Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 8, No. 2, Juni 2023 (121-128) **10.32493/informatika.v8i2.29306**

Penggunaan Kamus Singkatan Kata Bahasa Indonesia Sehari-Hari dalam Pembangkitan Fitur Teks

Citra Lestari[1], Kenny Jihiro[2], Andreas Lim[3], Daniel Aprilio[4], dan Franciscus Valentinus[5]

[1]Informatika, Universitas Cipura, CBD Boulevard CitraRaya Surabaya, Indonesia, 60219
e-mail: [1]caecilia.citra@ciputra.ac.id

[2,5]Informatika, Universitas Cipura, CBD Boulevard CitraRaya Surabaya, Indonesia, 60219
e-mail: [2] kjinhiro@student.ciputra.ac.id [3] andreaslim01@student.ciputra.ac.id,
[4]daprillio@student.ciputra.ac.id, [5]fvalentinus@student.ciputra.ac.id

Submitted Date: 2023-03-28 Reviewed Date: 2023-03-29
Revised Date: 2023-04-01 Accepted Date: 2023-04-17

Abstract

Natural Language Processing (NLP) research on Indonesian language is relatively slow compared to other languages, such as English or Chinese. Most of the research is dealing with Indonesian formal texts. Some NLP researches that are dealing with Indonesian informal texts are having quite difficulty since Indonesian informal language usually combines formal language, daily language, and local language. In addition, there is a habit in Indonesians to use abbreviations in texting. These cause great difficulty in the feature generation process, where machines fail to identify stopwords and form lemmas from the bag of words. There are actually dictionaries that can be used to do the lemming process for Indonesian forma language, daily language, local languages, and even Indonesian formal abbreviations. But there is still no dictionary for Indonesian informal abbreviations. This research made an Indonesian informal abbreviations dictionary from 4000 Indonesian tweets. The dictionary contains 706 unique abbreviations as its corpus. The dictionary then used to generate features. In this research, the feature generation only used this dictionary to measure its significance. The feature generation with the Indonesian informal abbreviations dictionary were tested with Indonesian tweets about Covid-19 Vaccine. The features generation process was able to identify 2262 abbreviations with 71,09% of them identified as stopwords. To take a further step, the features generated are then being tested to figure out their impact in sentiment analysis. The sentiment analysis used Multi-Layer Perceptron. Unfortunately, those features didn't increase the performance of the sentiment analysis. The accuracy decreased by 3,5% while the precision, recall, and F1-Score decreased in the range of 0,02 – 0,04. With this result, it can be concluded that the use of this dictionary alone for the lemming process is not enough. It needs to be combined with other dictionary to have a more optimal result.

Keywords: Indonesian informal; abbreviation; dictionary; lemming; feature generation

Abstrak

Penelitian Pemrosesan Bahasa Natural (Natural Language Processing, selanjutnya disingkat NLP) pada Bahasa Indonesia relatif lambat jika dibandingkan dengan bahasa lain seperti Bahasa Inggris atau Mandarin. Kebanyakan penelitian NLP pada Bahasa Indonesia masih mengenai teks formal/baku. Beberapa penelitian NLP dengan teks Bahasa Indonesia informal

mengalami kesulitan dikarenakan Bahasa Indonesia informal umumnya mengombinasikan Bahasa formal/baku, bahasa gaul, dan Bahasa daerah. Kebiasaan masyarakat Indonesia untuk menyingkat kata dalam penulisan teks juga menjadi permasalahan tambahan. Hal-hal tersebut memberikan kesulitan yang besar dalam proses pembangkitan fitur dimana mesin gagal untuk mengidentifikasi *stopwords* dan membentuk lemma dari kumpulan kata. Sebenarnya telah terdapat kamus Bahasa Indonesia baku, kamus Bahasa gaul, kamus Bahasa daerah, dan kamus singkatan kata baku yang dapat digunakan dalam proses lemmatisasi, yaitu proses pencarian bentuk dasar (*lemma*) sebuah kata. Namun belum ada kamus singkatan kata Bahasa Indonesia sehari-hari. Penelitian ini menghasilkan kamus singkatan kata Bahasa Indonesia sehari-hari yang terdiri dari 706 *corpus* singkatan yang ditemukan dari 4000 cuitan berbahasa Indonesia. Kamus itu kemudian digunakan dalam tahap lemmatisasi pada proses pembangkitan fitur. Pada penelitian ini tahap lemmatisasi hanya menggunakan kamus singkatan kata Bahasa Indonesia sehari-hari saja untuk mengetahui signifikansinya. Signifikansi pembangkitan fitur dengan lemmatisasi singkatan kata Bahasa Indonesia sehari-hari kemudian diujikan pada cuitan berbahasa Indonesia tentang vaksin Covid-19. Proses pembangkitan fitur mampu mengidentifikasi 2262 singkatan kata yang 71,09% di antaranya teridentifikasi sebagai *stopwords*. Mengambil langkah selanjutnya, pembangkitan fitur-fitur yang dihasilkan kemudian diujikan untuk melihat dampaknya pada performa analisis sentimen. Metode analisis sentimen yang digunakan adalah Multi-Layer Perceptron. Sayangnya, fitur-fitur yang dihasilkan tersebut tidak dapat meningkatkan performa analisis sentimen. Akurasi analisis sentimen justru mengalami penurunan sebanyak 3,5% sementara presisi, *recall*, dan F1-Score mengalami penurunan dalam kisaran 0,02 hingga 0,04. Dengan hasil ini dapat disimpulkan bahwa penggunaan kamus ini saja tidaklah cukup dalam proses lemmatisasi, melainkan perlu digabungkan dengan kamus-kamus lain untuk mendapatkan hasil yang lebih optimal.

Kata Kunci: singkatan kata; Bahasa Indonesia informal; kamus; lemmatisasi; pembangkitan fitur

1. Pendahuluan

Pemrosesan Bahasa Natural (*Natural Language Processing*, selanjutnya disingkat NLP) adalah salah satu bidang dari kecerdasan buatan yang bertugas untuk membuat komputer memahami bahasa natural atau sehari-hari dari manusia. Penelitian terkait NLP telah sangat banyak terutama dalam Bahasa Inggris, misalnya NLP dalam penelitian kualitatif kesehatan umum seperti yang dilakukan oleh Lesson, dkk (2019), model pemrosesan bahasa (Floridi & Chiriatti, 2020), atau pembangkitan deskripsi gambar (Kameswari, 2021)

Dalam Bahasa Indonesia, NLP banyak dimanfaatkan untuk teks-teks formal. Ratnasari , dkk (2014) membuat model NLP untuk keluhan pasien berdasarkan rekaman wawancara medis. Elcholiqi dan Musdholifa (2020) membuat *chatbot* untuk informasi perbankan. Nayoga, dkk (2021) mengembangkan NLP untuk analisis *hoax* pada berita berbahasa Indonesia.

Beberapa penelitian juga berusaha memanfaatkan NLP pada data teks informal seperti tulisan di sosial media, baik dalam bentuk *post*, ulasan, cuitan, atau komentar. Umumnya penelitian-penelitian tersebut melakukan analisis sentiment atau opini masyarakat terhadap suatu topik atau permasalahan.

Kendala yang terjadi dalam penelitian NLP untuk teks informal adalah rendahnya akurasi performa model klasifikasi yang dihasilkan. Lestari, dkk (2022) menengarai penyebab rendahnya akurasi performa model disebabkan karena mayoritas cuitan menggunakan percampuran dari bahasa Indonesia baku, bahasa sehari-hari atau disebut juga bahasa gaul, bahasa daerah, dan juga kebiasaan masyarakat Indonesia untuk menyingkat kata dalam penulisan teks.

Sebenarnya terdapat suatu tahap pada NLP yang dapat mengubah suatu kata menjadi kata yang bermaksa sama, disebut Lemmatisasi (*lemmatization*). Proses ini

mencarikan padanan dari sebuah kata dengan kumpulan kata yang diindeks menjadi sebuah kamus. Telah terdapat kamus lemmatisasi Bahasa Indonesia baku yang diambil dari Kamus Besar Bahasa Indonesia (Zaky, 2019). Juga telah terdapat kamus lemmatisasi untuk Bahasa Indonesia sehari-hari (Salsabila, dkk, 2018) dan Bahasa alay (Ibrahim, 2018) . Namun, belum terdapat kamus lemmatisasi singkatan kata sehari-hari dalam Bahasa Indonesia.

Artikel ini membahas pembuatan kamus singkatan kata bahasa Indonesia sehari-hari yang kemudian digunakan dalam lemmatisasi di tahap pembnagkitan fitur NLP. Pengujian kemudian dilakukan untuk mengetahui dampak lemmatisasi singkatan kata bahasa Indonesia sehari-hari pada identifikasi *stopwords* dan performa analisis sentimen. Pengujian dilakukan dengan perbandingan terhadap hasil penelitian sebelumnya dari Lestari, dkk (2022) yang mana melakukan pembangkitan fitur tanpa tahap lemmatisasi singkatan kata bahasa Indonesia sehari-hari.

2. Metodologi

Metodologi penelitian dibagi menjadi dua tahap besar, yaitu metodologi pembuatan kamus, dan metodologi pembangkitan fitur. Keduanya akan dibahas pada subbab di bawah ini.

2.1. Metodologi Pembuatan Kamus

Terdapat empat tahapan dalam pembuatan kamus singkatan kata bahasa Indonesia sehari-hari seperti yang ditampilkan pada Gambar 1.

Gambar 1. Metodologi Pembuatan Kamus Singkatan Kata

1. Pengumpulan Sumber Data

Sumber data singkatan kata diperoleh dari cuitan (*tweet)* masyarakat pada aplikasi Twitter. Cuitan diambil pada tanggal 15 September 2022 dengan metode *webcrawling*. Cuitan yang diambil hanyalah cuitan dengan bahasa Indonesia dengan topik bebas. Terdapat 4000 cuitan yang berhasil dikumpulkan dalam waktu satu menit 48 detik. Tabel 1 adalah beberapa cuitan yang terkumpul.

Tabel 1. Contoh Hasil Pengumpulan Sumber Data

No	Cuitan
1	Anjir akhirnya aku menemukan au namseok tidak angst😭 tapi plot twist banget sih ga nyangka seok jin kaya gitu
2	knp pas bikin sg muka sendiri jd byk viewersnya😊💄
3	Aku mencintaimu! Jika kamu benci aku, panah saja diriku. Tapi jangan di hatiku ya, karena di situ kamu berada #ForediSalatiga #ForediAmbarawa
4	kayanya gua bakal tobat dari perjokian tugas cok
5	Udah jalannya mereka, jalan kita mungkin masih jalan-jalan.

Dari 4000 cuitan yang dikumpulkan lebih dari 80% memiliki singkatan kata. Cuitan yang tanpa singkatan kata umumnya adalah cuitan berupa berita atau promosi produk.

2. Identifikasi Singkatan Kata

Identifikasi singkatan kata dilakukan secara manual oleh empat orang. Masing-masing orang melakukan identifikasi pada 1000 cuitan. Dengan rata-rata setiap cuitan memiliki tiga singkatan kata, ditemukan kurang lebih 7893 singkatan kata. Beberapa contoh singkatan kata yang teridentifikasi ditampilkan pada Tabel 2.

Tabel 2. Contoh Hasil Identifikasi Singkatan

No	Cuitan	Singkatan
1	Anjir akhirnya aku menemukan au namseok tidak angst😭 tapi plot twist banget sih ga nyangka seok jin kaya gitu	anjir, au, ga, gitu
2	knp pas bikin sg muka sendiri jd byk viewersnya😊💄	Knp, sg, jd, byk
3	Aku mencintaimu! Jika kamu benci aku, panah saja diriku. Tapi jangan di hatiku ya, karena di situ kamu berada #ForediSalatiga #ForediAmbarawa	-

4	kayanya gua bakal tobat dari perjokian tugas cok	Gua, cok
5	Udah jalannya mereka, jalan kita mungkin masih jalan-jalan.	Udah

3. Kurasi Singkatan Kata

Singkatan Kata yang berhasil diidentifikasi oleh empat orang tersebut kemudian dikumpulkan pada lembar sebaran (*spreadsheet*). Singkatan kata yang telah dikumpulkan tersebut kemudian dilakukan kurasi secara manual. Kurasi yang dilakukan adalah pengecekan redundansi dan kebenaran singkatan kata.

Setelah dilakukan kurasi, terdapat banyak singkatan kata yang teridentifikasi secara redundant. Contoh beberapa singkatan kata yang sering muncul ada pada Tabel 3.

Tabel 3. Singkatan Kata Yang Sering Muncul

No	Singkata Kata	Jumlah
1	yg	237
2	ga	153
3	tp	62
4	gw	37
5	km	30

Selain itu, terdapat beberapa kesalahan identifikasi, dimana yang teridentifikasi bukanlah singkatan kata melainkan kata gaul atau *alay*, seperti pada Tabel 4. Terdapat 38 istilah yang salah teridentifikasi sebagai singkatan kata. Tahap Kurasi ini menghasilkan luaran berupa 706 singkatan kata unik yang menjadi korpus dari kamus.

Tabel 4. Kata Gaul Yang Teridentifikasi Sebagai Singkatan Kata

No	Istilah
1	Aer
2	Ane
3	Skip
4	Tenan
5	Trust

4. Pemberian Arti Singkatan Kata

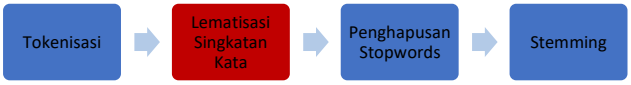
Singkatan-singkatan kata yang telah diakurasi tersebut kemudian dicarikan *lemma* atau istilah akarnya. Proses pemberian arti singkatan kata ini juga dilakukan secara manual. Tabel 5 adalah contoh hasil pengartian singkatan kata.

Tabel 5. Hasil Pengartian Singkatan Kata

No	Singkatan Kata	*Lemma*
1	yg	Yang
2	ga	Tidak
3	tp	Tapi
4	gw	Aku
5	km	Kamu

2.2. Metodologi Pembangkitan Fitur

Pembangkitan fitur yang dilakukan pada penelitian ini adalah pembangkitan fitur Lexicon dengan. metodologi yang engacu pada metodologi yang dibuat oleh Nazeer, dkk (2020) dengan sedikit perubahan/perkembangan. Adapun perubahan/ pengembangan yang dilakukan adalah lemmatisasi singkatan kata. Gambar 2 adalah metodologi pembangkitan fitur yang dilakukan pada penelitian ini. Kotak berwarna biru adalah tahapan dari Nazeer, dkk (2020), sedangkan kotak berwarna merah adalah tahapan tambahan yang dilakukan oleh peneliti.



Gambar 2. Metodologi Pembangkitan Fitur

1. Tokenisasi

Tokenisasi adalah proses pemecahan kalimat menjadi kumpulan kata. Pustaka TweetTokenizer digunakan untuk tahap tokenisasi pada penelitian ini. Hasil tokenisasi ditunjukkan pada Tabel 6.

Tabel 6. Contoh Hasil Tokenisasi

No	Kalimat	Hasil Tokenisasi
1.	serius la sejak budak skola da dbuka utk vaksin penuh ppv weyhsop da jadi soup	['serius', 'la', 'sejak', 'budak', 'skola', 'da', 'dbuka', 'utk', 'vaksin, 'penuh', 'ppv', 'weyhsop', 'da', 'jadi', 'soup']
2	adam benk yg disana itu vaksin apa amp utk apa \ncoba dijelaskan	['adam', 'benk', 'yg', 'disana', 'itu', 'vaksin', 'apa', 'amp', 'utk', 'apa', 'coba', 'dijelaskan']
3	adam benk yg disana itu vaksin	['imf', 'desak', 'negara', 'kaya', 'sumbang',

	apa amp utk apa \ncoba dijelaskan	'vaksin', 'covid', 'republika', 'online', 'covid', 'covid', 'httpstcoowqvqglvdf']
4	cahayaacb ayok vaksin utk diri kita sendiri	['cahayaacb', 'ayok', 'vaksin', 'utk', 'diri', 'kita', 'sendiri']
5	restamojokerto ayo vaksin biar kebal	['restamojokerto', 'ayo', 'vaksin', 'biar', 'kebal']

2. Lemmatisasi Singkatan Kata

Lemmatisasi adalah proses pencarian *lemma* atau bentuk dasar dari sebuah istilah pada sebuah kamus. Pada penelitian ini proses lemmatisasi hanya dilakukan pada singkatan kata bahasa Indonesia sehari-hari dengan menggunakan kamus yang telah dibuat pada metodologi pertama.

3. Penghapusan Stopwords

Pada tahap ini akan dilakukan penghapusan *stopword,* yaitu kata-kata umum yg sering muncul namun tidak penting dalam sebuah dokumen. Daftar *stopwords* yang digunakan diambil dari pustaka Natural Language toolkit (NLTK).

4. Stemming

Stemming adalah metode pencarian bentuk dasar kata dengan pemotongan awalan dan akhiran. Pada penelitian ini metode yang digunakan adalah Sastrawi sebab metode Sastrawi menggabungkan beberapa algoritma *stemming* dan adalah metode yang paling umum digunakan untuk *stemming* bahasa Indonesia.

3. Pengujian dan Hasil

Pengujian pada penelitian dilakukan untuk mengetahui dampak lemmatisasi singkatan kata bahasa Indonesia sehari-hari terhadap pembangkitan fitur dan dampak lemmatisasi singkatan kata bahasa Indonesia sehari-hari pada akurasi analisis sentimen.

Dataset yang digunakan pada pembangkitan fitur ini adalah *dataset* yang digunakan oleh Lestari, dkk (2022) yaitu cuitan berbahasa Indonesia yang memiliki frase ‘vaksin covid’. Terdapat 1394 cuitan pada *dataset* tersebut.

3.1. Pengujian Dampak Lemmatisasi Singkatan Kata pada Pembangkitan Fitur

Pengujian pembangikitan fitur dimulai pada tahap ketiga dari Metodologi Pembangkitan Fitur. Hal tersebut karena tahap ketiga ini, yaitu lemmatisasi singkatan kata adalah tahap yang membedakan antara penelitian ini dengan penelitian yang dilakukan Lestari, dkk (2022).

Pada tahap lemmatisasi dengan kamus singkatan kata bahasa Indonesia sehari-hari berhasil ditemukan 2262 singkatan kata yang kemudian digantikan dengan *lemma*-nya. Tabel 7 adalah beberapa contoh singkatan kata yang berhasil ditemukan.

Pada tahap penghapusan *stopwords*, dataset yang tidak dilakukan lemmatisasi memiliki 5851 *stopwords,* sementara data set yang telah dilakukan proses lemmatisasi teridentifikasi memiliki 7459 s*topwords*. Sehingga terdapat selisih 1608 *stopwords* lebih banyak teridentifikasi pada *dataset* yang telah melalui proses lemmatisasi, atau sekitar 27,48%. Adapun dari data tersebut di atas diketahui bahwa 71,09% singkatan kata yang teridentifikasi dalam proses lemmatisasi adalah *stopwords.*

Tabel 7. Contoh Lemmatisasi Singkatan Kata

Baris Ke-	Singkatan Kata yang Ditemukan	Hasil Lemmatisasi
1	‘blm’, ‘gabisa’, ‘liat’,	‘belum’, ‘tidak bisa’ ‘lihat’’
8	‘utk’	‘untuk’
10	‘yg’, ‘utk’	‘yang’, ‘untuk’
13	‘trs’, ‘bgt’, ‘blg’, ‘kak’, ‘kaka’, ‘liat’	‘terus’, ‘banget’, ‘bilang’, ‘kakak’, ‘kakak’, ‘lihat’
22	‘udah’	‘sudah’

Pada tahap *Stemming* tidak terdapat perbedaan antara dataset tanpa lemmatisasi singkatan kata dan dataset dengan lemmatisasi singkatan kata. Hal tersebut ditengarai karena singkatan kata yang ditemukan pada tahap Lemmatisasi Singkatan Kata telah diubah ke dalam bentuk dasarnya. Di lain pihak, singkatan kata pada dataset yang tanpa lemmatisasi singkatan kata tidak dapat dilakukan *stemming* karena tidak memiliki awalan dan akhiran.

3.2. Pengujian Dampak Lemmatisasi Singkatan Kata pada Analisis Sentimen

Pengujian ini bertujuan untuk mengetahui dampak yang dihasilkan oleh lemmatisasi singkatan kata pada analisis sentimen. Penelitian ini menggunakan metode Multi-Layer Perceptron (disingkat MLP). Dataset yang digunakan adalah fitur-fitur yang dihasilkan oleh Metodologi Pembangkitan Fitur, baik yang menggunakan tahap Lemmatisasi Singkatan Kata maupun yang tidak menggunakan tahapan tersebut.

Sebelum dilatih, kedua dataset mengalami perlakuan yang sama yaitu pembagian *dataset* menjadi data latih dan data uji, serta peneyimbangan jumlah kelas. Kedua *dataset* dibagi menjadi data latih dan data uji dengan pembagian 80% data latih dan 20% data uji.

Penyeimbangan jumlah kelas dilakukan dengan metode SMOTE. Setelah diterapkan metode SMOTE, jumlah kelas yang tadinya tidak seimbang, dimana klas positif berjumlah 7031 dan kelas negative berjumlah 412, telah menjadi seimbang. Masing-masing kelas memiliki jumlah 703.

Tabel 8 menunjukkan perbandingan performa analisis sentimen menggunakan MLP pada kedua dataset. Ternyata proses lemmatisasi singkatan kata tidak memberikan dampak positif pada proses klasifikasi analisis sentimen. Terdapat penurunan performa, meskipun tidak signifikan. Akurasi berkurang 3,5%, sedangkan presisi, recall, dan F1-Score mengalami penurunan dalam kisaran 0,02 hingga 0,04.

Tabel 8. Perbandingan Performa Analisis Sentimen

Kriteria	Dataset Tanpa Lemmatisasi Singkatan Kata	Dataset dengan Lemmatisasi Singkatan Kata
Akurasi	68,8%	65,32%
Presisi	0,82	0,79
Recall	0,63	0,61
F1-Score	0,73	0,69

4. Diskusi

Kamus yang telah dibuat pada penelitian ini telah mampu mengidentifikasi 2262 singkatan kata dan 71,09% diantaranya teridentifikasi sebagai *stopwords*. Meskipun demikian, masih banyak singkatan kata bahasa Indonesia sehari-hari yang belum ada di dalam kamus tersebut. Beberapa di antaranya adalah singkatan kata yang berulang, singkatan kata yang tidak lazim, maupun singkatan kata secara umum yang tidak teridentifikasi dalam 4000 cuitan yang menjadi sumber data.

Bahasa Indonesia banyak menggunakan perulangan kata. Misalnya: gadis-gadis, berlari-lari, tiba-tiba, malam-malam, buah-buahan, dan lainnya. Untuk penulisan kata yg berulang tersebut umumnya masyarakat Indonesia menggunakan karakter angka dua (‘2’) atau karakter tanda petik dua (“”). Terdapat beberapa singkatan perulangan kata yang ditemukan dalam sumber data namun sangat minim jika dibandingkan dengan jumlah perulangan kata yang ada.

Terdapat beberapa singkatan kata yang tidak lazim yang ditemukan dalam sumber data. Singkatan tersebut tidak dimasukkan ke dalam kamus karena penerjemah kurang memiliki pengetahuan akan kata sesungguhnya dari singkatan tersebut. Namun, karena singkatan tersebut tidak lazim, maka penulis beranggapan bahwa singkatan tersebut jarang digunakan oleh masyarakat Indonesia.

Sumber data yang hanya berjumlah 4000 cuitan tentulah merupakan jumlah yang minim untuk menemukan semua singkatan kata bahasa Indonesia yang sering digunakan sehari-hari. Oleh karena itu diyakini masih banyak singkatan kata yang belum tercakup dalam kamus yang berhasil dibuat dalam penelitian ini.

Adapun kegagalan peningkatan performa analisis sentimen dapat disebabkan karena di dalam *dataset* vaksin Covid tersebut, selain terdapat banyak singkatan kata, juga terdapat bahasa gaul dan bahasa daerah yang mendominasi. Tabel 9 adalah beberapa contoh cuitan dengan bahasa gaul atau bahasa daerah. Oleh karena itu penggunaan kamus singkatan kata bahasa Indonesia sehari-hari saja tidak cukup. Kamus tersebut perlu dikombinasikan dengan kamus bahasa gaul dan bahasa daerah saat digunakan dalam proses lemmatisasi.

Tabel 9. Contoh Cuitan dengan Bahasa Gaul atau Bahasa Daerah

No.	Cuitan
1	@Atuyy_yutaa hayuu atuh vaksin
2	pak lurahku kok jan ora masyarakate di edukasi
3	seronok je lahai tengok orang pergi vaksin rame
4	@gitajisung w ikut vaksin di sekul org huhu
5	ttp nggenjreng walaupun hbs vaksin 😀

Alasan lain yang dapat mempengaruhi performa analisis sentimen yang tidak optimal adalah jumlah kelas yang tidak seimbang. Sebelum

analisis dilakukan memang dilakukan penyeimbangan jumlah kelas, namun metode SMOTE melakukan penyeimbangan kelas hanya dengan melakukan replikasi dari data latih yang ada sehingga variasi data tetap sama atau tidak bertambah.

5. Simpulan

Berdasarkan hasil dari dua pengujian di atas dapat disimpulkan bahwa proses lemmatisasi singkatan kata bahasa Indonesia sehari-hari mampu mengidentifikasi 2262 singkatan kata pada dataset cuitan berbahasa Indonesia tentang vaksin covid, yang 71,09% di antaranya adalah *stopwords*. Simpulan lain adalah proses lemmatisasi singkatan kata bahasa Indonesia sehari-hari tidak meningkatkan performa analisis sentimen pada dataset cuitan berbahasa Indonesia tentang vaksin Covid. Akurasi analisis sentimen justru berkurang 3,5%, sedangkan presisi, recall, dan F1-Score berkurang antara 0,02 hingga 0,04.

6. Saran

Saran untuk pengembangan atau penelitian selanjutnya bagi topik ini adalah penggunaan kamus bahasa gaul, bahasa daerah, dan singkatan kata sehari-hari dalam proses lemmatisasi, serta penambahan *corpus* singkatan kata pada kamus singkatan kata sehari-hari.

Daftar Pustaka

Elcholiqi, A. & Musdholifah, A. (2020). Chatbot in Bahasa Indonesia using NLP to Provide Banking Information. *Indonesian Journal of Computing and Cybermetrics Systems. 14(1)*. https://doi.org/10.22146/ijccs.41289

Floridi, L. & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequence. *Minds & Machines* **30**, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Ibrahim, M.O. & Budi, I. (2018). A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media. *Procedia Computer Science, Vol 135 p. 222-229.* 1877-0509 https://doi.org/10.1016/j.procs.2018.08.169

Kameswari, A.V.N. (2021). Image Caption Generator Using Deep Learning. *International Journal for Research in Applied Science and Engineering Technology Vol 9 (10) pp. 1554-1564.*

Lesson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. *International Journal of Qualitative Methods vol 18 p. 1-9.* https://doi.org/10.1177%2F1609406919887021

Lestari, C., Saputri, T.R.D., & Siahaan, S.C.P. (2022). Analisis Sentimen Pandangan Netizen Indonesia terhadap Vaksin COVID-19 menggunakan Multi-Layer Perceptron. *Jurnal Teknik Informatika dan Sistem Informasi vol. 9(4).*

Nayoga, B.P., Adipradana, R., Suryadi, R., & Suhartono D. (2021). Hoax Analyzer for Indonesian News Using Deep Learning Models. *Procedia Computer Science vol 179 p. 704-712.* 1877-0509 https://doi.org/10.1016/j.procs.2021.01.059

Nazeer, I., Rashid, M., Gupta, Dr. S., & Kumar, A. (2020). Use of Novel Ensemble Machine Learning Approach for Social Media Sentiment Analysis. *Analyzing Global Social Media Consumption,* IGI Global.

Ratnasari, C.I., Kusumadewi, S., & Rosita, L. (2014). Model Natural Language Processing untuk Perumusan Keluhan Pasien. *Seminar Nasional Informatika Medis (SNIMed) V.*

Salsabila, N.A., Winatmoko. Y,A,, Septiandri, A.A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *International Conference on Asian Language Processing (IALP) p. 226-229.* 10.1109/IALP.2018.8629151.

Zaky, D. (2019). Kumpulan Kata Bahasa Indonesia. Github: https://github.com/damzaky/kumpulan-kata-bahasa-indonesia-KBBI diakses pada tanggal 26 Juli 2022.