Sentiment Analysis Before Presidential Election 2024 Using Naïve Bayes Classifier Based On Public Opinion In Twitter

Analisa Sentimen Jelang Pilpres 2024 Menggunakan Naïve Bayes Classifier Berdasarkan Opini Publik Di Twitter

Heri Prasetyo[1], Arif Senja Fitrani[2]
{191080200066@umsida.ac.id[1], asfjim@umsida.ac.id[2]}

[1,2] Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sidoarjo

Abstract. *This study aims to determine the performance of the Naïve Bayes Classifier algorithm and sentiment analysis tested on a dataset obtained from Twitter social media scrapping with the topic of 2024 presidential candidates. Three candidates frequently discussed in public spaces were used as keyword parameters in data mining: #anis, #ganjar, and #pilpres2024, resulting in 3021 tweets extracted from 12/1/2022 to 31/1/2023, which were successfully converted to ".csv" format documents. Public opinions extracted from the dataset were then pre-processed using the Python programming language, resulting in 2157 cleaned tweets. The data that passed the pre-processing stage was then labeled as positive or negative sentiment. Sentiment analysis was performed using the Naïve Bayes Classifier algorithm with three testing experiments using different training and testing data compositions in each experiment. The results of the study showed that the best Naïve Bayes model was obtained in the first experiment with a 10% testing data and 90% training data composition, resulting in 71% accuracy, 93% precision, 66% recall, and an f-measure score of 77%. The conclusion of the study is that the electability of the 2024 presidential candidates shapes public opinion and generates public sentiment in the form of positive and negative tweets. Positive tweets had a higher percentage of 71.5% (1543), while negative sentiment tweets accounted for 28.5% (614). Further research is expected to produce different information by using different classification algorithms and larger data sets*

Keywords *- Naïve Bayes Classifier, Public Opinion, Sentiment Analysis, Twitter*

Abstrak. *Penelitian ini bertujuan mengetahui performa algoritma naïve bayes classifier dan Analisa sentimen yang diuji pada dataset hasil scrapping pada sosial media twitter dengan topik kandidat pilpres 2024. 3 kandidat yang sering dibicarakan diruang publik menjadi parameter keyword pada penambangan data #anis, #ganjar dan #pilpres2024 sehingga didapatkan 3021 tweet dengan limit waktu 1/12/2022 sd. 31/1/2023 yang berhasil di ekstrak menjadi format dokumen '.csv'. Opini publik yang berhasil di ekstrak kemudian dilakukan proses pre-processing menggunakan Bahasa pemrogaman python yang kemudian menghasilkan data 2157 tweet yang telah dibersihkan, data yang berhasil melewati proses pre-processing kemudian di berikan label positif dan negatif sehingga mempunyai sifat sentimental. Analisa sentimen dilakukan menggunakan algoritma naïve bayes classifier dengan 3 kali percobaan pengujian data dengan komposisi pembagian data uji dan data latih berbeda pada tiap kali percobaan. Hasil penelitan menunjukkan bahwa model naïve bayes terbaik terdapat pada percobaan pertama dengan pembagian data 10% data uji dan 90% data latih, pada percobaan pertama didapatkan 71% keakurasian nya, dengan nilai precission 93%, recall 66% dan f - measure scored 77%. Kesimpulan penelitian elektabilitas pilpres 2024 membentuk opini publik dan menimbulkan sentiment publik berupa tweet yang bersifat positif, negatif, perbincangan/tweet pada laman sosial media twitter positif memiliki persentase lebih tinggi dengan 71,5% (1543) dan percakapan yang memiliki sentiment negatif 28,5% (614). Pada penelitian selanjutnya diharapkan menghasilkan sesuatu informasi yang berbeda dengan menggunakan algoritma klasifikasi yang lain dan jumlah data yang lebih banyak.*

Kata Kunci *- Analisa Sentimen, Naïve Bayes Classifier, Opini Publik, Twitter*

I. PENDAHULUAN

Kontestasi pemilihan presiden (pilpres) selalu menjadi topik yang sangat menarik diulas [1], pergerakan berita dari berbagai macam unsur selalu muncuk ke permukaan, kandidat yang sedang diusung sudah mulai diapungkan oleh partai – partai politik dan media [2]. Pencarian mengenai topik pilpres 2024 mulai ramai di halaman pencarian, google maupun social media, salah satunya media social twitter.

Twitter merupakan media berbagi informasi berupa foto, video dan narasi apasaja yang tidak dilarang oleh kebijakan perusahaan(twitter) [3]. Pejabat tinggi pemerintah bahkan lembaga di Indonesia banyak menggunakan social media ini [4] begitu pula dengan tokoh kandidat calon presiden pada pilpres 2024, banyak narasi yang dilontaran untuk menarik aspirasi public yang bertujuan mendeklarasikan opininya jelang pilpres 2024, banyak masyarakat pengguna twitter merespon kejadian tersebut di kolom komentar(re-tweet) sampai mengungkapkan narasinya pada laman pribadinya(tweet) untuk menanggapai narasi yang lontarkan kandidat pilpres 2024, dari yang pro kepada calon tersebut bahkan tanggapan negatif tentangnya [5].

Narasi yang dibangun selalu mempunyai sifat positif dan negative, itulah yang disebut dengan sentiment.

Analisis sentimen adalah salah satu bentuk analisis data yang bertujuan untuk mengetahui dan mengevaluasi opini, sentimen, dan emosi yang ada dalam suatu narasi/topik tertentu, seperti kasus pejabat publik, polemik atau suatu produk terntentu [6]. Analisis sentimen sering digunakan untuk memahami pandangan dan persepsi masyarakat terhadap suatu topik tertentu, termasuk juga dalam konteks politik, seperti pemilihan presiden.

Pemilihan presiden sendiri merupakan ajang pesta demokrasi untuk memilih suatu tokoh terbaik pada suatu negara sehingga dapat dipilih oleh masyarakat dan dijadikan suatu pimpinan tertinggi di negara tersebut [7]. Proses politik ini yang seringkali menimbulkan banyak polemik, sehingga menarik dikupas dan dijadikan suatu pembelajaran dalam taraf intelktual akademis maupun teoritis bagi para pengamat, pelaku, pakar politik [8].

Data yang didapatkan pada penelitian ini bersumber dan memanfaatkan opini public yang dituangkan pada laman twitter, proses penambangan datanya sendiri disebut dengan teks mining dengan metode scapping menggunakan Bahasa pemrogaman python. Teknik scrapping merupakan salah satu cara untuk mengunduh data di ruang media public yang cukup efektif [9], pada kasus ini digunakan untuk mengunduh opini public yang bersinggungan dengan kandidat capres pada pilpres 2024.

Hasil scrapping merupakan data mentah yang belum dapat kita gunakan untuk menganalisa suatu kasus, harus melalui beberapa tahapan seperti teks preprocessing, eksplorasi data, pemodelan topik menggunakan algoritma Naïve Bayes Classifier (NBC), evaluasi dan sehingga mendapatkan tujuan penelitian [10].

Algoritma Naïve Bayes Classifier sendiri merupakan metode klasifikasi berdasarkan teorema “bayes” menggunakan pembelajaran mesin, dengan perhitungan probabilitas frekuensi nilai yang sering muncul [11].

Oleh karena itu dalam penelitian ini ingin melakukan pengujian kinerja algoritma NBC pada study topik pilpres 2024 dan diketahui elektabilitas potensi kandidat bakal calon presiden pada pemilihan presiden 2024 mendatang berdasarkan opini publik.

II. METODE

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan proses sesuai kaidah pengolahan data yang sering digunakan pada penelitian – penelitan sebelumnya. Yakni meliputi penambangan data(*scrapping)*, preprocessing, labeling, pemodelan topik, analisis dan evaluasi. Seperti pada gambar 1 dibawah ini.

Gambar 1. Alur kerja penelitian

2.2 Penambangan Data

Scrapping (atau web scraping) adalah proses ekstraksi data dari sebuah website atau sumber informasi lainnya secara otomatis dengan menggunakan software atau bot tertentu. Proses ini dilakukan dengan cara mengekstrak data secara sistematis dari website, kemudian menyimpan data tersebut dalam format yang bisa diakses dan diolah dengan mudah. Tujuan dari scrapping bisa bermacam-macam, mulai dari pengambilan informasi produk pada situs e-commerce, penelitian pasar, hingga pengumpulan data untuk analisis dan penelitian ilmiah. Scrapping biasanya dilakukan dengan memanfaatkan tools atau bahasa pemrograman seperti Python, JavaScript, atau R untuk membaca dan mengambil data dari website [12].

2.3 Pre-Processing

Preprocessing merupakan proses menstandarkan data yang didapatkan dari public melalui tahapan – tahapan agar data terstandar, sehingga mengurangi *error-rate* pada proses Analisa data [13]. Tahapan preprocessing seperti berikut

a. Case Folding : adalah proses menyelaraskan karakter tweet yang tidak diperlukan. Karakter yang di proses hanyalah abjad yakni ‘a’ sampai ‘z’. Karakter selain itu akan dihilangkan.
b. Cleansing : merupakan proses untuk membersihkan kata yang tidak diperlukan dengan beberapa teknik seperti memperkecil noise, membetulkan data yang tidak konsisten, mengisi data kosong, mengidentifikasi atau membuang outlier. Kata yang akan dihilangkan meliputi URL, Hashtag (#), Username (@). Selain itu juga akan menghilangkan tanda baca seperti koma (,), titik (.), Tanda seru (!), dan tanda baca lainnya.
c. Tokenizing : merupakan proses pemisahan kalimat menjadi suatu token atau kata yang terpotong. Manfaat memisahkan menjadi perkata ini adalah untuk memudahkan pada proses selanjutnya, yakni proses stopword dan stemming, karena dua proses tersebut akan mencocokan kata per kata dengan library root nya.
d. Stopward : merupakan proses membersihkan kata yang tidak diperlukan, apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak, sehingga yang tersisa adalah kata penting saja atau disebut keywords.

e. Stemming : berfungsi untuk mereduksi kata yang bukan stopword menjadi ke-root word yang sesuai, dengan dilakukan nya proses *stemm* sehingga menghilangkan awalan dan akiran atau kata imbuhan.

2.4 Labeling

Merupakan tahapan untuk memberikan label pada setiap opini / tweet yang datanya berhasil di standarisasikan pada tahapan pre-processing, proses labeling sendiri bertujuan memberikan paradigma yang bersifat positif dan negative pada masing – masing kalimat, sehinggi mempunyai sentiment [14].

2.5 Pemodelan Topik

Pada penelitian ini yang digunakan adalah Algoritma Naive Bayes Classifier merupakan teknik klasifikasi berdasarkan Teorema Bayes dengan asumsi independensi di antara para prediktor. Naive Bayes Classifier memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Dalam istilah sederhana, penggolongan Naive Bayes menganggap bahwa kehadiran fitur tertentu di kelas tidak terkait dengan kehadiran fitur lainnya. Keuntungan penggunaan adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variabel independen, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

$$P\ (H|X) = \frac{P(X|H)P(H)}{P\ (X)} \quad (1)$$

Dengan P merupakan probabilitas, X adalah data dengan class yang belum diketahuai, H adalah hipotesis data X yang merupakan suatu class spesifik, P(H|X) merupakan probabilitas hipotesis H berdasarkan kondisi X (*posteriori prob.)*, P(H) adalah probabilitas hipotesis H (*prior prob*), P(X|H) adalah probabilitas X berdasarkan kondisi tersebut, P(X) merupakan probabilitas dari X [15].

2.6 Analisa dan evaluasi

Setelah melakukan preprocesing dan labeling, kemudian data dibagi menjadi data training dan data testing untuk melakukan pengujian model data, untuk mengetahui keakurasian pemodelan data atau metode yang digunakan maka dapat mengetahui nilai precission, recall & accuration serta f-measure. Proses Analisa sendiri bisa meliputi eksplorasi data seperti mengetahui karakteristik dan keberhasilan preprocesing yang dapat dilihat dari *wordcloud* seperti gambar 2 dibawah ini.

Gambar 2. Wordcloud

III. HASIL DAN PEMBAHASAN

Berdasarkan proses tahapan pengolahan data yang dilakukan pada penelitian ini, proses penambangan data yang berhasil dilakukan berjumlah 3021 dengan parameter *query* "pilpres 2024", "anis", "ganjar" untuk mengetahui secara umum sentiment opini public apa saja yang sedang di bicarakan di ruang publik, "anis" seorang politikus dan intelektual Indonesia yang saat ini menjabat sebagai Gubernur DKI Jakarta sejak 2017, "ganjar" yang merupakan seorang politikus dan gubernur Jawa Tengah yang saat ini menjabat untuk periode kedua sejak 2018, masing – masing dari kata kunci tersebut adalah kandidat calon presiden pada pemilihan presiden 2024 yang sedang ramai dibicarakan di ruang public, media social dan media cetak, data di ekstrak menjadi format .csv yang kemudian dilakukan proses selanjutnya.

Kemudian data dilakukan proses pre-processing untuk menstandarisasi, Kemudian data berkurang menjadi 2157 data setelah di bersihkan pada tahapan preprocessing, data berkurang akibat beberapa data ada yang redundant. Proses

pre-processing dapat dilihat pada table 1 dibawah ini.

Tabel 1. Proses Pre-processing

Proses	Sebelum	Sesudah
Case Folding	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	pilpres 2024 semoga mempunyai hasil yang baik, bagus dan berkualitas. indonesia maju !!!
Cleansing	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	pilpres 2024 semoga mempunyai hasil yang baik bagus dan berkualitas indonesia maju
Tokenizing	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘mempunyai’ ‘hasil’ ‘yang’ ‘baik’ ‘bagus’ ‘dan’ ‘berkualitas’ ‘indonesia’ ‘maju’
Stopword	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘mempunyai’ ‘hasil’ ‘baik’ ‘bagus’ ‘berkualitas’ ‘indonesia’ ‘maju’
Stemming	PILPRES 2024 Semoga mempunyai hasil yang baik, bagus dan berkualitas. Indonesia Maju !!!	‘pilpres’ ‘2024’ ‘semoga’ ‘punya’ ‘hasil’ ‘baik’ ‘bagus’ ‘dan’ ‘kualitas’ ‘indonesia’ ‘maju’

Kemudian data tersebut dilakukan proses labeling sehingga di dapatkan sentiment yang berisifat positif (1543) dan negative (614).

Setelah melalui proses pre-processing dan labeling selanjutnya dilakukan proses pemodelan topik yang pada penelitian ini menggunakan algoritma NBC, sebelum itu data dilakukan pembagian menjadi 2 bagian yakni data latih dan data uji yang selanjutnya dilakukan tes menggunakan algoritma NBC. Proses pengujian dilakukan sebanyak 3 kali dengan masing – masing komposisi pembagian data.

a. Percobaan Ke – 1

Pada percobaan pertama menggunakan komposisi pembagian data : 90% data latih dan 10% data uji, menghasilkan 71% keakurasian menggunakan metode NBC. Pada gambar 3 dijelaskan secara rinci hasil percobaan pertama.

	precision	recall	f1-score	support
negatif	0.45	0.85	0.59	53
positif	0.93	0.66	0.77	163
accuracy			0.71	216
macro avg	0.69	0.76	0.68	216
weighted avg	0.81	0.71	0.73	216

Gambar 3. Hasil percobaan ke – 1

b. Percobaan Ke – 2

Pada percobaan selanjutnya komposisi data yang digunakan adalah 80% data latih dan 20% data uji. Keakurasian pada percobaan ini didapatkan nilai 68%. Seperti pada gambar 4.

	precision	recall	f1-score	support
negatif	0.46	0.82	0.59	179
positif	0.90	0.63	0.74	469
accuracy			0.68	648
macro avg	0.68	0.72	0.67	648
weighted avg	0.78	0.68	0.70	648

Gambar 4. Hasil percobaan ke – 2

c. Percobaan Ke – 3

Pada percobaan terakir menggunakan pembagian data 70% data latih dan 30% data uji sehingga mendapatkan 69% akurasi data, seperti dijelaskan pada gambar 5.

negatif	0.48	0.82	0.60	246
positif	0.90	0.64	0.75	617
accuracy			0.69	863
macro avg	0.69	0.73	0.68	863
weighted avg	0.78	0.69	0.71	863

Gambar 5. Hasil percobaan ke – 3

Pada percobaan diatas dengan 3 kali uji data dengan masing – masing pembagian data uji dan data latih sehingga didapatkan nilai akurasi, *precission, recall* dan *f-measure* yang berbeda pada setia percobaan, seperti dijelaskan pada tabel 2 dibawah.

Tabel 2. Rekap percobaan

Percobaan	Kelas	Akurasi	*Precission*	*Recall*	*f-measure*
Ke – 1	Negatif	0.71	0.45	0.85	0.59
	Positif		0.93	0.66	0.77
Ke – 2	Negatif	0.68	0.46	0.82	0.59
	Positif		0.90	0.63	0.74
Ke – 3	Negatif	0.69	0.48	0.82	0.60
	Positif		0.90	0.64	0.75

IV. KESIMPULAN

Berdasarkan percobaan yang telah dilakukan sehingga mendapatkan hasil yang ingin dicari, hasil percobaan pengujian metode Naïve Bayes Classifier terhadap data hasil *scrapping* opini publik pada sosial media twitter dengan topik pilpres 2024 dan parameter pencarian kata "pilpres 2024", "anis", "ganjar" dan kata pendekatan atau yang berhubungan lainnya. Data hasil scrapping 3021 dilakukan pre-*processing* sehingga menjadi 2157 kata, kemudian diberikan label sentimen yang bermuatan positif berjumlah 1543 (71,5%) dan negatif berjumlah 614 (28,5%). Percobaan yang telah dilakukan berturut – turut sebanyak 3 kali percobaan dengan hasil terbaik diperoleh pada percobaan ke – 1 sehingga di dapatkan *accuracy scored* 71% dengan nilai *precission* 93% class positif 45% negatif, *recall* 66% class negatif 85% positif dan f - *measure scored* 59% pada class negative 77% pada class positif. Sehingga dapat disimpulkan bahwa percobaan terbaik yang dilakukan adalah pada percobaan ke – 1 dengan pembagian 10% data uji dan 90% data latih. Pada penelitian selanjutnya diharapkan menggunakan algoritma lain dengan data yang lebih banyak sehingga dapat mendapatkan hasil yang berbeda.

UCAPAN TERIMA KASIH

Terimakasih kepada Universitas Muhammadiyah Sidoarjo yang telah memberikan fasilitas laboratorium computer informatika sehingga penelitian ini dapat diselesaikan dengan baik, kemudian dapat memberikan pembelajaran dan pengetahuan secara umum pada publikasi artikel ini.

REFERENSI

[1] M. A. Firmansyah, D. Mulyana, S. Karlinah, and S. Sumartias, "Kontestasi Pesan Politik dalam Kampanye Pilpres 2014 di Twitter: Dari Kultwit Hingga Twitwar," *JIK*, vol. 16, no. 1, p. 42, Jan. 2018, doi: 10.31315/jik.v16i1.2681.

[2] A. Septiana, "Analisis Fungsi Partai Politik Pada Pilkada Musi Banyuasin 2017 (Studi Terhadap Partai Politik Pengusung Pasangan Dodi Reza Dan Beni Hernedi)," *jssp*, vol. 3, no. 1, pp. 28–41, Jun. 2019, doi: 10.19109/jssp.v3i1.4066.

[3] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web*, Hyderabad India: ACM, Mar. 2011, pp. 705–714. doi: 10.1145/1963405.1963504.

[4] P. Patmawati and M. Yusuf, "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *bits*, vol. 3, no. 3, pp. 122–129, Dec. 2021, doi: 10.47065/bits.v3i3.1012.

[5] I. W. D. Gafatia and N. Hadinata, "Analisis Pro Kontra Vaksin Covid 19 Menggunakan Sentiment Analysis Sumber

Media Sosial Twitter," *JPSII*, vol. 2, no. 1, pp. 34–42, Nov. 2021, doi: 10.47747/jpsii.v2i1.544.
[6] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.
[7] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Jurnal Eksplora Informatika*, vol. 9, no. 1, pp. 1–10, Sep. 2019, doi: 10.30864/eksplora.v9i1.237.
[8] E. I. Saptanti, "Analisis Manajemen Impresi Ma'ruf Amin dalam Debat Pilpres 2019," *ULTIMA Comm*, vol. 12, no. 2, pp. 262–284, Dec. 2020, doi: 10.31937/ultimacomm.v12i2.1573.
[9] Y. Sahria, "Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia," 2020.
[10] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *RESTI*, vol. 5, no. 1, pp. 123–131, Feb. 2021, doi: 10.29207/resti.v5i1.2587.
[11] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, Aug. 2018, doi: 10.33096/ilkom.v10i2.303.160-165.
[12] D. T. Anggraeni, "FORECASTING HARGA SAHAM MENGGUNAKAN METODE SIMPLE MOVING AVERAGE DAN WEB SCRAPPING," *jurnalmatrik*, vol. 21, no. 3, pp. 234–241, Dec. 2019, doi: 10.33557/jurnalmatrik.v21i3.726.
[13] D. Gunawan, "Metode Klasifikasi pada Data Preprocessing Data," no. 1, 2016.
[14] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," *Edutic*, vol. 7, no. 1, Nov. 2020, doi: 10.21107/edutic.v7i1.8779.
[15] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory." arXiv, Feb. 14, 2017. Accessed: May 25, 2023. [Online]. Available: http://arxiv.org/abs/1410.5329