Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics

JLK
Vol 5, No 1
Maret 2022

Perbandingan Penggunaan Kamus Normalisasi dalam Analisis Sentimen Berbahasa Indonesia

Firnanda Zuhad[*1], Nori Wilantika[*2]

[#] *Politeknik Statistika STIS*
Jl. Otto Iskandardinata No.64C Jakarta
[1] 221709702@stis.ac.id
[2]wilantika@stis.ac.id

***Abstrak*— Normalisasi merupakan salah satu tahapan *text preprocessing* dalam *Natural Language Processing*. Pengaruh normalisasi dengan kamus normalisasi dalam analisis sentimen berbahasa Indonesia belum diketahui. Dengan membandingkan data yang tidak dinormalisasi dan data yang dinormalisasi dari beberapa dataset, penelitian ini bertujuan untuk mengetahui bagaimana pengaruh normalisasi dan pengaruh kamus yang digunakan dalam analisis sentimen. Kamus yang digunakan pada penelitian ini antara lain *Colloquial Indonesian Lexicon* dan modul formalizer pada INANLP. Sebagai kontrol, dalam analisis sentimen, metode klasifikasi Multinomial Naïve Bayes diterapkan pada seluruh dataset. Performa dari delapan dataset kemudian dianalisis dan diuji secara statistik. Akurasi, presisi, dan recall diuji dengan menggunakan uji wilcoxon signed rank test untuk menentukan apakah normalisasi mampu meningkatkan performa dari analisis sentimen. Dari hasil uji hipotesis diperoleh bahwa hasil uji keseluruhan dari performa kamus menunjukkan nilai p-value kurang dari 0,05. Dengan begitu, normalisasi pada analisis sentimen berbahasa indonesia tidak menyebabkan kenaikan performa dari indikator akurasi, presisi, dan recall.**

***Kata Kunci*— Normalisasi, Lexicon, *Colloquial Indonesian Lexicon*, InaNLP, Wilcoxon Signed Ranks Test**

I. LATAR BELAKANG

Penggunaan media sosial dalam berkomunikasi telah menjadi kegiatan sehari-hari setiap orang di masa perkembangan teknologi yang begitu mutakhir. Pengguna membutuhkan sebuah gawai yang tersambung dengan internet agar bisa bersosialisasi dalam media sosial. Pada tahun 2020, Jumlah pengguna gawai di Indonesia sebanyak 81,87 juta orang [1]. Dengan berkomunikasi melalui media sosial, pengguna mampu mengekspresikan perasaan atau mengungkapkan opini terhadap topik tertentu. Salah satu media sosial yang sering dipakai dalam berkomunikasi adalah Twitter. Saat ini terdapat 13,2 juta pengguna Twitter di Indonesia [2].

Data twit yang berisi opini yang diunggah di Twitter dapat dimanfaatkan untuk analisis sentimen seperti yang dilakukan oleh [3], [4], dan [5]. Dengan analisis sentimen, kita dapat mencari tahu bagaimana opini individu dari sekumpulan data twit pengguna Twitter. Analisis sentimen menganalisis opini pengguna media sosial terhadap suatu topik tertentu sehingga dapat diketahui dominasi antara positif dan negatif dalam twit pengguna [6].

Sebelum melakukan analisis sentimen, data teks yang telah diperoleh harus dilakukan *preprocessing* terlebih dahulu agar data menjadi lebih terstruktur dan tertata sesuai format algoritma yang akan digunakan untuk analisis. Salah satu tahapan *preprocessing* adalah normalisasi, yaitu mengganti kata yang tak baku dengan konteks kata yang sesuai [7].

Kata tak baku biasa digunakan pada platform media sosial [8]. Pada umumnya, 85% twit di Twitter mengandung *slang* atau kata tak baku [9]. Contoh penggunaan kata tak baku dalam Bahasa Indonesia adalah *gue*(saya) dan *gaa*(tidak) [9]. Dalam bahasa Indonesia, penggunaan kata tak baku tersebut umunya disebut dengan bahasa "alay".

Kata tak baku dalam media sosial kebanyakan tidak ditemukan dalam kamus, sehingga kata tersebut tidak bisa diturunkan secara morfologis dari kata yang ada di kamus [7]. Kata tak baku dalam media sosial sering kali tidak sesuai untuk digunakan sebagai data dalam tugas *Natural Language Processing* seperti *Machine Translation, Information Retrieval,* dan *Opinion Mining* karena ketidakteraturan dari corak bahasanya [10]. Penelitian oleh [10], menunjukkan bahwa normalisasi dapat meningkatkan performa pada pemrosesan data berbahasa inggris. Sedangkan pada penelitian normalisasi dalam bahasa indonesia yang dilakukan oleh [11], menunjukkan bahwa normalisasi tidak memberikan pengaruh yang signifikan terhadap performa. Namun sayangnya penelitian yang diambil oleh penelitian tersebut hanya berdasarkan ujicoba pada satu dataset saja. Penelitian tersebut belum menguji coba pengaruh normalisasi dalam beberapa dataset sehingga penelitian ini bermaskud untuk menguji lebih lanjut apakah normalisasi memberikan pengaruh yang signifikan terhadap performa analisis teks berbahasa indonesia.

Submitted 05-11-2021; Revised 21-02-2022; Accepted 10-03-2022

Kata yang tak baku dapat dinormalisasi dengan menggunakan kamus *Colloquial* atau kamus normalisasi. Kamus yang tersedia untuk normalisasi teks berbahasa Indonesia antara lain Kamus *Colloquial Indonesian Lexicion* [11] dan Kamus Alay dalam InaNLP (*Indonesian Natural Language Processing*) [12]. Walaupun demikian, tidak semua kamus memiliki efektivitas dan performa yang sama. Kamus bahasa tak baku yang telah dibuat oleh penelitian sebelumnya memiliki kelebihan dan kekurangannya masing-masing sehingga perlu dilakukan pengujian untuk melihat efektivitas kamus itu sendiri maupun perbandingan performa antar kamus yang tersedia.

Penelitian ini bertujuan untuk menguji pengaruh normalisasi pada analisis sentimen dalam bahasa indonesia dengan cara membandingkan performa data yang tidak dinormalisasi dengan data yang dinormalisasi menggunakan kamus *Colloquial Indonesian Lexicon* dan Kamus pada INANLP. Untuk melihat performa kedua kamus tersebut, kami mengujinya pada beberapa dataset kemudian diproses lebih lanjut menggunakan analisis sentiment sehingga performa kedua kamus dapat dibandingkan. Dari perbandingan tersebut akan diambil kesimpulan apakah terdapat pengaruh yang signifikan dari penggunaan kedua kamus tersebut dalam normalisasi pada analisis sentimen. Selain membandingkan pengaruh (efektivitas kamus), kedua kamus dapat dibandingkan performanya untuk mencari tahu mana yang terbaik antara dua kamus tersebut.

II. Penelitian Terkait

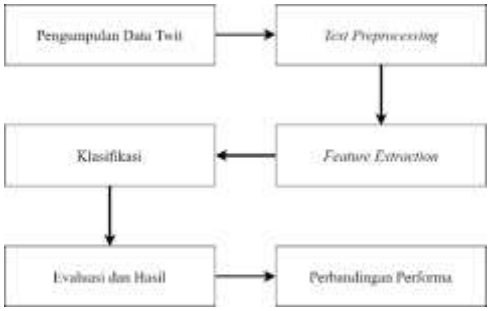
Pada penelitian sebelumnya yang dilakukan oleh [11], peneliti membangun sebuah kamus untuk normalisasi dalam Bahasa Indonesia. Peneliti memberi nama kamus tersebut dengan nama "Kamus Alay". Kata pada Kamus Alay dikumpulkan dari komentar Instagram. Penelitian tersebut bertujuan menguji pengaruh normalisasi dengan Kamus Alay dengan cara membandingkan F1-score dari kombinasi algortima sekumpulan fitur teks dengan algoritma Xgboost dan SVM. Dari uji evaluasi kamus tersebut, peneliti menunjukkan bahwa tidak ada perbedaan yang signifikan antara nilai F1-score dalam data yang dinormalisasi maupun yang tidak. Kesimpulan penelitian menyebutkan bahwa diperlukan penelitian lebih lanjut untuk memanfaatkan kamus dalam kasus yang lain seperti analisis sentimen.

Normalisasi pada Bahasa Indonesia bisa juga dilakukan dengan menggunakan kamus pada *toolkit* bernama InaNLP(Indonesian *Natural Language Processing*) yang dikembangkan oleh [12]. Penelitian tersebut bertujuan untuk membangun dan menguji sebuah *toolkit* yang bisa secara langsung digunakan dalam studi *natural language processing*. *Toolkit* yang dibuat dalam penelitian tersebut memiliki beberapa modul salah satunya adalah *word normalization*. Kami akan memakai kamus yang digunakan dalam modul tersebut untuk sebagai kamus kedua normalisasi.

Penelitian mengenai penggunaan algoritma multinomial naïve bayes telah dilakukan oleh [13], [14], dan [15]. Dalam penelitian tersebut, peneliti menggunakan algoritma multinomial naïve bayes untuk mengklasifikasikan sentimen positif dan negatif pada analisis sentimen. Metode naïve bayes dengan model multinomial inilah yang akan digunakan dalam penelitian ini.

Untuk membandingkan akurasi antara dua tipe data yang dihasilkan maka kami mengacu kepada penelitian oleh [16], tentang perbandingan antara dua *classifier* yang menyimpulkan apakah terdapat perbedaan yang signifikan antara dua *classifier* algoritma. Hasilnya menunjukkan bahwa Wilcoxon Signed-Ranks Test memberikan nilai p-value yang lebih kecil dibandingkan dengan uji lainnya sehingga uji tersebut akan cenderung menolak *null hypothesis* dibandingkan dengan uji yang lain. Walaupun begitu, uji Wilcoxon ini hanya bisa dipakai saat uji asumsi Paired t-test tidak terpenuhi. Pada penelitian ini, uji Wilcoxon Signed-Rank Test akan digunakan sebagai analisis pembanding antara hasil analisis sentimen dari data yang dinormalisasi dengan yang tidak dinormalisasi dan performa antara dua kamus.

III. Metode Penelitian



Gambar. 1 Alur tahapan penelitian

Dari Gambar. 1, tahapan penelitian dapat dijelaskan secara rinci sebagai berikut:

A. Pengumpulan Data Twit

Penelitian ini Tahapan awal penelitian adalah pengumpulun data dari platform media sosial yaitu Twitter. Data yang dikumpulkan berupa twit dari pengguna twitter dengan menggunakan *package* Twint dalam bahasa pemograman python. Data twit yang dikumpulkan adalah twit dengan kata kunci vaksin & indonesia. Dari data twit tersebut, diperoleh satu dataset dengan pengambilan data pada rentang waktu 1 minggu pada minggu pertama bulan Januari 2021. Karena dataset yang dibutuhkan lebih dari satu maka dataset yang lain menggunakan dataset dari penelitian sebelumnya. Dataset yang akan digunakan adalah data yang telah diberikan label positif ataupun negatif yang bersumber dari skripsi dan github.

B. Text preprocessing

Tahapan kedua pada penelitian ini adalah *text preprocessing*. Data twit akan di-*preprocessing* dengan beberapa langkah yaitu lowercasing, *cleaning*, menghapus whitespace, menghapus duplikat, tokenisasi, normalisasi,

menghapus stopword, dan stemming. Keseluruhan proses tersebut berfungsi mengurangi *noise* pada data dan menghasilkan data yang siap pakai untuk proses selanjutnya. Dengan berkurangnya *noise* maka diharapkan performa dari analisis akan mengalami kenaikan.

C. Feature Extraction

Data yang sudah di-*preprocessing* akan diproses lebih lanjut dengan *feature extraction* agar data yang berbentuk teks(*string*) diubah menjadi vektor. Data teks perlu diubah menjadi vektor agar bisa dianalisis dengan kalkulasi multinomial naïve bayes karena data teks tidak dapat secara langsung diproses. Pada proses ini menggunakan *package* sklearn dengan fungsi TfidfVectorizer.

D. Klasifikasi

Data yang sudah diubah dari teks menjadi vektor akan dijadikan sebagai input dalam algoritma naïve bayes. Metode klasifikasi yang digunakan pada analisis sentimen adalah metode naïve bayes *classifier* model multivariate [17]. Model seleksi data menggunakan metode kfold dengan nilai k=5. Dengan metode ini, setiap data twit yang telah siap dianalisis maka data yang telah diberi label dibagi menjadi 2 bagian yaitu data training dan testing. Data tranning yang berisi pasangan twit dan label akan digunakan sebagai acuan sumber pembentukan model. Probabilitas kemunculan kelas positif dan negatif dihitung dari setiap fitur yang mempresentasikan twit. Ketika mengklasifikasikan twit yang baru, maka tinggal mengalikan nilai probabilitas setiap fitur untuk masing-masing kelas. Nilai probabilitas terbesar diantara kelas positif dan negatif akan dijadikan dasar penentu kelas twit baru tersebut.

E. Evaluasi dan Hasil

Tahapan selanjutnya adalah evaluasi hasil performa algoritma klasifikasi yang digunakan. *Confusion matrix* digunakan untuk mengevaluasi performa dari algoritma naïve bayes. Dari *Confusion matrix* tersebut, akan didapatkan tiga indikator performa yaitu akurasi, presisi, dan recall dari masing masing dataset dan kategori dataset [18]. Setelah didapatkan indikator performa tersebut, kita dapat mengetahui seberapa baik model yang digunakan dalam penelitian ini. Walaupun begitu, hal tersebut bukanlah fokus utama dalam penelitian ini melainkan perbandingan performa kamus.

F. Perbandingan Performa

Tahapan terakhir adalah membandingkan performa yang dihasilkan dari masing-masing kategori dataset. Setelah analisis sentimen mengeluarkan hasil performa berupa nilai akurasi, presisi, dan recall dari setiap dataset dengan kamus maupun tanpa kamus, maka untuk analisis seberapa besar efektivitas kamus menggunakan uji Wilcoxon Signed-Ranks. Untuk menguji efektivitas kamus, Hipotesis nol untuk uji ini adalah bahwa performa analisis data yang tidak dinormalisasi sama dengan data yang dinormalisasi. Sedangkan Hipotesis alternatifnya adalah bahwa performa analisis data yang tidak dinormalisasi tidak sama dengan data yang dinormalisasi. Untuk membandingkan performa kedua kamus, Hipotesis nol untuk uji ini adalah bahwa performa analisis data yang dinormalisasi dengan kamus Colloquial Indonesian Lexicon sama dengan kamus InaNLP. Sedangkan Hipotesis alternatifnya adalah bahwa performa analisis data yang dinormalisasi dengan kamus Colloquial Indonesian Lexicon tidak sama dengan kamus InaNLP.

IV. HIPOTESIS PENELITIAN

Hipotesis dalam uji wilcoxon signed-rank dari penelitian adalah sebagai berikut.

A. Data Yang Tak Dinormalisasi Dengan Data Yang Dinormalisasi Dengan Kamus Normalisasi Satu

Indikator Akurasi

H_0: $M_0 = M_1$

H_1: $M_0 \neq M_1$

M_0 = nilai median akurasi dari data yang tidak dinormalisasi

M_1= nilai median akurasi dari data yang dinormalisasi dengan kamus satu

Indikator Presisi

H_0: $M_0 = M_1$

H_1: $M_0 \neq M_1$

M_0 = nilai median presisi dari data yang tidak dinormalisasi

M_1= nilai median presis dari data yang dinormalisasi dengan kamus satu

Indikator Recall

H_0: $M_0 = M_1$

H_1: $M_0 \neq M_1$

M_0 = nilai median recall dari data yang tidak dinormalisasi

M_1= nilai median recall dari data yang dinormalisasi dengan kamus satu

B. Data Yang Tak Dinormalisasi Dengan Data Yang Dinormalisasi Dengan Kamus Normalisasi Dua

Indikator Akurasi

H_0: $M_0 = M_2$

H_1: $M_0 \neq M_2$

M_0 = nilai median akurasi dari data yang tidak dinormalisasi

M_2= nilai median akurasi dari data yang dinormalisasi dengan kamus dua

Indikator Presisi

H_0: $M_0 = M_2$

H_1: $M_0 \neq M_2$

M_0 = nilai median presisi dari data yang tidak dinormalisasi

M_2= nilai median presisi dari data yang dinormalisasi dengan kamus dua

Indikator Recall
H_0: $M_0 = M_2$
H_1: $M_0 \neq M_2$
M_0 = nilai median recall dari data yang tidak dinormalisasi
M_2= nilai median recall dari data yang dinormalisasi dengan kamus dua

C. Data Yang Dinormalisasi Dengan Kamus Normalisasi Satu Dengan Data Yang Dinormalisasi Dengan Kamus Normalisasi

Indikator Akurasi
H_0: $M_1 = M_2$
H_1: $M_1 \neq M_2$
M_1 = nilai median akurasi dari data yang dinormalisasi dengan kamus satu
M_2= nilai median akurasi dari data yang dinormalisasi dengan kamus dua

Indikator Presisi
H_0: $M_1 = M_2$
H_1: $M_1 \neq M_2$
M_1 = nilai median presisi dari data yang dinormalisasi dengan kamus satu
M_2= nilai median presisi dari data yang dinormalisasi dengan kamus dua

Indikator Recall
H_0: $M_1 = M_2$
H_1: $M_1 \neq M_2$
M_1 = nilai median recall dari data yang dinormalisasi dengan kamus satu
M_2= nilai median recall dari data yang dinormalisasi dengan kamus dua.

V. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Pengumpulan data dibagi menjadi dua bahasan yaitu sebagai berikut:

1. Dataset dalam Penelitian

Pada tahap paling awal penelitian yaitu pengumpulan data telah berhasil mengumpulkan delapan dataset pada tabel 5 dengan rincian jumlah data sebelum di-*preprocessing* sebagai berikut:

TABEL I
DESKRIPSI DATASET

No.	Sumber	Penulis	Keterangan Dataset
1.	Pengumpulan Sendiri	-	Data twit dengan kata kunci vaksin dan indonesia dengan rentang waktu 1 Januari 2021 sampai dengan 8 Januari 2021
2.	Skripsi	Silvia Ni'matul Maula [19]	Data twit dengan kata kunci menggunakan #RUUPKS dengan rentang waktu 1 Januari 2018 sampai degan 31 Maret 2019
3.	Skripsi	Bambang Dwi Putra Nugraha [20]	Data twit dengan kata kunci "Ujian Nasional"
4.	Skripsi	Muhammad Firdaus [21]	Data twit dengan kata kunci "harga cabe" dan "harga cabai" dengan rentang waktu 1 Januari 2018 sampai 31 Desember 2018
5.	Github	Dhika Rangga [22]	Data twit tentang ujaran kebencian dari beberapa kata kunci dengan rentang waktu 20 Maret 2018 sampai 10 September 2018
6.	Github	Yahdi Indrawan [23]	Data twit tentang sentimen covid-19 di Indonesia
7.	Github	Rio Chandra [24]	Data twit tentang sentimen debat kedua pilkada DKI Jakarta
8.	Github	Ridi Ferdiana, Fahim Jatmiko, Desi Dwi Purwanti, Artmita Sekar Tri Ayu, Wiliam Fajar Dicka [25]	Data twit September sampai Desember 2018

Dari Tabel 1, dataset yang dikumpulkan terdiri dari satu data yang dikumpulkan sendiri, tiga dataset dari skripsi mahasiswa Politeknik Statistika STIS, dan empat data dari internet yang bersumber dari github.

TABEL II
DESKRIPSI JUMLAH DATA DALAM DATASET

No.	Nama Penulis	Jumlah Data
1.	-	3665
2.	Silvia Ni'matul Maula [19]	560
3.	Bambang Dwi Putra Nugraha [20]	1491
4.	Muhammad Firdaus [21]	2290
5.	Dhika Rangga [22]	14210
6.	Yahdi Indrawan [23]	636
7.	Rio Chandra [24]	1506
8.	Ridi Ferdiana, dkk [25]	5164

Dari Tabel 2, dataset 1 dikumpulkan dengan cara *scraping* dari twit dengan *package* twint. Dari *package* ini dapat dibuat fungsi untuk mengatur kata kunci, bahasa, rentang waktu, jenis file output, dan kostumisasi variabel dari target twit yang akan di-*scraping*. Dengan kata kunci vaksin dan indonesia menunjukkan bahwa setiap twit yang mengandung kata vaksin serta indonesia yang diunggah dari tanggal 1 Januari 2021 sampai 8 Januari 2021 dalam bahasa Indonesia akan dicetak ke dalam data csv dengan variabel output berupa nomor id, tanggal dan waktu, nama pengguna, isi twit, dan hashtag yang digunakan. Walapun demikian yang digunakan dalam proses selanjutnya hanya variabel twit.

TABEL III
CONTOH DATA TWIT DALAM DATASET 1

Twit	Hashtags
KEREN, pak @erickthohir benar memastikan kesiapan vaksin covid utk rakyat Indonesia, kita jg harus bersatu dukung kerja pemerintah dlm menyelesaikan pandemi ini, Masker, Mencuci tangan, menjaga jarak #pkpi #pkpindonesia #partaizamanWOW #indonesiabisalawancorona #coronagoaway	['pkpi', 'pkpindonesia', 'partaizamanwow', 'indonesiabisalawancorona', 'coronagoaway']
Yth. Pak @jokowi jika @MajelisUlamaID lamban & ogah2an terbitkn Sertifikasi,, Vaksin Corona yg sdh ada, mohon dahulukan saja Sdra2 kami Non Muslim yg juga sgt berhak sgra diobati. Kami ikhlas. @MajelisUlamaID agaknya lbh suka muslim Indonesia punah sblm waktunya.	[]

Data yang berhasil dikumpulkan pada proses *scraping* berjumlah 3665 data. Contoh data yang berhasil dikumpulkan ada pada Tabel 3. Pada Tabel 3, pada kolom *hashtags* "[]" menunjukkan bahwa twit tidak memiliki *hashtags* sama sekali.

2. Kamus dalam Penelitian

Rincian tentang jumlah kata yang terdapat dalam kamus adalah sebagai berikut:

TABEL IV
DESKRIPSI KAMUS

Nama Kamus	Penulis	Sumber Pengambilan	Disebut sebagai
Colloquial Indonesian Lexicon	Salsabila, Winatmoko, Septiandri, dan Jamal [11]	https://github.com/nasalsabila/kamus-alay	Kamus normalisasi Satu
Kamus Alay	Purwarianti, Andhika, Wicaksono, Afif, dan Ferdian [12]	https://github.com/panggi/pujangga	Kamus normalisasi dua

Dari Tabel 4, kedua kamus diperoleh dari internet yaitu dari github dalam bentuk format csv. Kamus *Colloquial Indonesian Lexicon* akan disebut sebagai kamus normalisasi satu. Sedangkan kamus INANLP akan disebut sebagai kamus normalisasi dua.

TABEL V
DESKRIPSI JUMLAH KATA DALAM KAMUS

No.	Nama Kamus	Jumlah Kata
1.	*Colloquial Indonesian Lexicon*	15006
2.	Kamus Alay	1147

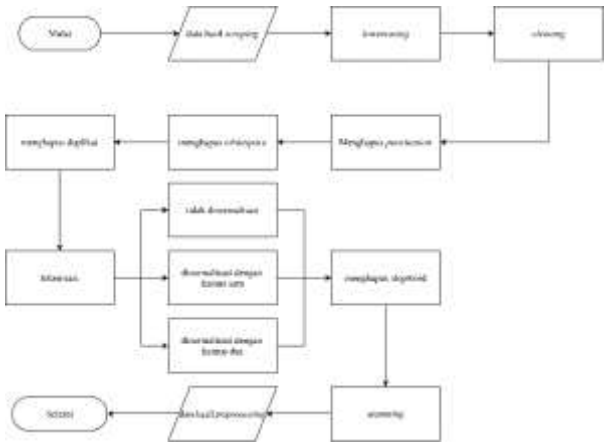
Dari Tabel 5, jumlah kata yang terdapat pada kamus 1, pada kamus yang diunduh dari github memiliki duplikat kata. Oleh karena itu duplikat kata yang ada di kamus dihapus sehingga jumlah kata yang *unique* dalam kamus sejumlah 13380 kata. Sedangkan pada kamus 2, semua kata dalam kamus telah *unique*. Dari sisi jumlah kata, kamus 1 lebih banyak daripada kamus 2 menunjukkan bahwa semakin banyak kata yang akan dinormalisasi dengan kamus 1.

B. Text preprocessing

Seperti yang dapat dilihat pada Gambar. 2, alur *preprocessing* dijelaskan secara detail sebagai berikut:

1. Lowercasing

Lowercasing dilakukan dengan menggunakan fungsi lower() dari tipe *dataframe* dimana setiap data bertipe string pada dataset akan diubah menjadi huruf kecil.



Gambar. 2 Flowchart tahapan pada proses *preprocessing*

2. Cleaning

Pada proses *cleaning*, ada beberapa elemen yang dihapus yaitu emoji, *emoticon, mentions, hashtag, retweet*, angka, dan *hyperlink*. Emoji adalah karakter khusus untuk menghasilkan *Mentions* adalah nama pengguna yang ditandai dengan karakter "@" pada awal kata dan diikuti dengan huruf, angka maupun karakter khusus seperti " _ ". *Hashtag* adalah penanda yang diawali dengan karakter "#" yang biasanya berkaitan dengan topik twit. *Retweet* adalah elemen dalam twit yang ditandai dengan kata rt yang maksudnya adalah memposting ulang twit dari akun pengguna lain. Angka pada data teks adalah elemen penganggu pada proses *text processing*. *Hyperlink* adalah alamat sebuah web yang ditandai dengan elemen "http:/" pada awal kata. *Package* yang digunakan dalam proses *cleaning* teks adalah *package regular expression*. Dengan memanfaatkan fungsi sub(), maka elemen string yang akan dihapus pada data dapat disubstitusi dengan elemen lain yang biasanya adalah spasi(" ").

3. *Menghapus Punctuation*

Pada proses ini, semua bentuk tanda baca seperti ".", ",", "!", dan sebaganya akan dihapus dari data. Karena elemen ini tidak dibutuhkan pada analisis. Dengan menggunakan fungsi replace() untuk mensubtitusi elemen pada data string. Apabila terdapat elemen tanda baca maka akan diganti dengan spasi(" ").

4. Menghapus Whitespace

Pada proses menghapus *whitespace*, *whitespace*(spasi) yang dihasilkan pada tahapan *cleaning* sebelumnya akan dihapus karena hanya akan menambah elemen yang tak diperlukan pada data. Proses ini dilakukan dengan menggunakan fungsi strip() dalam *package* string.

5. *Menghapus Duplikat*

Data yang telah dieksekusi pada tahap-tahap sebelumnya akan dicek apakah terdapat data twit yang terduplikasi. Jika ada data twit yang terduplikasi maka seluruh data yang terduplikasi akan dihapus dengan menyisakan satu data saja. Proses ini menggunakan fungsi drop_duplicates().

6. *Tokenisasi*

Pada tahap ini, setiap kata dalam kalimat pada setiap data twit akan diubah menjadi token-token(kata). Indikator untuk memisahkan kata dalam kalimat adalah *whitespace*(" ") dalam kalimat. Proses ini menggunakan fungsi word_tokenize() dalam *package* NLTK. Setiap baris data pada data twit akan dieksekusi dengan fungsi word_tokenize().

7. *Normalisasi*

Tahap yang paling penting dalam penelitian ini adalah tahap normalisasi dengan kamus yaitu kamus *Colloquial Indonesian Lexicon* dan kamus pada INANLP yang didapatkan dari github. Proses ini menginput kedua kamus tersebut kedalam coding python sebagai data dengan tipe *dictionary*. Konsep proses normalisasi dengan data tipe *dictionary* ini adalah apabila pada setiap token data twit terdapat kata pada kamus maka kata tersebut akan diubah menjadi kata pada bentuk bakunya. Dataset akan dibagi menjadi tiga versi pada tahap ini yaitu data yang tidak dinormalisasi, data yang dinormalisasi dengan kamus *Colloquial Indonesian Lexicon*, dan data yang dinormalisasi dengan kamus INANLP.

8. Menghapus Stopword

Proses ini adalah salah satu proses penting dalam *preprocessing*. *Stopword* adalah kata yang tidak dibutuhkan dalam analisis dan kemunculannya hanya akan memperburuk performa sehingga perlu dihapus. Daftar *stopword* yang dihapus dalam penelitian ini adalah daftar *stopword* yang ada di *package* Sastrawi. Proses ini menggunakan fungsi StopWordRemoverFactory() dari *package* Sastrawi.

9. Stemming

Proses *stemming* adalah proses mengubah kata pada data twit yang memiliki imbuhan menjadi kata dasar. Proses ini menggunakan fungsi StemmerFactory() pada *package* Sastrawi.

Keseluruhan proses *preprocessing* dapat diringkas pada Tabel 6.

TABEL VI
PERBANDINGAN SEBELUM DAN SESUDAH DATA *PREPROCESSING*

Sebelum Preprocessing	**Sesudah Preprocessing**
KEREN, pak @erickthohir benar memastikan kesiapan vaksin covid utk rakyat Indonesia, kita jg harus bersatu dukung kerja pemerintah dlm menyelesaikan pandemi ini, Masker, Mencuci tangan, menjaga jarak #pkpi #pkpindonesia #partaizamanWOW #indonesiabisalawancorona #coronagoaway	keren pak benar pasti kesiap vaksin covid rakyat indonesia satu dukung kerja perintah selesai pandemi masker cuci tangan jaga jarak

B. Pembangunan Model Klasifikasi Sentimen

1. Feature Extraction

Pada tahap ini, fitur dari data yang telah di-*preprocessing* sebelumnya akan diekstrak sehingga data yang berbentuk *string* tersebut akan dikonversi menjadi sebuah vektor dengan metode TFIDF. Dalam penelitian ini model bahasa n-gram yang dipakai adalah model unigram. Proses ini menggunakan fungsi TfidfVectorizer.

2. Model Multinomial Naïve Bayes

Proses ini membuat sebuah model klasifikasi dengan multinomial naïve bayes dengan menggunakan data twit *trainning* dan label *trainning*. Dengan menggunakan fungsi fit dalam *package* MultinomialNB dalam sklearn.

C. Performa Kamus Normalisasi Satu

1. Persentase Kata Tak Baku Dalam Dataset berdasarkan kamus normalisasi satu

TABEL VII
PERSENTASE KATA TAK BAKU DALAM DATASET MENURUT KAMUS NORMALISASI SATU

No.	Dataset	Jumlah Kata Total	Jumlah Kata Tak Baku	Persentase
1.	Dataset 1	44331	3018	6,80%
2.	Dataset 2	10076	614	6,09%
3.	Dataset 3	16584	1320	7,95%
4.	Dataset 4	32869	3882	11,81%
5.	Dataset 5	146286	380	0,25%
6.	Dataset 6	10059	860	8,54%
7.	Dataset 7	15046	1985	13,19%
8.	Dataset 8	57281	6886	12,02%
Rata-rata		41566,5	2368,125	8,33%

Dari Tabel 7, kedelapan dataset memiliki persentase kata tak baku terendah sebesar 0,25%, sedangkan persentase kata tak baku tertinggi sebesar 13,19%. Nilai persentase kata tak baku memiliki rata-rata sebesar 8,33%. Pada dataset 5 yang memiliki 146286 kata, hanya 380 kata yang termasuk kata tak baku. Walaupun jumlah kata dalam dataset sangat banyak, ternyata memiliki jumlah kata tak baku yang sedikit menurut kamus normalisasi satu. Kecilnya persentase kata tak baku pada dataset dapat disebabkan oleh beberapa faktor yaitu jumlah kata baku dalam dataset memang sangat banyak dibandingkan kata tak baku, jumlah kata tak baku dari dataset sebenarnya banyak tapi tidak terdapat pada kamus sehingga suatu kata akan dikenali kamus sebagai kata baku, ataupun penggunaan kata dalam bahasa lain sehingga kata tidak dikenali oleh kamus.

2. Frekuensi Kata Tak Baku Terbanyak Berdasarkan Kamus Normalisasi Satu

Dari Gambar. 3, terlihat bahwa sepuluh kata tak baku dengan frekuensi terbanyak adalah kata "yg"(yang) yang berjumlah 2086 kata pada keseluruhan dataset. Kata "yg" memiliki nilai frekuensi yang cukup berbeda secara signifikan dibandingkan dengan frekuensi kata "co", "aja", dan seterusnya. Dari kesepuluh kata tak baku, tidak terdapat kata yang memiliki makna sentimen(ungkapan emosi). Kesepuluh kata tersebut merupakan kata *stopword*(kata yang memiliki kemunculan tinggi). Dengan demikian, penggunaan kata tak baku sering digunakan pada kata yang termasuk *stopword* sehingga kemungkinan besar normalisasi memiliki pengaruh terhadap tahapan *stopword removal*.

Gambar. 3 Grafik 10 kata tak baku dengan frekuensi tertinggi menurut kamus normalisasi satu

D. Performa Kamus Normalisasi Dua

1. Persentase Kata Tak Baku Dalam Dataset berdasarkan kamus normalisasi dua

TABEL VIII
PERSENTASE KATA TAK BAKU DALAM DATASET MENURUT KAMUS NORMALISASI DUA

No.	Dataset	Jumlah Kata Total	Jumlah Kata Tak Baku Total	Persentase
1.	Dataset 1	44331	2946	6,64%
2.	Dataset 2	10076	649	6,44%
3.	Dataset 3	16584	959	5,78%
4.	Dataset 4	32869	3250	9,88%
5.	Dataset 5	146286	5973	4,08%
6.	Dataset 6	10059	845	8,40%
7.	Dataset 7	15046	1417	9,41%
8.	Dataset 8	57281	6375	11,12%
Rata-rata		41566,5	2801,75	7,71%

Dari Tabel 8, dari delapan dataset memiliki persentase kata tak baku terendah sebesar 4,08%. Sedangkan persentase kata tak baku tertinggi sebesar 11,12%. Nilai persentase kata tak baku memiliki rata-rata sebesar 7,71%. Jumlah kata tak baku dalam dataset berada pada rentang 649 sampai 6375 kata. Pada tabel, terlihat bahwa dataset 5 tetap memiliki persentase kata tak baku terkecil menurut kamus normalisasi dua. Bila dibandingkan dengan kamus normalisasi satu, normalisasi dengan kamus satu menghasilkan rata-rata persentase kata tak baku lebih tinggi sebesar 0,62% daripada normalisasi kamus dua.

2. Frekuensi Kata Tak Baku Terbanyak Berdasarkan Kamus Normalisasi Dua

Dari Gambar. 4, terlihat bahwa kata dengan frekuensi terbanyak dari kamus normalisasi satu dan dua adalah kata "yg"(yang). Dibandingkan dengan frekuensi kata tak baku pada kamus normalisasi satu, terlihat bahwa terdapat beberapa kata yang tak ada pada kamus normalisasi satu seperti kata "gue" dan "tau". Dari kesepuluh kata tersebut, terdapat dua kata yang memiliki makna sentimen(ungkapan emosi) yaitu "wkwk" dan "haha" yang berarti tertawa(senang). Delapan dari sepuluh kata pada Gambar. 4 merupakan kata *stopword*(kata yang memiliki kemunculan tinggi). Dengan demikian, penggunaan kata tak baku sering digunakan dalam kata *stopword* sehingga kemungkinan besar normalisasi memiliki pengaruh terhadap tahapan *stopword removal*.

Gambar. 4 Grafik 10 kata tak baku dengan frekuensi tertinggi menurut kamus normalisasi dua

E. Perbandingan Performa Kamus

1. Perbandingan Persentase Kata Tak Baku dari kedua kamus

Gambar. 5. Grafik Perbandingan Persentase Kata Tak Baku

Dari Gambar. 5, terlihat bahwa selisih antara persentase kata tak baku dari kedua kamus tidak terlalu signifikan pada beberapa dataset seperti pada dataset 1, 2, 6 dan 8. Sedangkan selisih kata tak baku paling signfikan berbeda ada pada dataset 5 dan dataset 7 dengan selisih masing-masing sekitar 3,7%. Selisih yang sangat besar ini disebabkan oleh perbedaan pengenalan kata tak baku oleh kedua kamus yang cukup banyak. Contohnya kata "minta"(meminta) tidak terdapat dalam kamus normalisasi satu tetapi terdapat pada kamus normalisasi dua dan kata "n"(dan) terdapat dalam kamus normalisasi satu tetapi tidak terdapat kamus normalisasi dua.

2. Perbandingan Performa

Dari Tabel 9, terlihat bahwa nilai median akurasi dari data yang tidak dinormalisasi mengalami kenaikan setelah dinormalisasi dengan kamus satu. Sedangkan nilai median akurasi data yang tidak dinormalisasi mengalami penurunan setelah dinormalisasi dengan kamus dua. Setelah dinormalisasi dengan kamus satu median nilai akurasi mengalami kenaikan sebesar 0,17% sedangkan dengan kamus normalisasi dua mengalami penurunan 0,15%.

TABEL IX
PERBANDINGAN AKURASI DARI TIGA TIPE DATA (DALAM %)

No.	Dataset	Data yang tidak dinormalisasi	Data yang dinormalisasi dengan kamus satu	Data yang Dinormalisasi dengan Kamus Dua
1.	Dataset 1	86,12	85,67	85,91
2.	Dataset 2	86,49	86,49	89,19
3.	Dataset 3	75,42	76,43	76,77
4.	Dataset 4	64,84	64,18	65,05
5.	Dataset 5	82,72	82,68	82,79
6.	Dataset 6	72,95	72,13	72,95
7.	Dataset 7	60,41	63,45	62,94
8.	Dataset 8	74,61	73,93	70,45
Median		75,01	75,18	74,86

TABEL X
PERBANDINGAN PRESISI DARI TIGA TIPE DATA (DALAM %)

No.	Dataset	Data yang tidak dinormalisasi	Data yang dinormalisasi dengan kamus satu	Data yang Dinormalisasi dengan Kamus Dua
1.	Dataset 1	86,12	85,63	85,91
2.	Dataset 2	83,64	84,91	88,46
3.	Dataset 3	74,91	75,70	75,97
4.	Dataset 4	76,00	66,67	75,00
5.	Dataset 5	88,84	88,90	89,08
6.	Dataset 6	72,50	72,27	72,50
7.	Dataset 7	59,06	61,79	61,48
8.	Dataset 8	79,95	79,02	74,22
Median		77,97	77,36	75,48

Dari Tabel 10, terlihat bahwa nilai median presisi dari data yang tidak dinormalisasi mengalami penurunan setelah dinormalisasi dengan kamus satu dan kamus dua. Setelah dinormalisasi dengan kamus satu median nilai presisi mengalami penurunan sebesar 0,61% sedangkan dengan kamus normalisasi dua mengalami penurunan 2,49%.

TABEL XI
PERBANDINGAN RECALL DARI TIGA TIPE DATA (DALAM %)

No.	Dataset	Data yang tidak dinormalisasi	Data yang dinormalisasi dengan kamus satu	Data yang dinormalisasi dengan kamus dua
1.	Dataset 1	100,00	100,00	100,00
2.	Dataset 2	88,46	86,54	88,46
3.	Dataset 3	99,54	99,54	99,54
4.	Dataset 4	10,98	11,56	12,14
5.	Dataset 5	72,57	72,41	72,27
6.	Dataset 6	100,00	98,85	100,00
7.	Dataset 7	74,26	75,25	74,26
8.	Dataset 8	64,50	63,91	61,22
Median		81,36	80,89	81,36

Dari Tabel 11, terlihat bahwa nilai median recall dari data yang tidak dinormalisasi setelah dinormalisasi dengan kamus satu mengalami penurunan. Sedangkan setelah dinormalisasi dengan kamus dua tetap sama. Setelah dinormalisasi dengan kamus satu median nilai presisi mengalami penurunan sebesar 0,47% sedangkan dengan kamus normalisasi dua tidak mengalami kenaikan ataupun penurunan.

F. Uji Hipotesis

1. Data yang tidak dinormalisasi dengan data yang dinormalisasi dengan kamus satu

TABEL XII
HASIL UJI HIPOTESIS ANTARA DATA YANG TIDAK DINORMALISASI DENGAN DATA YANG DINORMALISASI MENGGUNAKAN KAMUS SATU

No.	Indikator	Nilai p-value	Keputusan
1.	Akurasi	0,8657	Gagal Tolak H0
2.	Presisi	1,0000	Gagal Tolak H0
3.	Recall	0,3454	Gagal Tolak H0

Dari Tabel 12, kesimpulan pada pengujian terhadap ketiga indikator performa menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara ketiga indikator dari data yang tidak dinormalisasi dengan data yang dinormalisasi dengan kamus normalisasi satu.

Dari hasil pengujian p-value, terlihat bahwa memang tidak terdapat perbedaan nilai performa dari data yang tidak dinormalisasi dengan data yang dinormalisasi dengan kamus satu atau dengan kata lain perfoma data setelah dinormalisasi sama dengan data sebelum dinormalisasi. Hal ini memberikan hasil yang sama dengan penelitian sebelumnya yaitu [11] yang mana nilai F1-score dari penelitian tersebut memang tidak berbeda secara signifikan. Nilai presisi dan recall dari penelitian ini juga memberikan hasil yang sama dengan penelitian tersebut.

2. Data yang tidak dinormalisasi dengan data yang dinormalisasi dengan kamus dua

TABEL XIII
HASIL UJI HIPOTESIS ANTARA DATA YANG TIDAK DINORMALISASI DENGAN DATA YANG DINORMALISASI MENGGUNAKAN KAMUS DUA

No.	Indikator	Nilai p-value	Keputusan
1.	Akurasi	0,7531	Gagal Tolak H0
2.	Presisi	1,0000	Gagal Tolak H0
3.	Recall	0,5929	Gagal Tolak H0

Dari Tabel 13, kesimpulan pada pengujian terhadap ketiga indikator performa menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara ketiga indikator dari data yang tidak dinormalisasi dengan data yang dinormalisasi dengan kamus normalisasi dua.

3. Data yang dinormalisasi dengan kamus normalisasi satu dan kamus normalisasi dua

Dari Tabel 14, kesimpulan pada pengujian terhadap ketiga indikator performa menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara ketiga indikator dari data yang dinormalisasi dengan kamus normalisasi satu dengan kamus normalisasi dua.

TABEL XIV
HASIL UJI HIPOTESIS ANTARA DATA YANG DINORMALISASI MENGGUNAKAN KAMUS NORMALISASI SATU DENGAN DATA YANG DINORMALISASI MENGGUNAKAN KAMUS DUA

No.	Indikator	Nilai p-value	Keputusan
1.	Akurasi	0,9165	Gagal Tolak H0
2.	Presisi	0,7531	Gagal Tolak H0
3.	Recall	0,8927	Gagal Tolak H0

Performa yang tidak berbeda secara signifikan pada kedua kamus bisa disebabkan oleh beberapa faktor. Salah satu diantaranya adalah secara umum kata tak baku yang terdeteksi oleh kamus satu maupun kamus dua merupakan kata *stopword*, seperti yang ditunjukkan pada Gambar.3 dan Gambar.4. Adapun kata *stopword* ini akan dihapus pada tahapan *preprocessing* setelah normalisasi, sehingga pada akhrnya kata-kata yang telah dinormalisasi tidak diikutkan dalam analisis dan tidak mempengaruhi performa model yang dibangun.

VI. PENUTUP

Dari penelitian ini didapatkan hasil uji performa data yang tidak dinormalisasi dengan performa data yang dinormalisasi menggunakan kamus satu (*Colloquial Indonesian Lexicon*) dalam analisis sentimen yang menunjukkan bahwa bahwa tidak terdapat perbedaan performa yang signifikan baik akurasi, presisi, maupun recall. Demikian pula dengan hasil uji performa data yang tidak dinormalisasi dengan performa data yang dinormalisasi menggunakan kamus dua (kamus INANLP) dalam analisis sentimen menghasilkan kesimpulan yang sama. Normalisasi menggunakan kamus tidak menghasilkan kenaikan performa yang besar sehingga normalisasi dengan kamus tersebut dalam analisis sentimen bisa dilakukan ataupun tidak. Walaupun demikian pada pembahasan sebelumnya, terlihat bahwa kata yang dinormalisasi didominasi oleh kata *stopword* sehingga kemunculan kata ini kemungkinan tidak memberikan dampak yang besar terhadap performa. Disisi lain, dataset yang digunakan dalam penelitian memiliki keterbatasan karakteristik sehingga pengaruh normalisasi ini dapat diuji coba pada dataset lain yang memiliki kata *stopword* yang lebih sedikit dan bentuk kata tak baku lebih banyak.

Dari hasil uji performa data yang dinormalisasi dengan kamus satu(*Colloquial Indonesian Lexicon*) dengan performa data yang dinormalisasi dengan kamus dua(INANLP) menghasilkan kesimpulan bahwa tidak terdapat perbedaan yang signifikan dari performa kedua kamus dalam analisis sentimen baik akurasi, presisi, maupun recall. Dengan demikian, normalisasi pada analisis sentimen menggunakan kamus satu ataupun kamus dua tetap menghasilkan performa yang tidak memiliki selisih yang besar. Namun demikian normalisasi pada *tweet* berbahasa Indonesia tidak terbatas dilakukan menggunakan kedua kamus teresebut. Penelitian berikutnya dapat menguji normalisasi dengan menggunakan sumber lain berupa kamus ataupun algoritma lain.

Normalisasi pada penelitian ini hanya diimplementasikan pada analisis sentimen sehingga diperlukan penelitian lebih lanjut untuk menguji apakah hasil yang didapatkan pada penelitian ini berlaku pada bidang NLP yang lain seperti topic modelling, text summarization, named entity recognition, dan lain-lain.

DAFTAR PUSTAKA

[1] "Indonesia smartphone users," *Statista*. https://www.statista.com/statistics/266729/smartphone-users-in-indonesia/ (diakses Apr 30, 2021).

[2] "• Twitter: most users by country | Statista." https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (diakses Apr 30, 2021).

[3] I. F. Rozi, S. H. Pramono, dan E. A. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, hlm. 37–43, 2013.

[4] I. Sunni dan D. H. Widyantoro, "Analisis sentimen dan ekstraksi topik penentu sentimen pada opini terhadap tokoh publik," *J. Sarj. ITB Bid. Tek. Elektro Dan Inform.*, vol. 1, no. 2, 2012.

[5] N. Monarizqa, L. E. Nugroho, dan B. S. Hantono, "Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating," *J. Penelit. Tek. Elektro Dan Teknol. Inf.*, vol. 1, no. 3, 2014.

[6] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, hlm. 1–167, 2012.

[7] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, dan C. Richards, "Normalization of non-standard words," *Comput. Speech Lang.*, vol. 15, no. 3, hlm. 287–333, 2001.

[8] D. S. Maylawati dan G. P. Saptawati, "Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang," dalam *Journal of Physics: Conference Series*, 2017, vol. 801, no. 1, hlm. 012066.

[9] A. F. Hidayatullah, "Language tweet characteristics of Indonesian citizens," dalam *2015 International Conference on Science and Technology (TICST)*, 2015, hlm. 397–401.

[10] E. Clark dan K. Araki, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English," *Procedia-Soc. Behav. Sci.*, vol. 27, hlm. 2–11, 2011.

[11] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, dan A. Jamal, "Colloquial indonesian lexicon," dalam *2018 International Conference on Asian Language Processing (IALP)*, 2018, hlm. 226–229.

[12] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, dan F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," dalam *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 2016, hlm. 1–5.

[13] G. P. Wiratama dan A. Rusli, "Sentiment Analysis of Application User Feedback in Bahasa Indonesia Using Multinomial Naive Bayes," dalam *2019 5th International Conference on New Media Studies (CONMEDIA)*, 2019, hlm. 223–227.

[14] K. Hulliyah, N. S. A. A. Bakar, A. R. Ismail, dan M. O. Pratama, "A Benchmark of Modeling for Sentiment Analysis of The Indonesian Presidential Election in 2019," dalam *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, 2019, vol. 7, hlm. 1–4.

[15] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, dan A. Ahmed, "Multinomial Naive Bayes classification model for sentiment analysis," *IJCSNS*, vol. 19, no. 3, hlm. 62, 2019.

[16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, hlm. 1–30, 2006.

[17] C. C. Aggarwal dan C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[18] J. Han, M. Kamber, dan J. Pei, "Data mining concepts and techniques third edition," *Morgan Kaufmann Ser. Data Manag. Syst.*, vol. 5, no. 4, hlm. 83–124, 2011.

[19] S. N. Maula, "Kajian Perbandingan Algoritma Sentistrength dan Naïve bayes Analisis Sentimen Twitter (Studi Kasus: Pengesahan RUU Penghapusan Kekerasan Seksual)." Politeknik Statistika Sekolah Tinggi Ilmu Statistik, 2019.

[20] B. D. P. Nugraha, "Kajian Analisis Sentimen Data Twitter Menggunakan Metode Support Vector Machine Dengan Optimasi Pso dan Firefly." Jakarta: Politeknik Statistika Sekolah Tinggi Ilmu Statistik, 2017.

[21] M. Firdaus, "Kajian Perbandingan Metode Klasifikasi CNN Dan LSTM Pada Analisis Sentimen Berbahasa Indonesia (Studi Kasus: Data Twitter Tentang Harga Cabai)." Politeknik Statistika Sekolah Tinggi Ilmu Statistik, 2019.

[22] D. Rangga, "Twitter Sentiment Analysis Final Project," Sep 13, 2020. [Daring]. Tersedia pada: https://github.com/devildances/TwitterSentimentAnalysis_Final_Project

[23] Y. Indrawan, "yahdiindrawan/covid19-sentiment-dataset," Feb 09, 2021. https://github.com/yahdiindrawan/covid19-sentiment-dataset (diakses Apr 30, 2021).

[24] R. C. Rajagukguk, "riochr17/Analisis-Sentimen-ID," Mar 31, 2021. https://github.com/riochr17/Analisis-Sentimen-ID (diakses Apr 30, 2021).

[25] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, dan W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro Dan Teknol. Inf. JNTETI*, vol. 8, no. 4, hlm. 334–339, 2019.