3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

ANALISIS SENTIMEN ANIES BASWEDAN MENJADI CALON PRESIDEN2024 MENGGUNAKAN EKSTRASI FITUR COUNTVERCTORIZER DAN ALGORITMA KNN

Muhammad Ardhiansyah[1], Mohammad Syafrullah[2*]

[1]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia
[2]Manajemen Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia

Email: [1]1911500393@student.budiluhur.ac.id, [2*]mohammad.syafrullah@budiluhur.ac.id
(* : corresponding author)

Abstrak- Pemilihan umum (pemilu) diselenggarakan di Indonesia untuk memilih calon yang akan menjabat sebagai wakil rakyat. Khususnya pengguna Twitter menganggap isu presiden sangat menarik. Tweet publik tentang calon presiden dapat digunakan untuk menentukan sentimen publik, terlepas dari apakah mereka pendukung atau bukan. Anies Baswedan merupakan salah satu calon presiden indonesia 2024, dan tujuan penelitian ini adalah untuk mengetahui bagaimana pandangan orang-orang di Twitter tentangnya. Menyelidiki persepsi publik tentang Anies Baswedan sebagai calon presiden Indonesia di Twitter. Metode ini menggunakan algoritma K-Nearest Neighbor untuk klasifikasi, fitur kamus sentimen, pembelajaran mesin, ekstraksi fitur menggunakan CountVectorizer, dan fitur kamus sentimen. *Dataset* yang digunakan terdiri dari tweet berbahasa Indonesia yang ditemukan di Twitter menggunakan fungsi pencarian dan istilah "aniesbaswedan", "anies", "aniespresiden2024", "relawananies", "presiden2024", dan "#aniespresidenri2024". Analisis terhadap 664 tweet menghasilkan 51,42% untuk sentimen positif dan 48,58% untuk sentimen negatif. Menghasilkan nilai akurasi 71%, presisi 74%, dan recall 79% dengan nilai K = 5.

Kata Kunci: analisis sentimen, *k-nearest neighbor*, twitter, presiden2024

SENTIMENT ANALYSIS OF ANIES BASWEDAN AS A CANDIDATE FOR PRESIDENTIAL2024 USING COUNTVERCTORIZER FEATURE EXTRACTION AND KNN ALGORITHM

Abstract- *General elections are held in Indonesia to elect candidates who will serve as representatives of the people. Twitter users in particular find presidential issues very interesting. Public tweets about presidential candidates can be used to determine public sentiment, regardless of whether they are supporters or not. Anies Baswedan is one of the 2024 Indonesian presidential candidates, and the purpose of this study is to find out how people on Twitter view him. Investigating the public perception of Anies Baswedan as an Indonesian presidential candidate on Twitter. This method uses the K-Nearest Neighbor algorithm for classification, sentiment dictionary features, machine learning, feature extraction using CountVectorizer, and sentiment dictionary features. The dataset used consists of Indonesian tweets found on Twitter using the search function and the terms "aniesbaswedan", "anies", "aniespresiden2024", "relawananies", "president2024", and "#aniespresidenri2024". Analysis of 664 tweets resulted in 51.42% for positive sentiment and 48.58% for negative sentiment. Resulting in an accuracy value of 71%, precision of 74%, and recall of 79% with a value of K = 5.*

Keywords*: sentiment analysis, twitter, k-nearest neighbor,president2024*

1. PENDAHULUAN

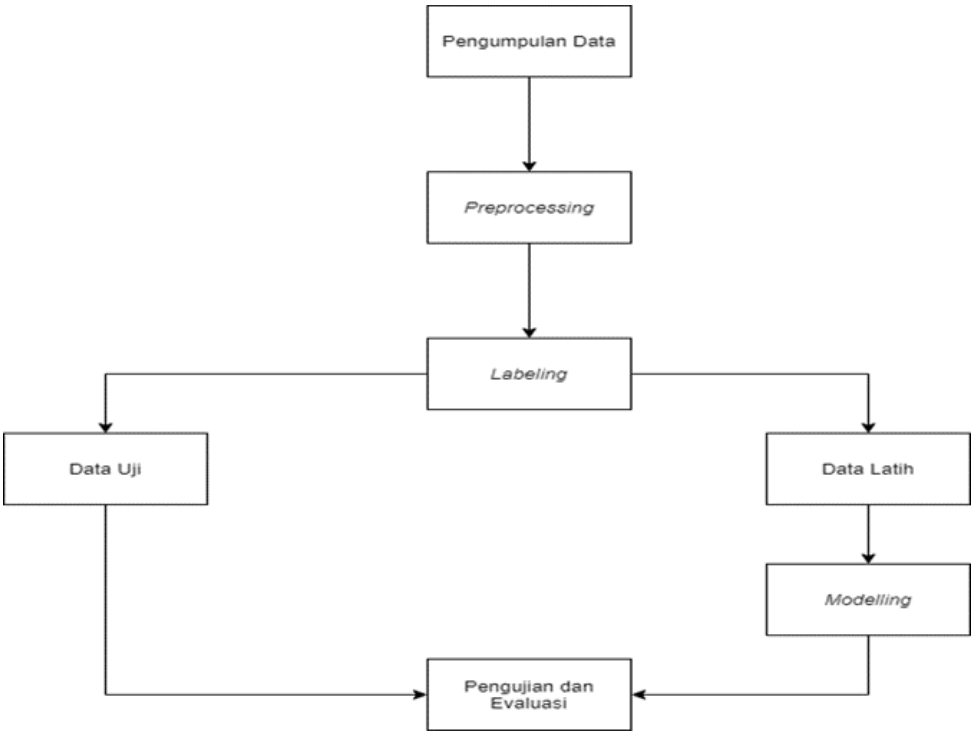
Indonesia menjunjung tinggi sistem demokrasi. Praktik demokrasi adalah penyelenggaraan pemilihan umum di Indonesia untuk memilih calon wakil rakyat. Sudah ada beberapa nama calon presiden yang telah diusung oleh partai pendukung dan pendukungnya masing-masing jelang pemilu 2024, antara lain Bapak H. Anies Rasyid Baswedan, S.E., M.P.P., Ph.D., Bapak H. Ganjar Pranowo, S.H, M.IP, dan Bapak Letnan Jenderal (Purn) H. Prabowo Subianto. Partai Nasdem mendukung Anies Baswedan mencalonkan diri sebagai presiden pada pemilihan umum 2024. Pengguna media sosial, khususnya Twitter, cukup vokal menyuarakan pemilihan presiden Anies Baswesdan. Platform media sosial yang digunakan oleh Twitter memungkinkan pengguna untuk

mengekspresikan pendapat mereka secara bebas. Berdasarkan permasalahan tersebut, diperlukan suatu sistem untuk analisis sentimen agar lebih mudah memahami sentimen dari setiap pendapat dan pendapat dalam setiap kalimat. Analisis sentimen merupakan teknik pengumpulan opini publik dengan memanfaatkan jejaring sosial yang berisi informasi tentang peristiwa terkini dan layanan publik [1]. Karena ada beragam pandangan pro dan kontra terhadap Pak Anies Baswedan yang mencalonkan diri sebagai presiden, akun Twitter resminya berpotensi digunakan sebagai sumber data untuk studi analisis sentimen. Berdasarkan penelitian analisis sentimen sebelumnya dengan judul "Analisis sentimen Twitter Capres Indonesia 2019 Menggunakan Metode *K-Nearest Neighbor*" [2]. Algoritma *K-Nearest Neighbor* (KNN) dapat mencapai skor akurasi klasifikasi sentimen sebesar 83,83%. Pada penelitian lain dengan judul "Analisis Sentimen Calon Presiden Indonesia 2019 dari media sosial Twitter menggunakan metode *Naïve Bayes*" [3]. Terdapat perbedaan metode dan perbedaan tahapan yaitu dengan menggunakan metode Naïve Bayes yang mendapatkan hasil terbaik pada pengujian dua kelas sentimen dengan hasil akurasi pada pasangan calon nomor urut satu dan pasangan calon nomor urut dua yaitu 77,7% dan 88% dengan performasi tertinggi pada calon presiden nomor urut dua dengan nilai f-measure sebesar 0,88. Kemudian pada tahap preprocessing juga terdapat perbedaan yaitu tokenization dan normalisasi

.

2. METODE PENELITIAN

Ada banyak langkah yang diambil dalam mengembangkan program analisis sentimen yang digunakan dalam penelitian ini. Dari awal hingga akhir program operasi, tahapan ini mewakili setiap prosedur dan desain studi. Gambar 1 di bawah ini mengilustrasikan langkah-langkah yang dilakukan:



Gambar 1 Alur Penelitian

Gambar 1 menggunakan proses *crawling* untuk mengumpulkan data dari *dataset tweet* sebagai sumber data. Untuk menyelesaikan tahap *preprocessing*, yang meliputi penyaringan, pengeksporan, dan pengisian koreksi, *tweet* yang diproses dalam format *excel* juga ditambahkan ke database. Prosedur penyusunan tersebut menghasilkan kalimat yang lebih terstruktur (*clean text*), yang kemudian digunakan pada tahap selanjutnya. Pada tahap pelabelan, *clean text* yang terkumpul akan diproses untuk memilih kelas (label) yang menunjukkan sentimen positif atau negatif. Tweet berlabel kemudian akan dibagi menjadi dua kategori: data uji dan data pelatihan. Data pelatihan adalah data yang berfungsi sebagai landasan pengetahuan dan membangun model pelatihan. Proses verifikasi tingkat akurasi dilakukan dengan pengujian dan penilaian, sedangkan pengecekan

akurasi model pelatihan yang dibuat proses pemodelan menggunakan data uji yang telah dihasilkan. Setelah proses pengujian dan penilaian, data akan digunakan untuk menunjukkan hasil dari penerapan pendekatan tersebut.

2.1 Pengumpulan data

Pada penelitian ini tahapan pengumpulan data dilakukan dengan menggunakan pendekatan *crawling.* Proses tersebut meliputi: mendapatkan API *key* Twitter melalui akun *developer* Twitter. API *key* Twitter yang diperoleh antara lain: Customer API *secret key* , Customer API *key, Access* Token, dan *Acces* Token *Secret.* Langkah selanjutnya melibatkan penggunaan pustaka Tweepy untuk menganalisis data yang diambil dari situs jejaring sosial Twitter. Data *tweet* akan dikumpulkan dan disimpan ke dalam file *excel* (*.xlsx*), setelah itu akan dimasukkan ke dalam database MySQL. Total 664 *tweet,* dikumpulkan dari 06 Juni hingga 11 Juni 2023, membentuk kumpulan data..

2.2 *Preprocessing*

Preprocessing merupakan tahapan persiapan data yang berupaya untuk mempermudah pengolahan data [4]. Pada tahap ini akan melibatkan penyaringan, penghapusan, dan koreksi data yang telah dikumpulkan dari Twitter dan media sosial lainnya. Ini berupaya untuk mendapatkan data twitter yang lebih terstruktur, atau "teks bersih". Teknik *preprocessing* berikut digunakan dalam penelitian ini:

a. *Case Folding*
Case folding adalah proses mengubah semua karakter kapital dalam dokumen menjadi huruf kecil [5]. Tujuan dari langkah ini adalah untuk membuat teks konsisten dan memudahkan analisis [3].
b. *Cleansing*
cleansing adalah proses metode untuk menghilangkan tanda baca, angka, simbol, tautan ke situs web eksternal, dan nama pengguna dari teks [6].
c. Mengubah *Slang Word*
Istilah *slang word* digunakan untuk mengubah kata-kata tidak baku dalam *tweet* menjadi kata-kata yang dimengerti dalam bahasa Indonesia [7].
d. Menghapus *Stop Word*
Stop word adalah kata-kata demonstratif, misalnya, adalah kata-kata dengan sedikit makna [8]. Karena tidak mempengaruhi keakuratan klasifikasi sentimen dan tidak relevan dengan subjek dokumen [3].
e. *Stemming*
Tahapan *stemming* adalah proses penghapusan prefiks dan sufiks untuk mereduksi konjungsi, pronomina, dan kata-kata yang mengandung preposisi menjadi bentuk dasarnya [9]

2.3 *Labeling*

Pelabelan adalah proses pengkategorian berdasarkan ciri-ciri atau ciri-ciri yang ada pada suatu dokumen atau kalimat. Setiap *tweet* akan diberi label positif atau negatif sebagai hasil dari proses pelabelan. Tweet dengan label positif cenderung setuju atau mendukung Anies Baswedan sebagai calon presiden 2024, sedangkan *tweet* dengan label negatif cenderung tidak setuju atau menolak Anies Baswedan sebagai calon presiden 2024. Ada dua (2) teknik untuk mengkategorikan tweet, antara lain pelabelan manual menggunakan frase yang telah diberi label subjektivitas dan pelabelan manual menggunakan pendekatan kamus sentimen.

2.4 Pembagian data

Pada tahap pembagian data, *tweet* yang diberi tag akan dibagi menjadi data pengujian dan data pelatihan selama tahap berbagi data. *Dataset* dibagi menjadi 90% data latih dan 10% data uji untuk melakukan pemisahan data. Hasil dari pembagian ini, sebagian besar data akan digunakan sebagai data latih untuk model latih dan sebagian kecil sebagai data uji untuk mengevaluasi akurasi model yang dikembangkan.

2.5 *Modeling*

Modeling adalah proses pembangunan pengetahuan berdasarkan data yang sudah tersedia. Dalam proses ini, data latih yang akan digunakan untuk membangun model dipilih menggunakan teknik *qouta sampling*. Teknik *qouta sampling* merupakan metode *sampling* yang menentukan jumlah sampel dari populasi yang memenuhi kriteria tertentu, sehingga jumlah sampel yang diinginkan atau ditetapkan dapat tercapai. Dengan menggunakan teknik ini, data latih yang akan digunakan dalam proses *modeling* dipilih berdasarkan kuota tertentu, sehingga model yang dihasilkan dapat mewakili karakteristik dari data tersebut [10]. Tahap pemodelan

dilakukan untuk mengekstrak *tweet* dari data latih menjadi representasi vektor menggunakan *CountVectorizer*. Setelah itu, data diubah menjadi representasi vektor dan disimpan sebagai model pelatihan dalam format file *JSON*(*.json*).

2.6 Pengujian dan Evaluasi

Pengujian dilakukan untuk mengevaluasi akurasi, presisi, dan *recall*, model pelatihan yang menggunakan algoritma yang disarankan. Membandingkan data antisipasi tertentu (data dari tahap klasifikasi) dengan data aktual (data dari tahap pelabelan) menjadi metode pengujian dalam penelitian ini. Untuk menghitung jarak antara data digunakan perhitungan euclidean distance dengan rumus sebagai berikut:

$$\sqrt{\sum_{i=0}^{n}(X_{1i} - X_{2i})^2} \quad (1)$$

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset atau data yang digunakan dalam penelitian ini terdiri dari 664 teks *tweet* yang dikumpulkan dari Twitter antara 6 Juni dan 11 Juni 2023. Data dikumpulkan melalui proses *crawling* menggunakan library Tweepy. Kumpulan data tersebut didasarkan pada kata kunci yang terkait dengan pencalonan Anies Baswedan sebagai presiden Indonesia pada tahun 2024, termasuk "aniesbaswedan", "anies", "aniespresiden2024", "relawananies", dan "#aniespresidenri2024".

3.2 *Preprocessing*

Dalam penelitian ini, data awal dari situs jejaring sosial Twitter mengalami beberapa tahapan *preprocessing.* Meliputi *case folding*, *cleansing*, mengubah *slang word*, menghapus *stop word*, dan *stemming.*

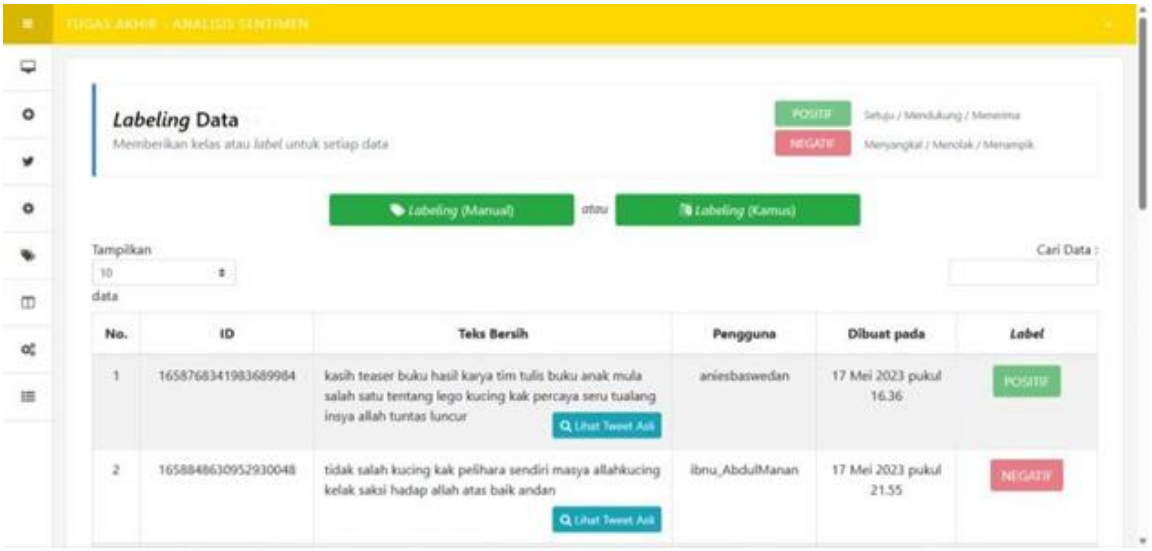
Tabel 1. Proses *preprocessing data*

Tahap	Input	Output
Case Folding	@CutSarina5 @Boediantar4 @AniesBaswedan Capres yg smart, Berwibawa kaya akan ide dan gagasan… Prestasinya luar biasa #AniesPresidenRI2024 #AniesPresidenRI2024 https://t.co/9q7M1wpPeb	@CutSarina5 @Boediantar4 @AniesBaswedan capres yg smart, berwibawa kaya akan ide dan gagasan… prestasinya luar biasa #AniesPresidenRI2024 #AniesPresidenRI2024 https://t.co/9q7M1wpPeb
Cleansing	@CutSarina5 @Boediantar4 @AniesBaswedan capres yg smart, berwibawa kaya akan ide dan gagasan… prestasinya luar biasa #AniesPresidenRI2024 #AniesPresidenRI2024 https://t.co/9q7M1wpPeb	capres yg smart, berwibawa kaya akan ide dan gagasan prestasinya luar biasa.
Mengubah *Slang Word*	capres yg smart, berwibawa kaya akan ide dan gagasan prestasinya luar biasa.	capres yang smart, berwibawa kaya akan ide dan gagasan prestasinya luar biasa
	capres yang smart, berwibawa	capres smart, berwibawa kaya ide

Menghapus *Stop Word*	kaya akan ide dan gagasan prestasinya luar biasa.	gagasan prestasi luar.
Stemming	capres smart, berwibawa kaya ide gagasan prestasi luar	capres smart, wibawa kaya ide gagas prestasi luar

3.3 *Labeling* Data

Tahap pelabelan tidak dapat diselesaikan kecuali satu atau lebih data teks bersih dari tahap *preprocessing* dibuat dapat diakses di database. Label untuk data positif dan negatif dibagi menjadi dua kelompok. Selama pelabelan, data yang ada diberi label secara manual.



Gambar 2. *Labeling* manual

3.4 Pengujian dan Evaluasi

Untuk menilai, menganalisis, dan menentukan tingkat akurasi yang telah dicapai oleh sistem yang telah dirancang, pengujian merupakan salah satu hal yang harus dilakukan. Pada penelitian ini penggunaan algoritma *K-Nearest Neighbor* (KNN) dalam memprediksi label untuk data uji diuji dari segi akurasi, presisi, dan *recall*. Ujian dalam penelitian ini menguji nilai K berdasarkan perubahan yang direncanakan, termasuk K=3, K=5, K=7, K=9, K=11, selain akurasi, presisi, dan *recall.* Tabel 2 Berikut ini menunjukkan hasil prediksi sampel dengan algoritma KNN dengan nilai K=3.

Tabel 2. Sampel data hasil prediksi

No	*Tweet*	*Label* aktual	*Label* prediksi
1	@VeelaaRhie perempuan2 kok ... ya gatega nerusin ucapan. emang udah trbukti korupsi???? T*l*l #AniesPresidenRI2024	negatif	positif
2	@MartinusButarb1 sepakat awakbang. #AniesPresidenRI2024	positif	positif

3	@msaid_didu Itulah orang2 yg ga punyamalu ya pak Said? #AniesPresidenRI2024	negatif	positif
4	@JarnasABWBpn Dengan keizinan Nyatidak ada yg tidak mungkin,,,INSYAA ALLAH #AniesPresidenRI2024 #IndonesiaButuhAnies	negatif	negatif
...	...	...	...
49	Luar biasa, Salut pada Ahlak Pak Anies Baswedan Ternyata Dia Mencontoh Ahlaknya Nabi Muhammad SAW. Selalu Membalas Kejahatan Orang Laindengan Senyum dan Kebaikan. #AniesBaswedan #AniesBaswedanLebihNKRI #AniesDijegalRakyatRevolusi #AniesPresidenRI2024 https://t.co/jHaEgjjk3W https://t.co/aE2zd7w1Bs	positif	positif

Pada Tabel 3 sampel data prediksi, kolom label prediksi merepresentasikan data label yang diperoleh melalui proses klasifikasi menggunakan KNN, sedangkan kolom label aktual merepresentasikan data label yang diperoleh melalui proses pelabelan. Data uji (49 data *tweet*) hasil prediksi keseluruhan kemudian ditampilkan dalam *Confusion matrix*. Tabel 4.3 dibawah ini menunjukkan penggambaran *Confusion matrix* K = 3 yang dihasilkan.

Tabel 3. *Confusion matrix* k=3

		Nilai Aktual	
		positif	negatif
Nilai Prediksi	positif	27	13
	negatif	2	7

Berikut dapat dilihat pada Tabel 4 untuk perolehan akurasi, presisi, dan recall berdasarkan Confusion matrix dengan K = 3:

Tabel 4. Nilai Pengujian K=3

Pengujian		
Akurasi	$= \frac{27+7}{27+7+13+2}$	0.69 (69 %)
Presisi	$= \frac{27}{27+13}$	0.68 (68 %)

Recall	$= \frac{27}{27+2}$	0.93 (93%)

Pengujian diatas diulangi dengan berbagai modifikasi nilai K. sehingga jelas dari Tabel 4.17 dibawah ini bahwa hasil pengujian secara keseluruhan adalah seperti yang ditunjukkan:

Tabel 5. Hasil pengujian dan evaluasi

	K=3	K=5	K=7	K=9	K=11
Akurasi	0.69	0.71	0.65	0.61	0.63
Presisi	0.68	0.74	0.66	0.61	0.62
Recall	0.93	0.79	0.86	0.97	0.97

Berdasarkan Tabel 5. Dengan menggunakan K = 5, algoritma KNN mampu mencapai nilai pengujian tertinggi, dengan akurasi 71%, presisi 74%. dan *recall* 79%. Sementara 664 *tweet* dianalisis sentimennya, hasilnya menunjukkan bahwa sentimen masyarakat Indonesia cenderung positif 51,42% dan negatif 48,58% pada priode juni 2023.

4. KESIMPULAN

Berdasarkan 664 *tweet* data yang diperoleh antara tanggal 6 sampai 11 juni 2024, pandangan (sentimen) masyarakat Indonesia terhadap Anies Baswedan menjadi calon presiden 2024 cenderung positif 51,42% dan negatif 48,58% pada periode bulan juni 2023. Algoritma *K-Nearest Neighbor* (KNN) dengan ekstraksi fitur *CountVectorizer* dapat digunakan untuk melakukan analisis sentimen secara efektif. Skor tes dan penilaian teratas adalah akurasi 71%, presisi 74%, dan *recall* 79% dengan nilai K = 5.

DAFTAR PUSTAKA

[1] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.

[2] A. Z. Malik, E. Utami, and S. Raharjo, "Analisis Sentiment Twitter Terhadap Capres Indonesia 2019 dengan Metode K-NN," *J. Inf. Politek. Indones. Surakarta*, vol. 5, no. 2, pp. 1–7, 2019.

[3] S. Nurul, J. Fitriyyah, N. Safriadi, E. Esyudha, and P. #3, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *(Jurnal Edukasi dan Penelit. Inform.*, vol. 5, no. 3, pp. 279–285, 2019, [Online]. Available: http://dev.twitter.com.

[4] D. T. Lukmana, S. Subanti, and Y. Susanti, "Analisis Sentimen Terhadap Calon Presiden 2019 Dengan Support Vector Machine Di Twitter," *Semin. Nas. Penelit. Pendidik. Mat. 2019 UMT*, no. 2002, pp. 154–160, 2019.

[5] L. A. Andika, P. A. N. Azizah, and R. Respatiwulan, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," *Indones. J. Appl. Stat.*, vol. 2, no. 1, p. 34, 2019, doi: 10.13057/ijas.v2i1.29998.

[6] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[7] N. Ruhyana, "Analisis Sentimen terhadap Penerapan Sistem Plat Nomor Gnajil/Genap pada Twitter dengan Metode Klasifikasi Naive Bayes," *J. IKRA-ITH Inform.*, vol. 3, no. 1, pp. 94–99, 2019, [Online]. Available: www.situs.com

[8] S. Informatika and J. Komputer, "Perbandingan algoritma dalam analisa sentimen krisis evergrande pada kanal berita youtube," vol. 11, no. 2, pp. 72–76, 2021.

[9] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm," *Jdmsi*, vol. 2, no. 1, pp. 31–37, 2021, [Online]. Available: https://t.co/NfhnfMjtXw

[10] M. Priandi and Painem, "Analisis Sentimen Masyarakat Terhadap Pembelajaran Daring di Era Pandemi Covid-19 pada Media Sosial Twitter Menggunakan Ekstraksi Fitur Countvectorizer dan Algoritma K-Nearest Neighbor," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. September, pp. 311–319, 2021.