JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Sentimen Terhadap Bakal Calon Presiden 2024 dengan Algoritma Naïve Bayes

Muhammad Raihan Fais Sya' bani*, Ultach Enri, Tesa Nur Padilah

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Singaperbangsa Karawang, Kabupaten Karawang, Indonesia
Email: [1,*]muhammad.raihan18129@student.unsika.ac.id, [2]ultach@staff.unsika.ac.id, [3]tesa.nurpadilah@staff.unsika.ac.id
Email Penulis Korespondensi: muhammad.raihan18129@student.unsika.ac.id
Submitted **05-04-2022**; Accepted **15-04-2022**; Published **29-04-2022**

Abstrak

Presiden Indonesia saat ini telah memegang jabatan yang sama selama 2 periode secara berturut – turut yang mana pada dasar peraturan untuk menjadi calon presiden sudah tidak bisa mencalonkan kembali menjadi presiden, dalam hal itu banyak lembaga survei yang telah mengeluarkan hasil survei terhadap beberapa tokoh yang memiliki elektabilitas untuk bisa menjadi calon presiden, berdasarkan hal tersebut juga banyak warganet yang menyampaikan pendapat, dari pendapat tersebut bisa dibuat kesimpulan mengenai sentimen warga masyarakat terhadap suatu tokoh bakal calon presiden tersebut dengan menggunakan metode *Knowledge Discovery from Data* dengan menggunakan algoritme *naïve bayes* dan perhitungan skor sentimen dengan harapan dari penelitian ini bisa memberikan bahan referensi kepada masyarakat dalam memilih presiden di pilpres yang akan datang. Hasil dari penelitian ini mendapatkan kesimpulan bahwa warganet memiliki sentimen positif terhadap setiap tokoh bakal calon presiden yang akan datang. Kemudian untuk hasil evaluasi dari algoritme *naïve bayes* yang didapatkan dari *dataset* pertama adalah 73,68 akurasi dan AUC 0,74 pada *fold* ke-7, *dataset* kedua adalah 71,43 untuk akurasi dan AUC 1,0 pada fold ke – 5, untuk *dataset* ketiga nilai akurasi yang didapat 60% dan AUC 0,92 pada *fold* ke-1, dan untuk *dataset* terakhir nilai akurasi yang didapatkan adalah 62,5% dan AUC 0,65 pada *fold* ke-3.

Kata Kunci: Bakal Calon Presiden; Pilpres; Knowledge Discovery from Data; Naïve bayes; Skor Sentimen

Abstract

President of Indonesia right now had been in position as president on 2 period in a row in this situation based on regulation President of Indonesia right now can't be candidate in the next election and based on this reason a lot of survey agency made some survey of some public figure that have good electability that could be candidate in the next election and for some reason a lot of netizen Indonesia made a comment on media social such as twitter, based on this comment conclusion can made drawn by sentiment analysis with knowledge discovery from data method with naïve bayes algorithm and sentiment score with the output of this research can give more knowledge to people for make a choice in next president election. The result of this research can made concluded that all public figures that had good electability before had good sentiment. For the evaluation result of naive bayes algorithm from first dataset is 73,68 for accuration and 0,74 for AUC at the seventh fold, on second dataset number of accuracy is 71,43 and AUC is 1,0 at the fifth fold, at the third dataset the number of accuracy is 60% and AUC is 0,92 at the first fold, and for the last dataset number of accuracy is 62,5% and AUC is 0,65 at the third fold.

Keywords: President candidate; President election; Knowledge discovery from data; Naïve bayes; Sentiment score

1. PENDAHULUAN

Presiden merupakan seseorang yang bertindak sebagai pemimpin dalam suatu negara dengan tugasnya yaitu sebagai kepala negara dan kepala pemerintahan. Selama 76 tahun Negara Indonesia telah berdiri, negara ini dipimpin oleh beberapa orang berbeda seperti Dr. Ir. H. Soekarno (1945 – 1967), Jendral TNI H.M. Soeharto (1967 – 1998), Prof. Dr. Ing. B. J. Habibie (1998 – 1999), K. H. Abdurrahman Wahid (Gusdur) (1999 – 2001), Megawati Soekarnoputri (2001 – 2004), Susilo Bambang Yudhoyono (2004 – 2009, 2009 – 2014), dan yang menjabat saat ini Ir. H. Joko Widodo (2014 – 2019, 2019 – Sekarang) [1]. Presiden di Indonesia dipilih melalui masyarakat dengan melalui proses demokrasi yaitu pemilihan presiden (pilpres) yang dilaksanakan setiap 5 tahun sekali. Menjadi seorang presiden memiliki beberapa persyaratan yang dimana persyaratan tersebut adalah seseorang tidak diperbolehkan menjadi presiden apabila orang tersebut sebelumnya telah menjadi presiden selama 2 periode secara berturut – turut [2], yang dalam hal ini presiden Indonesia saat ini sudah tidak bisa mencalonkan kembali menjadi Presiden pada pilpres selanjutnya yang akan terlaksana pada tahun 2024.

Berdasarkan tersebut banyak bermunculan survei elektabilitas terhadap beberapa tokoh publik yang memiliki elektabilitas baik yang menjadikan tokoh ini bisa dijadikan bakal calon presiden Indonesia di pilpres pada tahun 2024. Pada hasil dari beberapa survei tersebut memiliki hasil yang cukup beragam dimana peneliti mengambil 3 hasil survei dari lembaga survei berbeda yang ditemukan pada suatu lini berita yang memiliki judul "Zulhas Didorong Nyapres 2024, Survei Membuktikan..."[3]–[5]. Dari hasil ketiga survei tersebut peneliti melakukan pengambilan nilai rata – rata dari ketiga survei tersebut dengan hasilnya empat tokoh publik dengan nilai rata – rata elektabilitas terbesar adalah Ganjar Pranowo 19,2%, Anies Baswedan 14,2%, Prabowo Subianto 14%, Ridwan Kamil 9,84%. Dari hasil survei tersebut dan juga adanya lini berita yang sudah mempublikasi berita tersebut memiliki adanya kemungkinan penyampaian pendapat masyarakat atas hasil survei tersebut dengan media sosial.

Pengguna media sosial yang ada pada Indonesia saat ini sudah mencapai 170 juta pengguna di awal tahun 2021, dengan media sosial yang cukup favorit adalah twitter dengan pembuktiannya dengan hasil dari survei bahwa Indonesia menempati sebagai posisi ke-6 dengan peringkat dunia penggunaan twitter, selain dari itu di Indonesia pengguna dengan

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

jenjang umur 16 – 64 tahun mencapai tingkat 63,6% yang bersesuaian dengan sudah layak menjadi pemilih dalam pilpres [6].

Karena banyaknya pendapat yang disampaikan dari warganet Indonesia khususnya di media sosial twitter ini akan menghasilkan berbagai macam reaksi, maka dari hal itu diharapkan dari penyampaian pendapat ini dapat ditarik sebuah kesimpulan dari akumulasi reaksi pendapat warganet Indonesia dengan menggunakan metode sentimen analisis algoritme *naïve bayes*.

Berdasarkan hal tersebut banyak beberapa peneliti yang telah melakukan penelitian di bidang tersebut salah satu penelitian yang mendukung penelitian ini adalah penelitian dengan judul "*Top 10 Algorithm in Data Mining*" yang memberikan kesimpulan bahwa algoritme *naïve bayes* ini termasuk kedalam top 10 algorime [7], kemudian penelitian lainnya yang dilakukan oleh Yerik Afrianto Singgalen dengan penelitiannya yang berjudul "Pemilihan Metode dan Algoritma dalam Analisis Sentimen di Media Sosial : Sistematik Literature Review" dengan metodenya adalah sistematik literature review untuk melakukan pencarian terhadap beberapa penelitian tertentu dengan hasilnya disebutkan bahwa algoritme *naïve bayes* ini merupakan algoritme yang cukup dominan digunakan untuk penelitian analisis sentimen pada media sosial [8], yang dimana dalam hal ini bersesuaian dengan pengajuan penelitian ini.

Selain itu banyak juga beberapa peneliti yang telah melakukan penelitian pada bidang ini seperti penelitian yang dilakukan oleh Yulita dan kawan – kawan pada analisis sentimen opini masyarakat terhadap vaksin covid – 19 dengan menggunakan algoritme *naïve bayes* yang menghasilkan opini masyarakat lebih cenderung ber sentimen positif dengan opini positif memiliki persentase sebesar 60,3%, untuk negatif 5.4%, dan netral sebesar 34,4%. Kemudian hasil dari penggunaan algoritme ini menghasilkan akurasi sebesar 0,93 atau 93% [9].

Penelitian lainnya dilakukan oleh Muzaki dan Witanti yang melihat sentimen masyarakat terhadap pemerintah mengenai pelaksanaan pilkada pada tahun 2020 dengan adanya wabah covid – 19. Data yang didapatkan oleh Muzaki didapatkan pada media sosial twitter dengan hasil evaluasi dari penggunaan algoritme *naïve bayes* untuk data training adalah 92,3%, *validation* 50,79%, data testing 54,1% [10].

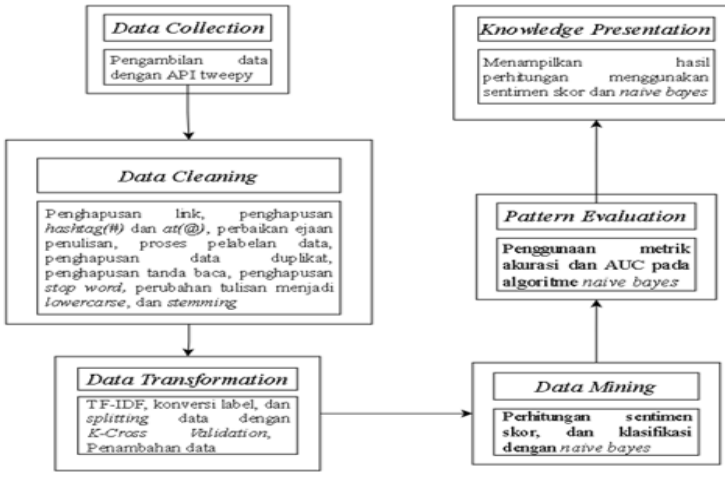
Penelitian lainnya yang dilakukan Winda dan kawan – kawan adalah melakukan analisa sentimen pada opini masyarakat terhadap vaksinasi covid – 19 dengan algoritme *naïve bayes* dengan data yang digunakna berasal dari *twitter* yang didapatkan total jumlah data sebanyak 3.780 *tweet*. Hasil evaluasi dari penelitian ini berupa opini masyarakat yang dominan terhadap sentimen positif sebesar 2.278 (60,3%), netral 1.299 (34,4%), dan negatif 203 (5,4%). Kemudian nilai evaluasi yang didapatkan setelah menggunakan algoritme *naïve bayes* adalah 0,93 (93%) [9].

Penelitian selanjutnya dilakukan oleh Santoso dan Nugroho mengenai analisis sentimen yang dilakukan terhadap calon presiden Indonesia pada tahun 2019 yang menggunakan sumber data dari media sosial *facebook*. Pengumpulan data pada media sosial tersebut dilakukan dengan menggunakan syarat dari tanggal 17 April 2019 hingga 22 May 2019. Hasil dari penelitian ini membuktikan bahwa Joko Widodo memiliki polaritas sentimen yang lebih unggul daripada Prabowo Subianto [11].

Berdasarkan penjelasan tersebut penelitian ini akan dilakukan sebuah analisa sentimen terhadap bakal calon presiden 2024 dengan menggunakan algoritme *naïve bayes* yang nantinya hasil penelitian ini dapat dimanfaat masyarakat sebagai bahan referensi dalam memilih pemimpinnya di kemudian hari pada tahun 2024.

2. METODOLOGI PENELITIAN

Pada proses penelitian ini peneliti menggunakan metode *Knowledge Discovery from Data* dengan tahapannya yakni *Data Cleaning, Data Transformation, Data Mining, Pattern Evaluation,* dan *Knowledge Presentation*. Selain penggunaan metode tersebut ditambahkan satu tahapan metode yakni *Data Collection* yang bertujuan untuk melakukan pengumpulan data pada media sosial *twitter*. Untuk mendapatkan pemahaman yang lebih mudah akan diberikan keterangan gambar dari alur penelitian ini pada gambar 1.



Gambar 1. Alur Penelitian

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

2.1 *Data Collection*

Untuk menunjan penelitian ini diperlukan data cuitan dari warganet yang berasal dari media sosial *twitter*, maka dari itu dilakukan pengumpulan data dengan menggunakan API yang telah disediakan oleh pihak *twitter* dengan bantuan dari library *tweepy* dari bahasa pemrograman python. Pengumpulan data ini dilakukan dengan menggunakan beberapa aturan seperti waktu cuitan yang dibatasi pada tanggal 1 Agustus 2021 sampai dengan 3 Februari 2022. Selain aturan tersebut peneliti menggunakan *hashtag* atau tagar untuk mempersempit pencarian data yang diperlukan. Untuk penggunaan rumus tagar yang digunakan terdapat persamaan 1.

(#capres2024 AND (#ganjarpranowo OR #aniesbaswedan OR #prabowo OR #ridwankamil) (1)

2.2 *Data Cleaning*

Dikarenakan data yang dikumpulkan pada penelitian adalah data langsung yang ditemukan pada media sosial twitter ditakutkan ada beberapa data yang masih bersifat kotor dan diperlukan pembersihan data lebih lanjut, maka dari itu peneliti melakukan pembersihan dengan beberapa tahapan yang akan dilalui sebagai berikut:

2.2.1 Penghapusan Link, *Hashtag*(#), dan *At*(@)

Terkadang pada suatu cuitan warganet masih banyak terdapat sebuah link seperti ini https://t.co/do53YD4avs, ataupun *hashtag* yang dimana proses sebelumnya peneliti melakukan pencarian dari sebuah cuitan dengan menggunakan tagar maka dari itu pasti akan terdapat sebuah tagar pada data yang akan digunakan seperti #Capres2024, selain dari kedua hal tersebut warganet masih melakukan penandaan terhadap beberapa akun lainnya seperti @ganjarpranowo. Dalam hal ini ketiga hal tersebut dapat mengganggu proses analisa yang akan dilakukan.

2.2.2 Perbaikan Ejaan

Pada proses pembersihan selanjutnya masih banyak ditemukan dari data yang tidak menggunakan ejaan yang disempurnakan (EYD), maka dalam hal ini perlu diperbaiki untuk mengurangi ambiguitas pada proses penelitian ini. Proses ini nantinya akan dibantu oleh guru bahasa indonesia tingkat SMA.

2.2.3 Pelabelan Data

Penelitian yang akan dilaksanakan nantinya menggunakan metode klasifikasi, yang dimana akan dilakukan proses pemetaan suatu himpunan atribut objek pada label kelas tertentu. Dikarenakan penggunaan metode tersebut maka diperlukan suatu label kelas dan dikarenakan data yang didapatkan masih bersifat mentah maka akan dilakukan pelabelan yang akan divalidasi oleh guru bahasa indonesia tingkat SMA sebagai ahli tata bahasa Indonesia.

2.2.4 Penghapusan Data Duplikat

Pada beberapa cuitan terdapat beberapa akun yang melakukan cuitan ulang terhadap cuitan yang pernah diposting oleh user lain sebelumnya ataupun ada beberapa cuitan yang diposting lebih dari satu kali oleh akun yang sama, dikarenakan hal tersebut mengakibatkan adanya data duplikat yang akan mengakibatkan hasil dari proses penelitian ini menjadi kurang baik, maka dari itu akan dilakukan penghapusan data duplikat tersebut.

2.2.5 Penghapusan Tanda Baca dan Perubahan *Lowercase*

Walaupun sudah melewati tahap pembersihan data pertama masih terdapat data yang memiliki tanda baca yang harus dihapus selain itu juga dilakukan perubahan bentuk huruf menjadi *lowercase* (huruf kecil) dilakukan agar menyamakan bentuk huruf dari setiap kalimat [12].

2.2.4 Menghapus *Stopword*

Terdapat beberapa kalimat yang tidak terlalu memiliki arti dalam suatu kalimat yang dalam hal ini juga dapat mempermudah dalam proses komputasi pada data mining maka bentuk kata (*stopword*) seperti ini akan dihapus [12].

2.2.5 Stemming

Pada suatu cuitan terdapat beberapa kata yang sudah ditambahkan afiks yang dalam hal ini akan menambahkan keragaman kamus kata, maka dalam hal ini kata tersebut diubah menjadi bentuk baku atau dasarnya dengan menggunakan metode *stemming* [12].

2.3 *Data Transformation*

Pada proses data yang telah melalui pembersihan tersebut masih bersifat tekstual sementara pada tahapan data mining diperlukan data yang bersifat numerik maka dari itu diperlukan proses perubahan data dengan menggunakan beberapa metode sebagai berikut

2.3.1 Transformasi Atribut

Pada prosesnya peneliti menggunakan sebuah metode yang dapat merubah kumpulan teks menjadi sebuah numerik menggunakan sebuah metode *term frequency – inverse document frequency* (TF – IDF). TF – IDF adalah sebuah metode

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

gabungan antara metode TF dan IDF yang dimana TF adalah rasio dari jumlah suatu kata pada kalimat dan dibandingkan dengan panjang dari kalimat tersebut, sementara IDF dari setiap kata adalah rasio berdasarkan total dokumen dengan jumlah dari dokumen tertentu yang terdapat teks tersebut [13]. Persamaan dari metode ini akan dituliskan pada persamaan 2 [9], [14].

$$tf - idf = tf \times \log \frac{N}{df} \quad (2)$$

Keterangan
tf : *Term Frequency*
idf : *Inverse Document Frequency*
N : Jumlah total dokumen dalam corpus N
df : Jumlah dokumen yang mengandung term sebuah kata

2.3.2 Transformasi Label

Selain atribut yang perlu diubah data dari label juga perlu diubah kedalam bentuk numerik yang dimana nanti data akan diubah menjadi sebuah angka yang dimulai 0 kemudian disesuaikan dengan jumlah label yang dimiliki yang dalam hal ini menggunakan *label encoder*[15].

2.3.4 Pembagian Data

Pada proses pembagian data akan digunakan metode *k – fold cross validation* dengan melakukan pembagian data ke dalam *k* grup atau *folds*, dari perkiraan pembagian nilai ukuran yang setara. Pada *fold* pertama akan dijadikan sebagai data validasi dan sisa *fold* lainnya akan dijadikan data *train*, hal tersebut akan terus beriterasi hingga seluruh bagian *fold* data telah digunakan sebagai data validasi [16], [17]. Penggunaan nilai *K* pada metode tersebut adalah 10 dan apabila ditemukan jumlah data yang tidak memungkinkan maka nilai *K* akan disesuaikan dengan jumlah data.

2.4 *Data Mining*

Pada proses *data mining* akan digunakan dua metode yakni penggunaan algoritme *naïve bayes* dan perhitungan skor sentimen untuk mendapatkan analisa sentimen yang diperlukan. Algoritme *naïve bayes* adalah salah satu teknik klasifikasi yang menggabungan penggunaan ilmu dari statistika dan teori kemungkinan (probabilitas) yang digunakan untuk menyelesaikan permasalahan kasus *supervised learning*. Perumusan dasar dari algoritme ini terdapat pada persamaan 3.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

Keterangan:
X : Sampel data belum diketahui labelnya
H : Hipotesis bahwa x merupakan data label
P(H) : Peluang dari hipotesa H
P(X) : Peluang data sampel
P(X|H) : Peluang data sampel apabila hipotesis benar
P(H|X) : Peluang hipotesis berdasarkan keadaan sampel

Selain dari penggunaan algoritme *naïve bayes* penggunaan skor sentimen juga digunakan dalam penelitian ini. Skor sentimen dilakukan dengan melakukan perhitungan jumlah kata sifat positif dalam suatu kalimat dikurang dengan jumlah kata sifat negatif dari suatu kalimat atau dapat dirumuskan pada persamaan 4 [18], [19].

$$Skor = \sum kata\ positif - \sum kata\ negatif \quad (4)$$

2.5 *Pattern Evaluation*

Setelah melalui proses data mining dengan menggunakan algoritme *naïve bayes* diperlukan evaluasi terhadap model yang telah dihasilkan dari penggunaan algoritme tersebut, proses evaluasi juga menggunakan beberapa metrik, peneliti menggunakan 2 nilai metrik untuk mengevaluasi yakni metrik akurasi dan *area under curve* (AUC).

2.5.1 Akurasi

Metrik akurasi adalah nilai persentase yang didapatkan dari set tupel data yang telah diklasifikasi secara benar oleh model. Metrik akurasi ini memiliki fungsi untuk menilai model dalam seberapa baik dalam mengenali tupel dari berbagai kelas. Persamaan dari metrik ini terdapat pada persamaan 5 [20].

$$akurasi = \frac{TP+TN}{P+N} \quad (5)$$

2.5.2 *Area Under Curve* (AUC)

Metrik ini menggunakan perhitungan dari sebagian nilai dari satuan luas persegi yang berasal pada hasil grafik *Receiver Operating Characteristic* (ROC) *Curve* yang hasil dari AUC akan selalu berada pada 0,0 dan 1,0. Pada sumbu-x ROC *Curve* menggunakan nilai dari *False Positive* (FP) *Rate* dan pada sumbu-y menggunakan nilai dari *true positive* (TP) *rate* dengan persamaan pada setiap nilai tersebut terdapat pada persamaan 6 dan 7 [21].

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$TPRate = \frac{TP}{TP+FN} \quad (6)$$

$$FP\ Rate = \frac{FP}{FP+TN} \quad (7)$$

Metrik AUC juga bisa dijadikan acuan sebagai penentu kualitas model dalam mengklasifikasikan suatu permasalahan yang dijelaskan pada tabel 1 [21].

Tabel 1. Kualitas Model dengan AUC

Nilai AUC	Kualitas
0.90 – 1.00	*Excellent classification*
0.80 – 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.70	*Poor classification*
0.50 – 0.60	*Failure*

2.6 *Knowledge Presentation*

Setelah dilakukan pencarian hasil evaluasi terbaik kemudian akan digunakan model dengan nilai evaluasi terbaik kemudian digunakan model tersebut untuk mengklasifikasi dari setiap data yang hasil dari klasifikasi ini akan ditampilkan perbandingannya dengan klasifikasi *naïve bayes.* Kemudian hasil dari skor sentimen tersebut juga akan dibandingkan pada setiap tokoh yang telah diusulkan.

3. HASIL DAN PEMBAHASAN

Pelaksanaan penelitian ini didasari dengan metode KDD yang sebelumnya telah dijelaskan dengan menggunakan algoritme *naïve bayes* dan juga menggunakan skor sentimen untuk memperhitungkan nilai sentimen dari setiap cuitan dengan menggunakan data cuita dari setiap tokoh publik bakal calon presiden 2024 yang diharapkan dari hasil penelitian ini bisa dijadikan sebagai bahan referensi masyarakat untuk pilpres 2024.

3.1 *Data Collection*

Pengumpulan data yang dilakukan menggunakan rentang waktu cuitan dari tanggal 1 Agustus 2021 dan 3 Februari 2022 dengan hasil jumlah data yang didapatkan yakni 533 cuitan dengan 274 cuitan dari Ganjar Pranowo, 120 cuitan dari Anies Baswedan, 72 cuitan dari Prabowo Subianto, dan 67 dari Ridwan Kamil. Hasil dari pengambilan data ini akan ditampilkan pada tabel 2.

Tabel 2. Hasil Penarikan Data

Dataset	Jumlah Data
Ganjar Pranowo	274
Anies Baswedan	120
Prabowo Subianto	72
Ridwan Kamil	67
Total	533

Pada proses penarikan data peneliti juga mengambil beberapa atribut data lainnya yang dianggap penting untuk diperlukannya peninjauan ulang terhadap data cuitan yang telah dikumpulkan sebelumnya, akan tetapi untuk penggunaan pada tahapan selanjutnya atribut yang digunakan hanya *text*. Jenis atribut yang digunakan akan disajikan dalam tabel 3.

Tabel 3. Atribut *Dataset* Cuitan Twitter

Nama Atribut	Penjelasan Atribut	Tipe Data Atribut
created_at	Waktu dan tanggal dari pembuatan cuitan tersebut	*Date and time*
name	Nama dari akun yang membuat cuitan tersebut	*String*
screen_name	Username dari akun twitter yang membuat cuitan tersebut	*String*
text	Cuitan dari warganet	*String*
description	Deskripsi dari akun yang membuat cuitan tersebut	*String*
hashtags	Kumpulan hashtag yang digunakan oleh cuitan tersebut	*List*

3.2 *Data Cleaning*

Pada proses ini akan dilakukan beberapa tahapan pembersihan yang sudah disebutkan dalam metodologi penelitian. Pada hasil dari tahapan ini akan ditampilkan dalam bentuk tabel. Pada tahapan pembersihan dari Awal hingga pembersihan perbaikan ejaan akan ditampilkan pada tabel 4.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 4. Pembersihan Data

Tahapan	Sebelum	Sesudah
Menghapus link, *hashtag(#)*, dan *at(@)*	Ganjar Pranowo Terus Berkomitmen Turunkan Kemiskinan Di Jateng, Maksimalkan Program RSLH . . #GanjarUntukIndonesia #GanjarPranowo #Ganjar #GanjarKu #GanjarPedia #JatengGayeng #BeritaTerkini #beritahariini @ganjarpranowo #Capres2024 #KerjaBarengGanjar https://t.co/do53YD4avs	Ganjar Pranowo terus berkomitmen turunkan kemiskinan di Jateng, maksimalkan program RSLH
Perbaikan ejaan	Top News Koran Rakyat Merdeka Belum Ada Kendaraan Buat **Nyapres** Politisi Gerindra Yakin Anies Maju **Nyagub** Lagi	Top News Koran Rakyat Merdeka belum ada kendaraan buat **mencalonkan presiden** dari Politisi Gerindra yakin Anies maju **mencalonkan gubernur** lagi
Penghapusan tanda baca dan perubahan bentuk huruf	Merdeka PDIP-Gerindra kian lengket semakin kuatkah Prabowo-Puan?	merdeka pdipgerindra kian lengket semakin kuatkah prabowopuan
Penghapusan *stopword*	siap menjadi capres lahir batin ridwan kamil warnanya **saya** kabari **nanti**	siap menjadi capres lahir batin ridwan kamil warnanya kabari
stemming	sidak jalan tol banyak lubang ganjar minta **pengelola lakukan perbaikan**	sidak jalan tol banyak lubang ganjar minta **kelola laku baik**

Tahapan selanjutnya adalah pemberian label terhadap data dengan menggunakan bantuan dari ahli tata bahasa guru bahasa indonesia tingkat SMA, dengan mendapatkan hasil sentimen yang cukup beragam dari berbagai tokoh yang diusulkan, dengan hasil data pada tabel 5.

Tabel 5. Jumlah Label Dataset Validator

Data	Label			Jumlah Data
	Positif	Netral	Negatif	
Ganjar Pranowo	178	54	31	263
Anies Baswedan	51	16	15	82
Prabowo Subianto	25	5	29	59
Ridwan Kamil	48	8	4	60

Selanjutnya pada tahapan lainnya terdapat penghapusan data duplikat dengan salah satu bentuk contoh duplikat seperti “Cuitan ulang dari Andesepen menyebutkan bahwa tinggal dikawinkan Anies dan Emil pasangan lengkap”, kemudian hasil dari perubahan kuantitas jumlah data disajikan pada tabel 6.

Tabel 6. Hasil Penghapusan Data Duplikat

Data	Sebelum	Sesudah	Selisih
Ganjar	237	189	74
Anies	81	58	23
Prabowo	59	47	12
Ridwan	60	33	27

3.3 *Data Transformation*

Perubahan data atribut yang telah melewati metode akan berubah menjadi matrik numerik dengan dimensi jumlah cuitan × jumlah korpus kata yang menghasilkan sebuah matriks dengan dimensi yang cukup besar pada setiap data dan terdapat kesulitan dalam menampilkan data dalam jumlah besar tersebut. Setelah atribut di ubah menjadi bentuk numerik dilanjutkan dengan perubahan data dari label sendiri dengan menggunakan metode dari *label encoder* setiap label akan di ubah menjadi numerik dimana pemetaan perubahan data tersebut akan ditampilkan pada tabel 7.

Tabel 7. Hasil Pemetaan Konversi Label

Sebelum	Sesudah
negatif	0
netral	1
positif	2

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Setelah data atribut dan label diubah menjadi bentuk numerik, tahapan selanjutnya yang dilakukan adalah melakukan pembagian data dengan menggunakan metode *k-fold cross validation* akan tetapi setelah beberapa tahapan dari *data cleaning* banyak jumlah data yang berkurang menyebabkan ada perbedaan nilai *K*. Pada data Ganjar memiliki nilai *K* = 10, untuk data Anies *K* = 8, data Prabowo *K* = 5, dan dataset Ridwan memiliki nilai *K* = 4.

3.4 *Data Mining*

Pada tahapan ini nantinya akan dilakukan dua kali tahapan proses yakni dengan menggunakan algoritme *naïve bayes* untuk tugas klasifikasi dan perhitungan skor sentimen dengan melakukan perhitungan terhadap jumlah kata sifat positif dan jumlah kata negatif.

3.4.1 *Naïve Bayes*

Pada penggunaan algoritme ini akan menggunakan beberapa iterasi yang disesuaikan dengan jumlah *fold* yang telah didefinisikan bersesuaian dengan jumlah data pada tahapan *data transformation* sebelumnya, hasil dari proses ini nantinya akan berbentuk nilai numerik dengan 3 hasil, yang merupakan nilai probabilitas dari setiap label tersebut, kemudian dicari nilai terbesar dari ketiga probabilitas tersebut dan ditentukan nilai dari label tersebut. Ringkasan hasil dari penggunaan algoritme ini akan ditampilkan pada tabel 8.

Tabel 8. Ringkasaan Dataset dengan Algoritme Naïve Bayes

Cuitan	Prediksi				Label
	Kelas 0	Kelas 1	Kelas 2	Final	
['suara' 'tingkat' 'pengaruh' 'juang' 'pdi' '2024' 'pilpres' 'prabowo' 'calon' 'wacana']	0,39	0,11	0,49	2	positif
['prematur' 'terlalu' 'anggap' 'gerindra' 'imin' 'prabowocak' 'pasang' 'wacana']	0,41	0,14	0,45	2	positif
['bicara' 'asal' 'nilai' 'negara' 'bangsa' 'aset' 'selamat' 'maju' 'partai' 'umum' 'ketua' 'tuju' 'gerindra' '2024' 'pilpres' 'prabowo']	0,45	0,11	0,44	0	negatif
['deh' 'dulu' 'cak' 'kalau' 'nahdliyin' 'ideal' 'wapres' 'imin' 'prabowo']	0,39	0,15	0,46	2	positif
['tentu' 'efek' 'jokowi' 'utakatik' '2024' 'pilpres']	0,36	0,14	0,50	2	positif
['banget' 'girang' 'favorit' 'capres' 'ppi' 'prc' 'sigi' 'survei' 'gerindra' 'prabowo']	0,32	0,10	0,57	2	positif

3.4.2 Skor Sentimen

Pada implementasi dari perhitungan ini, peneliti melakukan pengumpulan kata dengan melakukan ekstraksi dari semua *dataset* sebelumnya yang didapatkan 1063 kata, kemudian dilakukan penghapusan beberapa kata yang tidak memiliki makna ataupun mengandung unsur nama orang atau nama tempat dan dihasilkan jumlah data menjadi 671 kata, kemudian dilakukan kembali pemberian label terhadap setiap kata yang dilakukan oleh ahli tata bahasa dan dihasilkan kata yang memiliki makna positif sebanyak 508 kata dan 163 kata yang memiliki makna negatif.

3.5 *Pattern Evaluation*

Setelah dilakukan proses tahapan data mining diperlukannya evaluasi terhadap penggunaan algoritme *naïve bayes* sebelumnya untuk bisa menentukan mana model terbaik yang bisa digunakan dari setiap model untuk setiap *dataset*, maka hasil dari setiap model tersebut akan ditampilkan pada tabel 9.

Tabel 9. Hasil Evaluasi Setiap *Dataset*

Dataset	*Fold* ke -	Akurasi	AUC
Ganjar	7	73,68%	0,74
Anies	5	71,43%	1,0
Prabowo	1	60%	0,92
Anies	3	62,5%	0,65

3.6 *Knowledge Presentation*

Hasil dari kedua metode yang digunakan sebelumnya akan ditampilkan pada tahapan ini dimana hasil dari klasifikasi pada algoritme *naïve bayes* akan dilakukan perhitungan dari setiap label dengan memberikan nilai dikalikan satu setiap data sentimen positif, dikalikan negatif satu setiap sentimen negatif, dan dikalikan dengan nol jika bersentimen netral kemudian di normalisasi dengan dibagi jumlah data pada setiap *dataset.* Hasil implementasi ini akan ditampilkan pada tabel 10.

Tabel 10. Hasil Perhitungan Nilai Sentimen setelah Menggunakan Naïve Bayes

Dataset	Perhitungan Nilai Sentimen				Normalisasi
	Positif × 1	Negatif × -1	Netral × 0	Total	
Ganjar	177 × 1 =177	0 × -1 = 0	12 × 0 = 0	177	177/189 = 0,9365

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3989
Hal 265−273
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Anies	56 × 1 = 56	2 × -1 = -2	0 × 0 = 0	54	54/58 = 0,931
Prabowo	32 × 1 = 32	15 × -1 = -15	0 × 0 = 0	17	17/47 = 0,3617
Ridwan	33 × 1 = 33	0 × -1 = 0	0 × 0 = 0	33	33/33 = 1

Kemudian untuk tahapan sentimen skor setelah ditemukan jumlah sentimen skor pada setiap dataset akan dilakukan pembagian kembali dengan jumlah data setiap *dataset* untuk melihat nilai skor sentimen yang bersesuaian dengan jumlah data setiap dataset. Hasil tersebut ditampilkan pada tabel 11.

Tabel 11. Hasil Normalisasi Skor Sentimen

Dataset	Jumlah Skor Sentimen	Rata – rata Skor Sentimen
Ganjar	732	3,873
Anies	232	4
Ridwan	161	4,878
Prabowo	120	2,533

4. KESIMPULAN

Setelah dilakukan proses penelitian dapat dilihat bahwa warganet memiliki sentimen positif dari setiap dataset karena terlihat bahwa jumlah label sentimen terbanyak yang dimiliki oleh setiap dataset adalah sentimen positif, kemudian penggunaan algoritme *naïve bayes* memiliki nilai evaluasi yang beragam dari setiap *dataset* yang terburuk dimiliki oleh *dataset* Ridwan dengan nilai akurasi 62,5% dan AUC 0,65, kemudian dataset Prabowo dengan akurasi 60% dan AUC 0,92, dataset dari Anies dengan akurasi 71,43% dan AUC 1,0, dan dataset dari Ganjar dengan akurasi sebesar 73,68% dan AUC 0,74. Hasil dari nilai sentimen yang dihasilkan melalui algoritme *naïve bayes* juga beragam dengan nilai terbesar dari *dataset* Ridwan dengan nilainya 1, kemudian *dataset* Ganjar 0,9365, *dataset* Anies senilai 0,931, dan *dataset* Prabowo 0,3617. Selain dari hasil tersebut digunakan juga nilai dari skor sentimen dengan nilai terbesar dari *dataset* Ridwan dengan nilainya adalah 4,878, kemudian *dataset* Anies dengan nilai 4, *dataset* Ganjar dengan nilai 3,879, dan *dataset* Prabowo dengan nilai skor sebesar 2,533.

REFERENCES

[1] Apif Supriadi and Fatmasari, "Implementasi metode klasifikasi naive bayes pada sistem analisis opini pengguna twitter berbasis web," *J. Sist. Inf.*, vol. 10, no. 1, pp. 46–54, 2021, doi: 10.51998/jsi.v10i1.356.
[2] Pemerintah Indonesia, *Undang - Undang Republik Indonesia Nomor 42 Tahun 2008 Tentang Pemilihan Umum Presiden dan Wakil Presiden*. Jakarta, Indonesia: Sekretariat Negara, 2008.
[3] E. Safitri, "Zulhas didorong nyapres 2024, survei membuktikan...," *detikNews*, 2021. https://news.detik.com/berita/d-5705142/zulhas-didorong-nyapres-2024-survei-membuktikan (accessed Dec. 06, 2021).
[4] NewIndonesia Research & Consulting, "Baliho puan bertebaran, ganjar kian berkibar," *Liputan Media*, 2021. https://newindoresearch.com/baliho-puan-bertebaran-ganjar-kian-berkibar/ (accessed Dec. 06, 2021).
[5] N. A. Akbar and N. Nasrullah, "Nama-nama ini konsisten di 6 besar survei capres 2024," *republika.co.id*, 2021. https://www.republika.co.id/berita/qxvpxm440/namanama-ini-konsisten-di-6-besar-survei-capres-2024 (accessed Dec. 06, 2021).
[6] we are social and Hootsuite, "Digital 2021 Indonesia," Canada, New York, 2021. [Online]. Available: https://www.slideshare.net/DataReportal/digital-2021-indonesia-january-2021-v01.
[7] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008, doi: 10.1007/s10115-007-0114-2.
[8] Y. A. Singgalen, "Pemilihan metode dan algoritma dalam analisis sentimen di media sosial : sistematic literature review," *J. Inf. Syst. Informatics*, vol. 3, no. 2, pp. 278–302, 2021, doi: 10.33557/journalisi.v3i2.125.
[9] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Analisis sentimen terhadap opini masyarakat tentang vaksin covid - 19 menggunakan algoritma naïve bayes classifier," *JDMSI*, vol. 2, no. 2, pp. 1–9, 2021.
[10] A. Muzaki and A. Witanti, "Sentiment analysis of the community in the twitter to the 2020 election in pandemic covid-19 by method naive bayes classifier," *J. Tek. Inform.*, vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.
[11] E. B. Santoso and A. Nugroho, "Analisis sentimen calon presiden indonesia 2019 berdasarkan komentar publik di facebook," *Eksplora Inform.*, vol. 9, no. 1, pp. 60–69, 2019, doi: 10.30864/eksplora.v9i1.254.
[12] V. Kalra and R. Aggarwal, "Importance of text data preprocessing & implementation in rapidminer," *Proc. First Int. Conf. Inf. Technol. Knowl. Manag.*, vol. 14, no. July, pp. 71–75, 2018, doi: 10.15439/2017km46.
[13] A. Kulkarni and A. Shivananda, *Natural language processing recipes*. New York: Springer Science+, 2019.
[14] T. I. Saputra and R. Arianty, "Implementasi algoritma k-means clustering pada analisis sentimen keluhan pengguna indosat," *J. Ilm. Inform. Komput.*, vol. 24, no. 3, pp. 191–198, 2019, doi: 10.35760/ik.2019.v24i3.2361.
[15] F. Pedregosa *et al.*, "Scikit-learn: machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Second Edi. New York: Springers, 2021.
[17] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.
[18] K. L. Bansal and S. V. S. Pathania, "A dynamic approach to improve the sentiment score using machine learning algorithms on twitter," *Purva Mimaansa*, vol. 12, no. September, pp. 1–8, 2021.

[19] N. Singh, N. Sharma, and A. Juneja, "Sentiment score analysis for opinion mining," *Adv. Intell. Syst. Comput.*, vol. 748, pp. 363–374, 2019, doi: 10.1007/978-981-13-0923-6_32.
[20] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Third. United States of America: Elsevier Inc., 2012.
[21] F. Gorunescu, *Data mining concepts, model, and techniques*. Chennai, India: Springer, 2010.