Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 4 (2020) 648 – 654 ISSN Media Elektronik: 2580-0760

Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia

Siti Khomsah[1], Agus Sasmito Aribowo[2]
[1] Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[2]Teknik Informatika, Fakultas Teknologi Industri, UPN Veteran Yogyakarta
[1]siti@ittelkom-pwt.ac.id , [2]sasmito.skom@upnyk.ac.id

Abstract

YouTube is the most widely used in Indonesia, and it's reaching 88% of internet users in Indonesia. YouTube's comments in Indonesian languages produced by users has increased massively, and we can use those datasets to elaborate on the polarization of public opinion on government policies. The main challenge in opinion analysis is preprocessing, especially normalize noise like stop words and slang words. This research aims to contrive several preprocessing model for processing the YouTube commentary dataset, then seeing the effect for the accuracy of the sentiment analysis. The types of preprocessing used include Indonesian text processing standards, deleting stop words and subjects or objects, and changing slang according to the Indonesian Dictionary (KBBI). Four preprocessing scenarios are designed to see the impact of each type of preprocessing toward the accuracy of the model. The investigation uses two features, unigram and combination of unigram-bigram. Count-Vectorizer and TF-IDF-Vectorizer are used to extract valuable features. The experimentation shows the use of unigram better than a combination of unigram and bigram features. The transformation of the slang word to standart word raises the accuracy of the model. Removing the stop words also contributes to increasing accuracy. In conclusion, the combination of preprocessing, which consists of standard preprocessing, stop-words removal, converting of Indonesian slang to common word based on Indonesian Dictionary (KBBI), raises accuracy to almost 3.5% on unigram feature.

Keywords: YouTube comments, sentiment analysis, text preprocessing, slang-word, N-Gram

Abstrak

YouTube merupakan media sosial yang paling banyak digunakan di Indonesia, mencapai 88 % dari seluruh pengguna internet di Indonesia. Komentar dalam bahasa Indonesia yang dihasilkan dari pengguna YouTube bertambah masif setiap hari, hal ini dapat digunakan untuk mengelaborasi polarisasi opini masyarakat terhadap kebijakan pemerintah. Tantangan utama dalam analisis opini adalah memproses *noise* terutama mengolah *stop words* dan bahasa slang. Penelitian ini bertujuan menerapkan berbagai model *preprocessing* pada analisis sentimen dari teks komentar YouTube, kemudian melihat pengaruhnya pada akurasi model *classifier*. Jenis-jenis *preprocessing* yang digunakan antara lain standar pemrosesan teks bahasa Indonesia, penghapusan *stop words* dan kata yang menyatakan subyek atau obyek, dan mengganti bahasa slang sesuai kosakata KBBI. Skenario pengujian dirancang untuk melihat pengaruh setiap jenis *preprocessing* pada akurasi model. Fitur yang digunakan adalah unigram dan kombinasi unigram-bigram. Ekstraksi fitur yang digunakan yaitu Count-Vectorizer dan TF-IDF-Vectorizer. Validasi menunjukkan penggunaan fitur unigram lebih baik akurasinya daripada kombinasi fitur unigram dan bigram. Sedangkan konversi kosakata slang mampu menaikkan akurasi. Penghapusan *stop words* juga menyumbang kenaikan akurasi. Kesimpulannya, kombinasi *preprocessing* yang terdiri *preprocessing* standar, penghapusan *stop words*, konversi kosakata slang menjadi kosakata standar KKBI menaikkan akurasi sekitar 3,5 % pada fitur term tunggal (unigram).

Kata kunci: komentar YouTube, sentiment analysis, text preprocessing, bahasa slang, N-Gram

1. Pendahuluan

Penelitian-penelitian analisis sentimen awalnya lebih banyak dilakukan pada teks twitter. Baru-baru ini, para peneliti mulai menganalisis sentimen pada komentar YouTube [1]–[4]. Data YouTube menarik diteliti karena YouTube menjadi media sosial paling populer di seluruh dunia. Data 2020 menunjukkan 88% warganet di Indonesia mengakses YouTube [5]. YouTube menjadi alternatif saluran transfer informasi bahkan setiap televisi di Indonesia mempunyai kanal YouTube untuk menjangkau warga yang lebih suka menonton melalui YouTube. Reaksi masyarakat terhadap program yang disiarkan dapat di *mining*. Hal ini menarik terutama pada siaran-siaran yang memberitakan kebijakan pemerintah. Reaksi warga Indonesia terhadap kebijakan pemerintah

Diterima Redaksi : 15 Juni 2020 | Selesai Revisi : 08 Juli 2020 | Diterbitkan Online : 20-08-2020

dapat digali dari komentar-komentar yang ditinggalkan pada kolom komentar setiap video terkait [6]. Pada situasi Pandemi COVID-19, ribuan komentar warga terhadap kebijakan pemerintah menjadi sumber data sangat berharga untuk menganalsis pendapat masyarakat terhadap kebijakan seperti PSBB (Pembatasan Sosial Berskala Besar), BLT (Bantuan Langsung Tunai), dan listrik gratis. Berdasarkan dataset komentar tersebut, polarisasi pendapat warga dapat diidentifikasi menggunakan metode analisis sentimen, yaitu metodologi untuk mengekstrak informasi dari data tidak terstruktur [7]. Penelitian sentimen analisis pada komentar YouTube juga sudah pernah dilakukan, antara lain dengan Deep Neural Network [4].

Tantangan utama dalam analisis sentimen adalah preprocessing. Kumpulan komentar pada YouTube mengandung banyak noise, antara lain *stop words* dan bahasa slang. *Stop words* adalah fitur kata yang tidak mengandung unsur sentimen, misalnya kata hubung "lagipula" "atau" "dari", "tetapi", "dan", dan sebagainya. Dalam tata bahasa Indonesia terdapat 16 macam kata penghubung. Selain kata penghubung, kata ganti orang, keterangan waktu, kata depan, dan kata-kata yang tidak mempunyai informasi bermakna juga masuk kategori stopword. Kamus stopword tidak tersedia baku sehingga memerlukan database indeks berisi daftar kata-kata *stop words* (*stopword list*). Beberapa peneliti telah membuat *stopword list* bahasa Indonesia antara lain Fadillah Z. Tala, Damian Doyle, dan Yudi Wibisono[8].

Pradana dan Hayaty [9] mencoba menghapus *stop words* dan kosakata slang saat *preprocessing* pada dataset Twitter, tetapi teknik ini tidak meningkatkan akurasi model secara siginifikan. Abidin, et al[10] menerapkan enam fitur N-Gram Word yang berbeda untuk meningkatkan pengklasifikasi KNN pada model analisis sentimen, dan ini meningkatkan akurasi menjadi 81,2%. Prahasiwi dan Kusumaningrum [11] mengolah kata-kata negasi menggunakan aturan POS Tagging yang dimodifikasi. Dengan klasifier Naive Bayes, akurasi model meningkat sebesar 3,3% [11]. Penelitian-penelitian tersebut hanya menangani stopword. Menggabungkan beberapa teknik dalam setiap langkah preprocessing mungkin akan meningkatkan model akurasi secara signifikan.

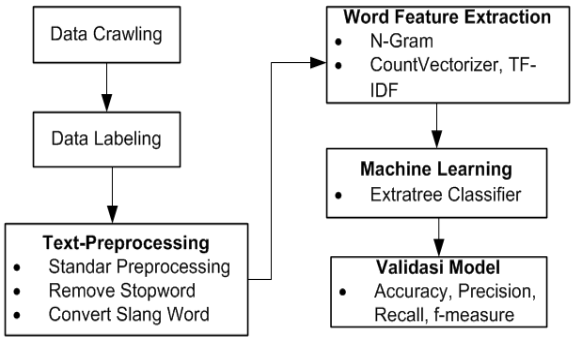
Penelitian ini mengusulkan hipotesis, yaitu akurasi model analisis sentimen akan meningkat jika kosakata slang diubah menjadi kosakata standar, dan *stop words* dihapus. Kata slang adalah kata yang tidak memenuhi standar kamus Indonesia (KBBI), biasanya dalam bentuk singkatan atau istilah gaul yang muncul di masyarakat. Istilah slang muncul hampir di setiap kalimat opini di media sosial. Tahap awal persiapan data ditemukan bahwa 88% kalimat dalam korpus penelitian ini mengandung kata slang. Setidaknya ada satu kata slang dalam setiap kalimat tersebut. Secara umum, warganet lebih suka menggunakan kata slang daripada kosakata standar. Konversi kosakata slang menjadi kosakata standar yang benar berdasarkan KBBI bisa menjadi salah satu langkah prapemrosesan yang dapat menghasilkan performa model *classifier* yang lebih baik.

Oleh karena itu, menangani kata slang juga merupakan tantangan dalam analisis sentimen. Maylawati, et al [12] mengembangkan porter stemmer untuk data teks bahasa Indonesia dengan menambahkan fungsi untuk memproses bahasa slang. Fungsi yang dikembangkan tersebut meningkatkan akurasi yang signifikan, dari 64,1% menjadi 88,7%. Hasil penelitian tersebut perlu diuji dalam domain dataset lain dengan jumlah dan jenis kosakata slang yang lebih beragam. Penelitian Ardi, et al [13] mengkonversi kosakata slang menggunakan kamus bahasa Indonesia. Deteksi kata slang menggunakan fungsi filter kamus slang. Penelitian tersebut juga menganalisis efek penggunaan fitur N-Gram, ekstraksi fitur TF-IDF pada model *classifier* sentimen analisis menggunakan SVM. Konversi kata slang dan pemilihan fitur unigram menghasilkan model dengan akurasi tertinggi 80,87%.

Berbagai *preprocessing* untuk menangani kata slang pada analisis sentimen masih belum menemukan *state of the art*-nya. Penelitian ranah ini masih terbuka lebar. Bagaimana jika semua metode preprocessing terlibat, mulai dari penerpaan preprocessing standar, penghapusan *stop words*, dan konversi kata slang. Preprocessing yang efektif perlu dibuktikan melalui eksperimen. Oleh karena itu, tujuan penelitian ini adalah menerapkan berbagai model *preprocessing* data komentar kemudian melihat pengaruhnya pada akurasi model analisis sentimen. Contoh sampel data yang digunakan adalah komentar pada video YouTube kebijakan listrik gratis di masa pandemi COVID-19.

2. Metode Penelitian

Tahapan penelitian mengikuti langkah-langkah standar proses pada analisis sentimen dilengkapi metode konversi kata slang, ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1. Data Crawling

Dataset penelitian ini merupakan teks komentar dalam bahasa Indonesia yang mengandung opini. Data tersebut diperoleh dari data komentar yang di-*crawling* dari kolom komentar pada video kebijakan pemerintah Indonesia terkait listrik gratis pada masa pandemi COVID-19. Komentar diunduh dari kolom komentar video yang diunggah dari Maret hingga Mei 2020, sebanyak 6096 komentar seperti Tabel 1. *Crawling* menggunakan algoritma yang dikembangkan peneliti dalam bahasa pemrograman Python.

Tabel 1. Sumber Data

No	URL Video and channel	
1	https://www.YouTube.com/watch?v=Z_poLb3Rzhg	
	Metro TV News	
	Tanggal upload	3 April 2020
	Jumlah viewer	460.196
	Tanggal crawling oleh peneliti	3 Mei 2020
	Jumlah komentar	1.882
2	https://www.YouTube.com/watch?v=ka3RDOWZeBc.	
	Tribunnews.com	
	Tanggal upload	31 Maret 2020
	Jumlah viewer	163.027
	Tanggal crawling oleh peneliti	3 Mei 2020
	Jumlah komentar	1.464
3	https://www.YouTube.com/watch?v=Mr57Vfikb0g.	
	Tribunnews.com	
	Tanggal upload	31 Maret 2020
	Jumlah viewer	85.838
	Tanggal crawling oleh peneliti	3 Mei 2020
	Jumlah komentar	750
4	https://www.YouTube.com/watch?v=tO1A3F4Wc9A.	
	Viva.co.id	
	Tanggal upload	31 Maret 2020
	Jumlah viewer	111.462
	Tanggal crawling oleh peneliti	3 Mei 2020
	Jumlah komentar	649
5	https://www.YouTube.com/watch?v=wZzvsbS-vXk.	
	Tribunnews.com	
	Tanggal upload	31 Maret 2020
	Jumlah viewer	231.820
	Tanggal crawling oleh peneliti	3 Mei 2020
	Jumlah komentar	1.351

Komentar-komentar tersebut diseleksi relevansinya dengan video-video tersebut. Komentar yang tidak relevan dengan video, misalnya percakapan antar user, pembicaraan diluar topik video, dan komentar yang tidak mengandung unsur opini tidak dipilih sebagai data dan tidak disertakan dalam proses berikutnya. Data yang sesuai sebanyak 3469 opini, memenuhi kriteria untuk proses selanjutnya.

2.2. Labeling

Kalimat-kalimat opini yang terpilih tersebut diberi label. Label opini terdiri sentimen positif dan negatif, seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Jenis Polaritas Sentimen

Polarity	Information	Symbol
Positive	Positive sentiment	1
Negative	Negative sentiment	-1

Semua komentar dalam dataset diberi label secara manual oleh dua *annotator* yang berpengalaman dalam penilaian sentimen. Pemberian label dilakukan dengan mengamati konteks keseluruhan kalimat dalam komentar tersebut. Kalimat komentar yang mengandung unsur kata-kata bermakna positif, diberi label positif. Sebaliknya kalimat mengandung kata-kata negatif, diberi label negatif. Jika kalimat mengandung unsur kata positif dan negatif maka dihitung berapa cacah kata bersentimen positif dan berapa cacah kata bersentimen negatif, kemudian polaritas yang dominan ditetapkan sebagai sentimen kalimat tersebut. Sentimen positif ditandai dengan kode 1, dan negatif kode -1. Contoh pemberian label ada pada Tabel 3.

Tabel 3. Pemberian Label

No	Teks	Label
1.	Terimakasih atas kebijakannya Pak Jokowi	1
2.	Yg pake token listrik di cuekin	-1
3.	saya sendiri tidak membutuhkan.potongan harga..(-1) keputusan itu sangat meringankan bagi masyarakat kecil...(1) terimakasih pak Jokowi (1)	1
4.	Terimakasih atas program listriknya pak...(1), tapi kami masih tidak puas..(-1)	-1

Contoh kalimat nomor satu dan dua pada Tabel 3 menunjukkan sebuah opini yang mempunyai polaritas tunggal. Pada opini pertama, kata ”terimakasih” mengandung unsur positif, sedangkan unsur kata bermakna negatif tidak terdapat didalamnya. Opini nomor dua terdapat kata ”cuekin” yang memberikan unsur negatif sehingga kalimat tersebut dianggap bersentimen negatif.

Opini yang terdiri beberapa kalimat atau paragraf seperti nomor 3 sangat mungkin mengandung sentimen majemuk. Pada opini seperti tersebut, perlu memisahkan paragraf tersebut menjadi kalimat-kalimat tunggal kemudian menentukan sentimennya masing-masing. Paragraf opini nomor tiga jika dipisahkan akan menghasilkan tiga kalimat. Pada kalimat pertama terdapat frasa ”tidak membutuhkan” yang memberikan unsur negatif. Namun kalimat selanjutnya terdapat frasa ”sangat menguntungkan” sehingga dalam opini tersebut juga terdapat sentimen positif. Sampai disini jenis sentimen seimbang yaitu satu positif dan satu negatif. Tetapi, kata ”terimakasih” pada kalimat ketiga menyebabkan sentimen positif pada opini nomor tiga bertambah menjadi dua. Karena unsur sentimen positif lebih banyak maka opini nomor tiga diberi label positif. Opini nomor empat terdapat sentimen positif dan negatif yang seimbang sehingga untuk menentukan sentimen yang dominan harus melihat konteks utama kalimat yaitu pada ungkapan ”tidak puas”. Maka opini nomor empat bersentimen negatif. Dari hasil proses pelabelan

diperoleh sebuah dataset opini yang terdiri atas 2258 komentar positif dan 1211 komentar negatif.

2.3. Text Preprocessing

Preprocessing mengacu metode standar yang digunakan dalam studi analisis sentimen pada teks bahasa Indonesia. Teknisnya terdiri dari empat langkah, yaitu penghapusan *stop words, case folding, tokenizing*, dan *stemming* [14]. Proses-proses ini telah diverifikasi dan menjadi standar umum dalam analisis sentimen [15], [16]. Secara umum, *preprocessing* dalam penelitian ini terdiri langkah- langkah berikut:

(1) *Stop words Removing.* Jika ditemukan kata yang masuk dalam daftar *stop words* maka dihapus. Daftar *stop words* menggunakan *stopwords list* Sastrawi.

(2)Konversi kosakata slang terdiri dari langkah-langkah berikut, (2.i) hapus karakter yang berulang-ulang secara berurutan sehingga tersisa satu karakter tunggal. Hal ini dilakukan karena kosakata slang biasanya mengandung banyak karakter berulang. Contoh, kata “siiiaaappppp” akan berubah menjadi "siap". "Waaaahhh" menjadi "wah", dan "maantuuulll" menjadi "mantul". (2ii) Hapus kata berisi satu karakter saja. Misalnya, kata yang terdiri dari satu karakter yaitu y, t, dan seterusnya.

(3)Konversi kosakata slang kedalam kosakata standar KBBI. Misalnya "guwee" dikonversi menjadi "saya", "eloo" dikonversi ke "kamu", "laen" dikonversi ke "lain", "pengen" dikonversi ke "ingin", "knp" dikonversi ke "kenapa", dan seterusnya. Pada proses konversi ini, peneliti membuat kamus slang berisi 4421 kata slang.

(4) Hapus subyek atau obyek. Contoh subyek adalah nama tokoh politik atau nama lembaga atau nama benda yang tidak memiliki unsur sentimen seperti: "jokowi", "prabowo", "maruf amin", "sandiaga", "pln", "listrik", " watt ”. *Pseudecode* untuk mengubah kata slang yang ditulis seperti berikut.

Algorithm for converting slang word

```
sub conv_slang(sentences,slang_dict)
newsentences←""
if length(sentences)>0
  for word in sentences.split()
     new_word←word
     for i ← 1 to length(slang_dict)
         if new_word==slang_dict[i]
           new_word←slang_dict[i]
           exit for
         endif
     next i
     newsentences←newsentences+new_word
  next word
end if
return newsentences

dict=[['slangword1','standardword1',
['slangword2','standardword2'],..]
sentence='sentences with slank word'
cleansentences=conv_slang(sentence,dict)
```

Algoritma dalam *pseudocode* di atas akan bekerja mengkonversi kosakata slang dalam satuan pengecekan per-kalimat. Kalimat yang mengandung kosakata slang akan dikirim ke fungsi *conv_slang* bersama dengan kamus slang *dict*. Pada fungsi *conv_slang*, kalimat yang dikirim akan dipisahkan per-kata, kemudian dicocokkan apakah terdaftar dalam kamus slang tersebut. Jika kalimat yang dikirim menemukan padanan kata dalam kamus slang maka akan diganti dengan kata baku sesuai dalam kamus slang. Kalimat baru yang sudah bersih akan dirangkai kembali di bagian akhir fungsi.

2.4. Word Feature Extraction

Kumpulan kata dalam dokumen seringkali disebut dengan *Bag of Word* (BoW) atau sekeranjang kata. Analisis dokumen teks harus mempertimbangkan pemilihan fitur. Studi ini memilih fitur N-Gram Word, N-Gram Word adalah jumlah kata yang merepresentasikan fitur tunggal. Sedangkan, ekstraksi fitur dari BoW menjadi vektor, menggunakan algoritma Count-Vectorizer dan TF-IDF-Vectorizer. Kedua algoritma ini juga akan dilihat, mana yang palling baik meningkatkan akurasi.

1) N-Gram Word

Hilangnya konteks adalah salah satu masalah dalam representasi (BoW) karena representasi BoW hanya berfokus pada kata-kata yang disajikan secara terpisah. BoW tidak memperhatikan keterkaitan suatu kata dengan kata sebelum dan sesudahnya dalam satu kalimat. Informasi semantik yang signifikan hilang ketika kalimat dipisahkan menjadi sekumpulan kata mandiri [12]. Dalam beberapa model, N kata-kata berturut-turut digunakan sebagai fitur. Dalam model bi-gram, N = 2, dua kata berturut-turut akan digunakan sebagai fitur dalam representasi vektor dokumen. Jelas bahwa sementara fitur N-gram memberikan konteks dan akibatnya hasil yang lebih baik dalam akurasi model analisis sentimen.

2). Count-Vectorizer

Count-Vectorizer mengubah BoW menjadi vektor. Cara kerjanya adalah mengekstrak kalimat-kalimat dalam dokumen ke dalam satu kata yang menyusunnya, dan menghitung seberapa sering setiap kata hadir dalam setiap dokumen. Setiap dokumen diwakili oleh vektor yang ukurannya sama dengan jumlah kosa kata, dan entri dalam vektor untuk dokumen tertentu menunjukkan jumlah kata dalam dokumen tersebut.

3). TF-IDF-Vectorizer

Penggunaan jumlah kata sebagai nilai fitur tidak mencerminkan pentingnya kata tersebut dalam dokumen. Pentingnya sebuah kata adalah nilai fiturnya. Kata dalam sebuah dokumen tidak hanya tergantung pada seberapa sering ia hadir dalam dokumen tersebut, tetapi juga bagaimana bobot kehadirannya terhadap keseluruhan dokumen yang digunakan sebagai dataset.

Gagasan tentang pentingnya sebuah kata dalam dokumen ditangkap oleh suatu skema yang dikenal sebagai algoritma pembobotan dokomen "*term-frequency inverse document frequency*" (TF-IDF). *Term frequency* (TF) adalah rasio dari jumlah kemunculan kata dalam dokumen dan IDF adalah kemunculan kata terhadap keseluruhan dokumen dalam database. Dengan demikian, TF-IDF adalah ukuran yang dinormalisasi yang mempertimbangkan panjang dokumen.

2.5. Machine Learning

Dataset dalam penelitian ini dikategorikan data tidak seimbang (*imbalanced data*). Hasil pelabelan dengan cara yang dijelaskan pada sub bagian 2.2, jumlah komentar positif sebanyak 2258 dan negatif 1211, hal ini menandakan perbandingan jumlah anggota kelas positif dan negatif tidak seimbang. Data yang tidak seimbang dapat mempengaruhi hasil klasifikasi. Para peneliti menangani data yang tidak seimbang menggunakan pendekatan *ensemble*, seperti *oversampling* atau *undersampling*. *Over* atau *undersampling* menyeimbangkan jumlah sampel positif dan negatif menjadi sama besar [17][18]. Namun, penanganan *imbalanced data* tidak selalu pada sampel. Algoritma *ensemble* bisa juga efektif, contohnya algoritma *extremily randomize tree* (Extra Trees).

Extra Trees *clasifier* adalah algoritma *ensemble* yang bekerja sangat baik untuk analisis sentimen mengungguli Adboost, KNN, Naive Bayes, Decision Tree dan Extra Trees mengakomodir data yang tidak seimbang [18]. Extra Trees merupakan *ensemble* dari algoritma Decision Tree. Algoritma Extra Trees bekerja dengan membuat pohon keputusan yang banyak berdasarkan dataset pelatihan, tanpa menerapkan *pruning*. Prediksi dibuat berdasar voting mayoritas kelas prediksi. Misalnya, mayoritas kelas prediksi setiap pohon adalah "True" maka keputusan akhir juga "True".

2.6. Validasi

Performa model analisis sentimen yang dibangun di diukur dengan matrik *confusion*. Matrik *confusion* berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh model *classifier*. Pengukuran menggunakan tingkat akurasi, presisi, recall, dan nilai F-measure, tujuannya menganalisis dampak setiap skenario *preprocessing* terhadap kinerja model analisis sentimen. Matriks *confusion* untuk prediksi dua kelas ditunjukkan pada

Tabel 3.

Akurasi adalah ukuran seberapa banyak peringkat sentimen benar dalam klasifikasi. Klasifikasi yang benar adalah true positive (TP) dan true-negative (TN). Akurasi menggunakan persamaan (1).

$$Akurasi = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})} \quad (1)$$

Tabel 3. Matrik Confusion

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negatif (FN)	True Negative (TN)

Presisi mengukur ketepatan classifier. Presisi tinggi berarti lebih sedikit false positive (FP), sedangkan presisi yang rendah berarti lebih banyak false positive (FP). Rumus presisi menggunakan persamaan (2).

$$Presisi = \frac{TP}{(TP+FP)} \quad (2)$$

Recall mengukur kelengkapan, atau sensitivitas *classifier*. Peningkatan nilai recall menunjukkan kemampuan model menemukan informasi yang relevan dari semua dataset. Rumus recall dalam persamaan (3).

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

Presisi dan Recall dapat digabungkan untuk menghasilkan metrik tunggal yang dikenal sebagai ukuran-F, yang merupakan rata-rata harmonik tertimbang dari presisi dan daya ingat. Rumus ukuran-F ada dalam persamaan (4). Pengukuran-F berguna sebagai akurasi.

$$f1 - measure = \frac{2\ x\ (Precision\ x\ Recall)}{(Precision+Recall)} \quad (4)$$

Implementasi Percobaan

Eksperimen ini menggunakan fitur unigram dan bigram. Ekstraksi fitur yang digunakan dalam penelitian ini adalah Count-Vectorizer dan TF-IDF Vectorizer. Setelah proses pelabelan selesai, data diproses oleh beberapa skenario *preprocessing* berikut.

1. SKENARIO 1. Tahap *preprocessing* tanpa menghapus *stop words*, kata slang tidak dikonversi ke kosakata standar, dan objek atau subyek tidak dihapus. Skenario ini bertujuan untuk menentukan model analisis sentimen akurasi tanpa ketiga *preprocessing* tersebut.
2. SKENARIO 2. Tahap *preprocessing* menggunakan proses penghapusan *stop words* tetapi tidak melakukan proses konversi kata slang dan menghapus subjek. Tujuannya adalah untuk mengetahui keakuratan jika hanya menghapus *stop words* dan tidak mengkonversi kosakata slang serta tidak menghapus subjek/objek.
3. SKENARIO 3. Tahap *preprocessing* menggunakan proses penghapusan *stop words*, mengubah kata-kata slang kedalam kosakata standar, tetapi tidak menghapus subjek. Tujuannya adalah untuk

mengetahui efek penghapusan *stop words* dan mengonversi kata slang pada akurasi model.

4. SKENARIO 4. Tahap *preprocessing* menggunakan penghapusan *stop words*, mengubah kata-kata slang kedalam kosakata standar, dan menghapus subjek/objek. Tujuannya adalah untuk mengetahui akurasi jika menggunakan semua metode *preprocessing* yang diusulkan.

3. Hasil dan Pembahasan

Model *classifier* dibangun menggunakan 80% data dari dataset, sedangkan 20% data digunakan untuk menguji model. Pengamatan dilakukan pada pengaruh empat skenario tahapan *preprocessing* dan penggunaan fitur N-gram terhadap performa model. Untuk pembandingan setiap tahapan skenario ini, penelitian hanya memakai satu machine learning saja yaitu Extra Tree Classifier dengan *estimator* (jumlah pohon) 100 karena berdasarkan pengujian memberikan akurasi tertinggi pada dataset tersebut.

Hasil pengujian ditunjukkan pada Tabel 4, Tabel 5, Tabel 6, dan Tabel 7. Kolom ”UNI” pada Tabel 4 -7 adalah simbol untuk fitur unigram, sedangkan kolom ”UNI + BI” berarti fitur unigram dan bigram.

Tabel 4. Skenario 1 (Tanpa Konversi Slang Word dan Hapus Stop Words dan Subyek/ Obyek Tidak Dihapus)

Parameter	CountVectorizer		TF-IDF	
	UNI	UNI+BI	UNI	UNI+BI
Accuracy	**0,856**	0,851	**0,856**	0,836
Precision	0,856	0,856	0,865	0,843
Recall	0,856	0,851	0,856	0,836
F-measure	0,859	0,853	0,859	0,838

Tabel 5. Skenario 2 (Standar Preprocessing dan Hapus Stop Words)

Parameter	CountVectorizer		TF-IDF	
	UNI	UNI+BI	UNI	UNI+BI
Accuracy	**0,870**	0,852	**0,858**	0,846
Precision	0,875	0,857	0,866	0,850
Recall	0,870	0,852	0,858	0,846
F-measure	0,871	0,854	0,861	0,850

Tabel 6. Skenario 3 (Standar Preprocessing, Hapus Stop Words, dan Konversi Kosakata Slang)

Parameter	CountVectorizer		TF-IDF	
	UNI	UNI+BI	UNI	UNI+BI
Accuracy	**0,885**	**0,879**	**0,884**	0,864
Precision	0,889	0,883	0,893	0,869
Recall	0,885	0,879	0,884	0,864
F-measure	0,886	0,880	0,886	0,865

Tabel 7. Skenario 4 (Standart Preprocessing, Hapus Stop words, Konversi Kosakata Slang, dan Hapus Subyek/Obyek)

Parameter	CountVectorizer		TF-IDF	
	UNI	UNI+BI	UNI	UNI+BI
Accuracy	**0,888**	0,883	**0,882**	**0,875**
Precision	0,894	0,887	0,890	0,881
Recall	0,888	0,883	0,882	0,875
F-measure	0,890	0,884	0,884	0,877

Parameter Accuracy menunjukkan semua skenario tahapan *preprocessing* mencapai akurasi lebih dari 85%. Nilai F-measure menunjukkan nilai tinggi, yang berarti tingkat ketepatan model memprediksi dan meghasilkan jawaban yang tepat sudah sangat baik.

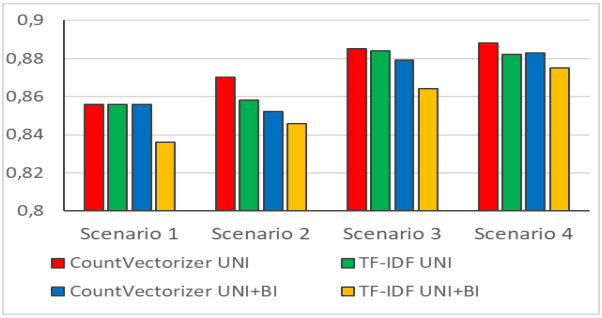
Skenario *preprocessing* model pertama mampu menghasilkan akurasi model *classfier* sebesar 85,5% (Tabel 4). Hal ini menjelaskan bahwa meskipun tidak menghapus *stop words*, tidak mengubah kosakata slang, dan tidak menghapus objek/subjek namun model prediksi ternyata lebih baik daripada penelitian serupa, yaitu model analisis sentimen pada komentar YouTube menggunakan Jaringan Saraf Tiruan [4].

Skenario kedua, seperti pada Tabel 5 menunjukkan akurasi yang lebih baik daripada percobaan pertama, meskipun hanya meningkat 1,4%, dan itu membuktikan bahwa *preprocessing* standar dan menghilangkan *stop words* dapat meningkatkan kinerja model. Sedangkan hasil skenario ketiga pada Tabel 6 yaitu mengubah kata slang dalam bentuk kata baku, meningkatkan akurasi Hal ini membuktikan bahwa mengubah kosakata slang efektif meningkatkan akurasi. Tabel 7 menunjukkan akurasi akan mencapai maksimal 88% jika menerapkan *preprocessing* standar, menghapus *stop words*, mengubah kosakata slang ke dalam kosakata standar KBBI, dan menghapus kata dalam bentuk subjek atau objek. *Preprocessing* yang lengkap ini mampu meningkatkan akurasi sebanyak 3.5% (dari 85,6% menjadi 88,8%) jika menggunakan ekstrkasi fitur Count-Vectorizer dengan unigram sebagaimana pada Tabel 8.

Tabel 8. Perbandingan Hasil Akurasi Setiap Jenis *Feature Extraction* dan *N-Gram*

Skenario	CountVectorizer UNI	TF-IDF UNI	CountVectorizer UNI+BI	TF-IDF UNI+BI
1	85,6%	85,6%	85,6%	83,6%
2	87,0%	85,8%	85,2%	84,6%
3	88,5%	88,4%	87,9%	86,4%
4	**88,8%**	88,2%	88,3%	87,5%

Percobaan dengan fitur N-Word memperlihatkan hasil bahwa fitur unigram lebih baik daripada kombinasi fitur unigram dan bigram. Grafik berjenjang pada Gambar 2 mengungkapkan bahwa setiap skenario meningkatkan akurasi 0,75 hingga 1%. Skenario keempat menunjukkan bahwa akurasi meningkat antara 3% hingga 3,5%. Hal ini membuktikan bahwa fitur unigram lebih baik dalam memberikan informasi yang relevan jika ekstraksi fitur mengunakan Count-Vectorizer daripada TF-IDF-Vectorizer, ditunjukkan oleh Gambar 2.



Gambar 2. Dampak Penerapan Skenario Preprocessing dan Fitur N-Gram Terhadap Kenaikan Akurasi

4. Kesimpulan

Penelitian ini menyimpulkan bahwa model *preprocessing* terbaik pada dataset yang dipakai dalam penelitian ini terdiri dari tiga langkah, yaitu menghapus *stop words*, mengubah kata slang menjadi kata baku berdasarkan kamus bahasa Indonesia (KBBI), dan menghilangkan kata berjenis subjek/objek. Memanfaatkan *preprocessing* tersebut akan meningkatkan akurasi cukup signifikan, setidaknya antara 3% hingga 3,5% jika menggunakan fitur unigram dan ekstraksi fitur Count-Vectorizer. Khusus pada dataset ini, jika *preprocessing* terbaik tersebut diterapkan maka akan mencapai akurasi sebesar 88,8%.

Disaranka skenario *preprocessing* terbaik tersebut diterapkan pada data lintas domain. Eksplorasi beberapa jenis *preprocessing* seperti menambahkan kosakata ke kamus kata slang dan membuat proses pelabelan secara otomatis dapat diteliti lebih lanjut.

Daftar Rujukan

[1] H. Bhuiyan, J. Ara, R. Bardhan, and Md. Rashedul Islam, "Retrieving YouTube Video by Sentiment Analysis on User Comment," in *International Conference on Signal and Image Processing Applications (IEEE ICSIPA)*, 2017, no. 1, pp. 474–478.

[2] M. Thelwall, "Social Media Analytics for YouTube Comments : Potential and Limitations," *International Journal of Social Research Methodology*, vol. 5579, no. October, pp. 1–14, 2017, doi: 10.1080/13645579.2017.1381821.

[3] A. Musdholifah and E. Rinaldi, "FVEC Feature and Machine Learning Approach for Indonesian Opinion Mining on YouTube Comments," in *Proceeding of EECSI*, 2018, pp. 724–729.

[4] A. A. L. Cunha, M. C. Costa, and M. A. C. Pacheco, "Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks," *International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pp. 561–570, 2019, doi: 10.1007/978-3-030-20912-4.

[5] D. H. Jayani, "Orang Indonesia Habiskan Hampir 8 Jam untuk Berinternet," *26 February 2020*, 2020. [Online]. Available: https://databoks.katadata.co.id/datapublish/2020/02/26/indonesia-habiskan-hampir-8-jam-untuk-berinternet. [Accessed: 20-Mar-2020].

[6] A. S. Aribowo *et al.*, "Systematic Literature Review : Sentiment And Emotion Analysis Techniques On Twitter Political Domain," *Opcion*, vol. 34, no. 86, pp. 2051–2060, 2018.

[7] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypoll Publisher, 2012.

[8] F. Rahutomo and A. R. T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," vol. 6, no. 1, 2019, doi: 10.25126/jtiik.201861226.

[9] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," *Kinetik*, vol. 4, no. 4, pp. 375–380, 2019, doi: 10.22219/kinetik.v4i4.912.

[10] T. F. Abidin, M. Hasanuddin, and V. Mutiawani, "N-grams Based Features for Indonesian Tweets Classification Problems," in *International Conference on Electrical Engineering and Informatics (ICELTICs)*, 2017, pp. 307–310.

[11] T. G. Prahasiwi and R. Kusumaningrum, "Implementation of negation handling techniques using modified syntactic rule in Indonesian sentiment analysis," *Journal of Physics: Conference Series*, vol. 1217, no. 1, 2019, doi: 10.1088/1742-6596/1217/1/012115.

[12] C. S. Dian Sa'adillah Maylawati, Wildan Budiawan Zulfikar, "An Improved of Stemming Algorithm for Mining Indonesian Text with Slang on Social Media," in *International Conference on Cyber and IT Service Management (CITSM)*, 2018, doi: 10.1109/CITSM.2018.8674054.

[13] H. L. Ardi, E. Sediono, and R. Kusumaningrum, "Support Vector Machine Classifier for Sentiment Analysis of Feedback Marketplace with a Comparison Features at Aspect Level," *International Journal of Innovative Research in Advanced Engineering*, vol. 4, no. 11, 2017.

[14] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," in *SENTIKA*, 2016, pp. 49–56, doi: ISSN: 2089-9815.

[15] A. S. Aribowo, H. Basiron, and N. S. Herman, "Systematic Literature Review : Sentiment And Emotion Analysis Techniques On Twitter Political Domain," *Opcion*, vol. 86, pp. 2051–2060, 2018.

[16] A. S. Aribowo, H. Basiron, N. S. Herman, and S. Khomsah, "Fanaticism Category Generation Using Tree-Based Machine Learning Method Fanaticism Category Generation Using Tree-Based Machine Learning Method," *Journal of Physics:Conference Series*, vol. 1501 01202, 2020, doi: 10.1088/1742-6596/1501/1/012021.

[17] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," in *ICSITech*, 2019, pp. 217–222.

[18] D. Tiwari and N. Singh, "Ensemble Approach for Twitter Sentiment Analysis," *I.J. Information Technology and Computer Science*, no. August, pp. 20–26, 2019, doi: 10.5815/ijitcs.2019.08.03.