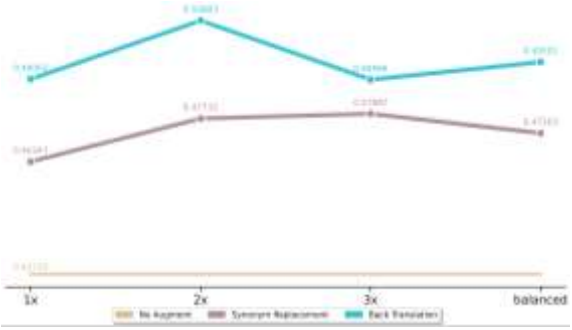
3.3.2. Kondisi *Imbalance*

Gambar 5 dan 6 menunjukkan skor F1 label minoritas hasil klasifikasi SVM pada berbagai skenario augmentasi kedua *dataset*. Berdasarkan

hasil eksperimen, augmentasi yang dilakukan mampu meningkatkan skor F1 label minoritas pada kedua *dataset*. Selain itu, teknik *back translation* lebih baik dibandingkan dengan teknik *synonym replacement* dalam meningkatkan performa model terhadap label minoritas. Pada skenario terbaik, *back translation* mampu meningkatkan skor F1 pada *dataset clickbait* sebesar 7,84% dan pada *dataset hate speech* sebesar 8,16%. Sementara itu, teknik *synonym replacement* mampu meningkatkan skor F1 pada *dataset clickbait* sebesar 4,41% dan pada *dataset hate speech* sebesar 5,16%. Tren dari skenario augmentasi menunjukkan bahwa semakin banyak augmentasi yang dilakukan belum tentu menghasilkan skor F1 yang semakin baik. Setiap *dataset* dengan penerapan teknik augmentasi yang berbeda memiliki skenario jumlah augmentasi terbaiknya masing-masing, tetapi cenderung memiliki pola penurunan yang sama pada skenario *balanced*. Skenario *balanced* memanfaatkan data sintetis paling banyak. Hal tersebut mengindikasikan terjadinya *overfitting* yang dikarenakan terlalu banyak data sintetis yang ditambahkan pada *train set* atau saat proses pelatihan.



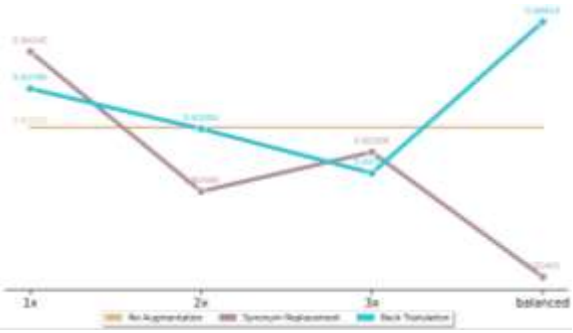
Gambar 5. Skor F1 Klasifikasi *Dataset Clickbait* dengan SVM



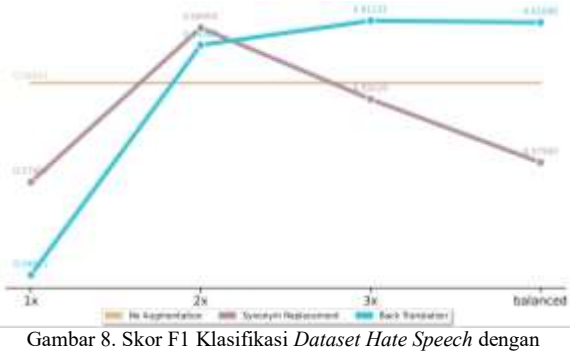
Gambar 6. Skor F1 Klasifikasi *Dataset Hate Speech* dengan SVM

Selanjutnya dilakukan eksperimen dengan model klasifikasi berbasis *pretrained transformer* IndoBERT. Hasil klasifikasi ditunjukkan pada Gambar 7 dan 8. Sesuai dengan klasifikasi menggunakan SVM, augmentasi yang diterapkan mampu meningkatkan skor F1 pada label minoritas. Teknik *back translation* mampu meningkatkan skor F1 label minoritas sebesar 1,31% pada *dataset clickbait* dan 1,59% pada *dataset hate speech.* Sementara itu, teknik *synonym replacement* mampu meningkatkan skor F1 sebesar 0,94% pada *dataset clickbait* dan 1,40% pada *dataset hate speech.* Oleh karena itu, teknik terbaik untuk klasifikasi kedua *dataset* dengan IndoBERT adalah *back translation*. Tren dari skenario jumlah augmentasi juga menunjukkan penurunan pada jumlah augmentasi yang besar kecuali pada *dataset clickbait* teknik *back translation* yang memperoleh hasil terbaik pada kondisi *balanced*. Hal tersebut menunjukkan indikasi *overfitting* ketika teks sintetis terlalu banyak ditambahkan pada *train set.*

Pada kondisi *imbalance*, augmentasi yang dilakukan lebih efektif pada SVM dibandingkan dengan IndoBERT. Peningkatan skor F1 yang diperoleh pada SVM sekitar 4-8% sementara pada IndoBERT hanya sekitar 1%. Meskipun peningkatan skor F1 lebih tinggi pada SVM, nilainya masih di bawah hasil klasifikasi dengan IndoBERT walaupun tanpa dilakukan augmentasi. IndoBERT merupakan *pretrained model* sehingga telah memiliki pengetahuan dari pelatihan sebelumnya pada *dataset* yang lebih besar sehingga hasil klasifikasi awal sudah baik. Sementara itu, SVM merupakan model *machine learning* yang melakukan klasifikasi berdasarkan fitur yang diekstrasi hanya dari data latih yang tersedia. Penambahan teks sintetis pada data dengan ketidakseimbangan yang parah berfungsi memperkaya model klasifikasi SVM.



Gambar 7. Skor F1 Klasifikasi *Dataset Clickbait* dengan IndoBERT



Gambar 8. Skor F1 Klasifikasi *Dataset Hate Speech* dengan IndoBERT

Pada model IndoBERT juga terjadi penurunan skor F1 terutama pada teknik *synonym replacement*. Selain indikasi *overfitting*, proses *generate* teks baru misalnya pada pemilihan sinonim pengganti yang

kurang sesuai secara konteks kalimat dapat mengganggu kinerja model IndoBERT yang mengklasifikasikan teks berdasarkan konteks kalimat. Hal tersebut sejalan dengan penelitian oleh Kobayashi (2018) yang menyatakan teknik *synonym replacement* sering kali menghasilkan teks sintetis yang belum cukup baik. Selain itu, augmentasi lebih berpengaruh pada *dataset hate speech* yang dapat dilihat dari skor F1 terbaik yang selalu lebih tinggi nilainya dibandingkan *dataset clickbait. Dataset hate speech* memiliki jumlah *instance* yang jauh lebih kecil dibandingkan *dataset clickbait* sehingga teks sintetis yang ditambahkan signifikan dalam memperkaya pengetahuan model.

Berdasarkan hasil dua kondisi yang diterapkan yaitu *slightly imbalanced* dan *imbalanced*, terdapat kesamaan yaitu augmentasi yang dilakukan mampu meningkatkan skor F1. Selain itu, augmentasi signifikan berpengaruh pada *dataset hate speech*. Kondisi *dataset* yang kurang ideal seperti jumlah *instance* yang jauh lebih sedikit dibandingkan *dataset clickbait* dan kata-kata informal yang termuat di dalamnya membuat augmentasi dapat lebih memperbaiki kondisi *dataset* tersebut. Kedua kondisi ketidakseimbangan juga menunjukkan teknik terbaik yang sama yaitu *back translation* karena mampu meningkatkan skor F1 yang jauh lebih tinggi.

Peningkatan skor F1 akibat augmentasi untuk model SVM lebih signifikan pada kondisi *imbalanced,* sedangkan model IndoBERT lebih signifikan pada kondisi *slightly imbalanced.* Kondisi ketidakseimbangan yang parah di mana jumlah label minoritasnya jauh lebih sedikit membuat model IndoBERT lebih rentan terhadap *overfitting.* Model IndoBERT merupakan model berbasis *transformer* yang memiliki banyak parameter dan *layer* untuk perhitungan yang kompleks dibanding SVM sehingga dapat *overfitting* ketika digunakan pada *dataset* yang kecil (Fan *et al.*, 2020; Wang *et al.*, 2020; Zhu *et al.*, 2021). Oleh karena itu, IndoBERT lebih signifikan pada kondisi *slightly imbalanced.* Sementara itu, SVM merupakan algoritma *machine learning* yang hanya dilatih dari korpus *dataset* yang digunakan sehingga ketika kondisinya semakin tidak seimbang (*imbalanced*), maka augmentasi akan semakin berpengaruh karena berperan dalam memperkaya model

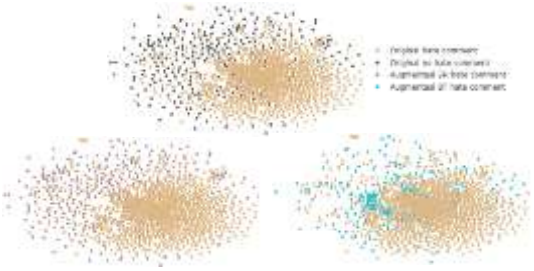
3.4. Evaluasi Augmentasi

3.4.1. Evaluasi Otomatis

a. Plot t-SNE

Vektor teks diperoleh dengan TF-IDF *vectorizer* kemudian diproses menggunakan t-SNE untuk mendapatkan representasi vektor dua dimensi. Teks asli dan teks sintetis hasil augmentasi dari kedua *dataset* divisualisasikan pada Gambar 9 dan 10. Terlihat bahwa kedua label cenderung berada pada wilayah terpisah walaupun terdapat *overlapping* terutama pada *dataset clickbait*. Teks sintesis hasil augmentasi terlihat menempati wilayah label teks aslinya yaitu label minoritas. Hal tersebut menunjukkan bahwa hasil augmentasi baik karena tidak signifikan mengubah semantik kalimat sintetis berdasarkan pendekatan visual (Van Der Maaten, 2015).

Gambar 9. Plot t-SNE *Dataset Clickbait*



Gambar 10. Plot t-SNE *Dataset Hate Speech*

b. BLEU

Hasil evaluasi dengan BLEU tertera pada Tabel 3. Skor BLEU diperoleh dari perhitungan teks sintesis dengan teks aslinya. Sementara *self*-BLEU (SBLEU) diperoleh dengan menghitung skor BLEU setiap pasangan teks sintesis dari teks asli yang sama.

Tabel 3. Skor Evaluasi BLEU

Dataset	**Teknik Augmentasi**	**Skor BLEU**	**Skor SBLEU**
Clickbait	*Synonym replacement*	0,33615	0,68846
	Back translation	0,38508	0,69832
Hate speech	*Synonym replacement*	0,26318	0,30276
	Back Translation	0,03850	0,11782

BLEU dan SBLEU memiliki pola yang sama. Pada kedua *dataset* terjadi perbedaan hasil di mana *dataset clickbait* memiliki aspek keragaman yang lebih baik pada teknik *synonym replacement*, sementara *dataset hate speech* beragam pada teknik *back translation*. Kondisi teks informal pada *dataset hate speech* membuat perbedaan yang berarti karena normalisasi oleh Google Translate memberi hasil terjemahan yang formal atau baku.

c. BERTScore

Pengukuran kelestarian konten dilakukan dengan menggunakan BERTScore dan menghasilkan pola yang berbeda untuk kedua *dataset*. Hasil evaluasi BERTScore dapat dilihat pada Tabel 4.

Evaluasi dengan BERTScore menunjukkan bahwa teknik terbaik untuk *dataset clickbait* adalah teknik *back translation*. Sementara itu, teknik terbaik untuk *dataset hate speech* adalah teknik *synonym replacement.* BERTScore dihitung pada pasangan teks sintetis dengan teks asli. Kata-kata informal yang

dinormalisasi menyebabkan teknik *synonym replacement* lebih cocok pada *dataset hate speech* secara kelestarian semantik konten.

Tabel 4. Skor Evaluasi BERTScore

Dataset	Teknik Augmentasi	BERTScore
Clickbait	*Synonym replacement*	0,84465
	Back translation	0,89642
Hate speech	*Synonym replacement*	0,85613
	Back Translation	0,76301

3.4.2. *Human Evaluation*

Human evaluation dilakukan oleh tiga orang evaluator yang kemudian skor yang diperoleh dirata-ratakan. Dipilih sebanyak 30 *instance* secara acak untuk dievaluasi. Hasil evaluasi indikator *readability* tertera pada Tabel 5 yang menunjukkan teknik augmentasi *back translation* lebih baik dibandingkan dengan teknik *synonym replacement*. Hal tersebut dikarenakan teknik *synonym replacement* menghasilkan teks yang belum cukup baik (Kobayashi, 2018). Nilai *synonym replacement* yang negatif dan mendekati nilai -100 menunjukkan kualitas keterbacaan teks hasil augmentasinya belum baik. Sementara itu, teks asli tidak memperoleh skor penuh, yang artinya teks sintetis berhasil mengungguli teks asli atau mengelabuhi evaluator pada beberapa *instance*.

Tabel 5. Skor BWS Indikator *Readability*

Dataset	Asli	*Synonym Replacement*	*Back Translation*
Clickbait	62,22	-74,44	12,22
Hate speech	63,33	-86,67	23,33

Indikator *label agreement* dinilai oleh evaluator dengan memilih label yang paling sesuai dengan teks sintetis. Hasilnya ditunjukkan dengan persentase teks yang sesuai dengan label teks aslinya atau label minoritas. Tabel 6 menunjukkan hasil evaluasi *label agreement*. Terdapat perbedaan dari kedua teknik augmentasi untuk kedua *dataset*. Hasil terjemah Google Translate yang menormalisasi kata-kata informal pengguna yang banyak mengandung kata kasar, makian, dan ujaran kebencian. Normalisasi tersebut dapat mengurangi intensitas kebencian/*hate* sehingga dapat menjadi penyebab ketidaksesuaian label.

Tabel 6. Hasil Evaluasi Indikator *Label Agreement*

Dataset	*Synonym Replacement*	*Back Translation*
Clickbait	73,33	83,33
Hate speech	92,22	72,22

Berdasarkan hasil evaluasi indikator *readability,* augmentasi teks informal belum cukup baik dari segi kualitas keterbacaannya. Kata-kata informal yang tidak baku dari *dataset* dapat menjadi permasalahan saat augmentasi dilakukan. Contohnya pada teknik *synonym replacement* yang mengidentifikasi POS *tag* kata dan pencarian sinonim di Tesaurus. Kedua hal tersebut dapat dilakukan dengan baik jika kata yang menjadi input merupakan kata baku. Selain itu, pada teknik *back translation*, penerjemah Google Translate akan menormalisasi kata tidak baku pada teks informal dan jika penerjemah kurang dapat menangkap konteks kalimat, maka penerjemah akan mencoba mencari konteks tersendiri dari proses belajarnya. Contoh implementasinya adalah munculnya *suggestion* saat menggunakan Google Translate via web ketika Google Translate tidak memahami kata input. Interpretasi dari Google Translate dapat menjadi kurang tepat dan menghasilkan teks terjemahan balik yang kurang sesuai dengan teks aslinya.

Augmentasi yang dilakukan menghasilkan teks sintetis yang baik untuk *dataset* teks formal karena kata-kata baku di dalamnya mendukung augmentasi untuk dapat diidentifikasi POS *tag* dan juga pencarian sinonimnya. Saat dilakukan terjemahan pun hasilnya baik karena teks formal umum digunakan sehingga mudah diinterpretasikan dengan benar oleh penerjemah. Sebaliknya, augmentasi yang dilakukan belum dapat menghasilkan teks sintetis yang cukup baik pada *dataset* informal. Kriteria *dataset* yang mengandung kata tidak normal membuat sulit untuk mengidentifikasi POS *tag* dan pencarian sinonim. Ditambah lagi teks informal ketika diterjemahkan sering kali menghasilkan teks terjemahan yang memiliki konteks dan makna yang berbeda dari teks asli karena penerjemah kurang dapat memahami konteks sehingga memberi interpretasinya sendiri sesuai proses belajarnya yang kerap kali tidak sesuai. Kata-kata tidak normal menjadi penyebab augmentasi yang dilakukan kurang maksimal.

Selain itu, evaluasi kelestarian semantik konten dengan BERTScore dan *label agreement* dengan *human evaluation* menunjukkan bahwa teks formal cocok jika menggunakan teknik *back translation,* sementara teks informal cocok jika menggunakan teknik *synonym replacement.* Normalisasi oleh Google Translate yang dijelaskan sebelumnya menjadi penyebab teks informal lebih baik jika menggunakan teknik *synonym replacement.* Namun, evaluasi aspek keragaman menunjukkan hal sebaliknya. Semakin beragam teks sintetis menunjukkan augmentasi yang dilakukan semakin kuat (Okimura *et al.*, 2022). Teks formal memiliki augmentasi yang lebih beragam jika menggunakan teknik *synonym replacement,* sementara teks informal lebih beragam jika menggunakan teknik *back translation.* Setelah dilakukan analisis lebih mendalam pada *dataset,* teks formal yang memiliki kata-kata baku sehingga dapat diidentifikasi dan dilakukan penggantian sinonim sehingga hasilnya lebih beragam dibanding teks informal. Sementara itu, untuk *back translation* yang menormalisasi teks informal membuat keragamannya lebih tinggi dibanding teks formal. Analisis lebih lanjut memperoleh fakta bahwa terdapat kata-kata yang tidak sesuai konteks kalimat asli. Oleh karena itu, semakin beragam teks sintetis belun tentu semakin

baik jika konteksnya tidak sesuai karena dapat membahayakan model.

4. KESIMPULAN

Berdasarkan hasil dari penelitian yang dilakukan, dapat disimpulkan beberapa hal sebagai berikut.

1. Augmentasi teks dapat menjadi salah satu alternatif untuk pengembangan *dataset* pelatihan pada permasalahan data yang tidak seimbang. Penelitian ini menerapkan teknik augmentasi *synonym replacement* dengan bantuan Tesaurus dan *back translation* menggunakan API Google Translate pada teks berbahasa Indonesia.
2. Hasil klasifikasi menunjukkan augmentasi teks dapat meningkatkan skor F1 pada data yang tidak seimbang. Pada penelitian ini, augmentasi mampu meningkatkan skor F1 sekitar 0,1% hingga 8%. Selain itu, augmentasi teks lebih efektif pada *dataset* kecil. Berdasarkan hasil eksperimen, jumlah augmentasi memengaruhi peningkatan atau penurunan skor F1. Jumlah augmentasi yang besar belum tentu semakin baik karena model terindikasi *overfitting* pada *train set*.
3. Berdasarkan hasil klasifikasi dan evaluasi augmentasi, teknik augmentasi *back translation* lebih baik dibandingkan dengan teknik *synonym replacement*. Augmentasi yang dilakukan menghasilkan teks sintetis yang baik untuk data teks formal. Kata-kata tidak normal pada teks informal menjadi penyebab proses augmentasi kurang maksimal. Teks formal cocok menggunakan teknik *back translation,* sementara teks informal cocok dengan teknik *synonym replacement.*

Selain itu, saran untuk penelitian yang dapat dilakukan selanjutnya adalah optimasi teknik *synonym replacement* dalam menentukan sinonim yang sesuai konteks kalimat agar dapat meningkatkan akurasi. Selanjutnya, perlu dilakukan eksperimen mengenai penerapan augmentasi teknik *synonym replacement* dan *back translation* pada teks panjang.

DAFTAR PUSTAKA

ABDURRHAMAN dan PURWARIANTI, A., 2019. Effective Use of Augmentation Degree and Language Model for Synonym-based Text Augmentation on Indonesian Text Classification. In 2019 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2019), pp. 217-222.

ALI, A., SHAMSUDDIN, S.M., dan RALESCU, A.L., 2015. Classification with Class Imbalance Problem: A Review. International Journal of Advances in Soft Computing and its Applications, 7(3), pp. 176-204.

BARUA, S., ISLAM, M.M., YAO, X., dan MURASE, K., 2014. MWMOTE – Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. IEEE Transactions of Knowledge and Data Engineering, 26(2), pp. 405-425.

CAHYAWIJAYA, S., WINAYA, G.I., WILIE, B., VINCENTIO, K., LI, K., KUNCORO, A., RUDER, S., LIM, Z.Y., BAHAR, S., KHODRA, M., PURWARIANTI, A., dan FUNG, P., 2021. IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 8875-8898.

CHOWDHARY, K.R., 2020. Natural Language Processing BT – Fundamentals of Artificial Intelligence. Jodhpur: Springer.

COULOMBE, C., 2018. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs, [online] Tersedia di: <http://arxiv.org/abs/1812.04718> [Diakses 1 November 2022]

DINAKARAMANI, A., RASHEL, F., LUTHFI, A., dan MANURUNG, R., 2014. Designing an Indonesian Part of Speech Tagset and Manually Tagged Indonesian Corpus. In Proceedings of the International Conference on Asian Language Processing 2014 (IALP 2014), pp. 66-69.

ESTABROOKS, A., JO, T., JAPKOWICZ, N., 2004. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, 20(1), pp. 18-36.

FAN, A., GRAVE, E., dan JOULIN, A., 2019. Reducing Transformer Depth on Demand with Structured Dropout. ArXiv, abs/1909.11556.

FENG, STEVEN Y., GANGAL, V., KANG, D., MITAMURA, T., dan HOVY, E., 2020. GenAug: Data Augmentation for Finetuning Text generators. In Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architecture, pp. 29-42, Online. Association for Computational Linguistics.

GU, Q., WANG, X.M., WU, Z., NING, B., dan XIN, C.S., 2016. An Improved SMOTE Algorithm Based on Genetic Algorithm for Imbalanced Data. Journal of Digital Information Management, 14(2), pp. 92-103.

GUDIVADA, V.N., BAEZA-YATES, R., dan RAGHAVAN, V.V., 2015. Big Data: Promises and Problems. Computer (Long. Beach Calif), 48(3), pp. 20-23.

HAN, H. dan JIANG, X., 2014. Ovecome Support Vector Machine Diagnosis Overfitting. Cancer Informatics, 13(S1), pp. 145-158.

HARYWANTO, G.N., VERON, J.S., dan SUHARTONO, D., 2022. A BERTweet-based Design for Monitoring Behaviour Change Based on Five Doors Theory on Coral Bleaching Campaign. Journal of Big Data, 9(1), pp. 1-22.

HIRSCHBERG, J. dan MANNING, C.D., 2015. Advanced in Natural Language Processing. Science, 349(6245), pp. 261-266.

JIANG, Z., PAN, T., ZHANG, C., dan YANG, J., 2021. A new Oversampling Method Based on the Classification Contribution Degree. Computer Science and Symmetry/Asymmetry, 13(2), pp. 1-13.

JUNGIEWICZ, M. dan SMYWINSKI-POHL, A., 2019. Towards Textual Data Augmentation for Neural Networks: Synonym and Maximum Loss. Computer Science, 20(1), pp. 57-84.

KOBAYASHI, S., 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (NAACL HLT 2018), vol. 2, pp. 452-457.

KURNIA, N., 2020. *Big Data* untuk Ilmu Sosial: Antara Metode Riset dan Realitas Sosial. Yogyakarta: UGM PRESS.

LIU, R., XU, G., JIA, C., MA, W., WANG, L., dan VOSOUGHI S., 2020. Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9031-9041, Online. Association for Computational Linguistics.

LOGIT, I. dan ORME, B., 2009. MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB. Sawtooth Software Research Paper Series, 98382(360), pp. 1-7.

LU, Q., DOU, D., dan NGUYEN, T.H., 2021. Textual Data Augmentation for Patient Outcomes Prediction. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2817-2821, Houston, TX. USA.

MADUKWE, K.J., GAO, X., dan XUE, B., 2022. Token Replacement-based Data Augmentation Methods for Hate Speech Detection. World Wide Web, 25(3), pp. 1129-1150.

NATASYA dan GIRSANG, A.S., 2023. Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis. Emerging Science Journal, 7(1), pp. 256-272.

OKIMURA, 1., REID, M., KAWANO, M., dan MATSUO, Y., 2022. On the Impact of Data Augmentation on Downstream Performance in Natural Language Processing. In Proceedings of the Third Workshop on Insights from Negative results in NLP, pp. 88-93.

PANDEY, S., AKHTAR, M.S., dan CHAKRABORTY, T., 2021. Syntactically Coherent Text Augmentation for Sequence Classification. In IEEE Transactions on Computational Social Systems, 8(6), pp. 1323-1332.

PAPINENI, K., ROUKOS, S., WARD, T., dan ZHU, W.J., 2002. Bleu: A Method for Automatic Evaluation for Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318.

RAZNO, M., 2019. Machine Learning Text Classification Model with NLP Approach. In Computational Linguistics and Intelligent Systems, Proceedings of the 3rd International Conference, vol. 2, pp. 71-73.

RUPAPARA, V., RUSTAM, F., SHAHZAD, H.F., MEHMOOD, A., ASHRAF, I., dan CHOI, G.S., 2021. Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. IEEE Access, 9, pp. 78621-78634.

SAEZ, J.A., LUENGO, J., STEFANOWSKI, J., dan HERRERA, F., 2015. SMOTE-IPF: Addressing the Noisy and Borderline Examples Problem in Imbalanced Classification by A Resampling Method with Filtering. Information Sciences, vol. 291, pp. 184-203.

SANYA, A.D. dan SUADAA, L.H., 2022. Handling Imbalanced Dataset on Hate Speech detection in Indonesian Online News Comments. In 10th International Conference on Information and Communication Technology (ICoICT), pp. 380-385.

SOLTANZADEH, P. dan HASHEMZADEH, M., 2021. RCSMOTE: Range-Controlled Synthetic Minority Oversampling Technique for Handling the Class Imbalance Problem. Information Sciences, vol. 542, pp. 92-111.

SUN, A., LIM, E.P., dan LIU, Y., 2009. On Strategies for Imbalanced text Classification using SVM: A Comparative Study. Decision Support Systems, 48(1), pp. 191-201.

SUTOYO, E. dan FADLURRAHMAN, M.A., 2020. Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. Jurnal Edukasi dan Penelitian Informatika, 6(3), p. 379.

TESFAGERGISH, S.G., DAMASEVICIUS, R., dan KAPOCIUTE-DZIKIENE, J., 2021. Deep Fake Recognition in Tweets Using Text Augmentation, Word Embeddings, and Deep Learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12954 LNCS, pp. 523-538.

VAN DER MAATEN, L., 2015. Accelerating t-SNE Using Tree-Based Algorithm. Journal of Machine Learning Research, vol. 15, pp. 3221-3245.

VERDIKHA, N.A., ADJI, T.B., dan PERMANASARI, A.E., 2018. Komparasi Metode Oversampling untuk Klasifikasi Teks Ujaran Kebencian. Seminar Nasional Teknologi Informasi dan Multimedia, 1(2), pp. 85-90.

WANG, S., LI, Z., CHAO, W., dan CAO, Q., 2012. Applying Adaptive Oversampling Technique Based on Data Density and Cost-sensitive SVM to Imbalanced Learning. In Proceedings of the International Joint Conference on Neural Networks, pp. 10-15.

WANG, Y., LIU, F., VERSPOOR, K., dan BALDWIN, T., 2020. Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 105–111, Online. Association for Computational Linguistics.

WILLIAM, A. dan SARI, Y., 2020. CLICK-ID: A Novel Dataset for Indonesian Clickbait Headlines. Data In Brief, vol. 32, p. 106231.

XIANG, R., CHERSONI, E., LU, Q., HUANG, C.R., LI, W., dan LONG, Y., 2021. Lexical Data Augmentation for Sentiment Analysis. Journal of the Association for Information Science and Technology, 72(11), pp. 1432-1447.

ZHANG, T., KISHORE, V., WU, F., WEINBERGER, K.Q., dan ARTZI, Y., 2019. Bertscore: Evaluating Text Generation with Bert, [online] Tersedia di: <https://arxiv.org/abs/1904.09675v3> [Diakses 6 Januari 2023]

ZHANG, X. dan LECUN, Y., 2015. Text Understanding from Scratch, [online] Tersedia di: <http://arxiv.org/abs/1502.01710> [Diakses 1 November 2022]

ZHU, D., LIN, W., ZHANG, Y., ZHONG, Q., ZENG, G., WU, W., dan TANG, J., 2021. AT-BERT: Adversarial Training BERT for Acronym Identification Winning Solution for SDU@AAAI-21. CEUR-WS, 2831(28).