SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK IV - Surabaya, 27 April 2024
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2024.5806

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Penerapan Text Mining untuk Menganalisis Sentimen di Twitter Mengenai Vaksin Covid 19 Sinovac

Budanis Dwi Meilani, Ang Anthony Purnomo, Anggi Yhurinda Perdana Putri

Sistem Informasi Institut Teknologi Adhi Tama Surabaya
budanis@itats.ac.id

ABSTRACT

Sinovac Vaccine or CoronaVac refers to a coronavirus vaccine that has been developed by a private company in China. Its implementation in Indonesia has pros and cons. Although the Drug and Food Supervisory Body has issued the legal permit for its emergency use, many people still doubt the effectiveness of the Sinovac vaccine and even complain about the symptoms caused by this vaccine. People give various comments concerning the Sinovac vaccine on social media, including Twitter. Basically, those comments can be managed by sentiment analysis so that useful information for numerous parties can result. This study employed orange data mining software to collect data on Twitter using the keyword Sinovac vaccine. Following that, the data were manually labeled and went through a pre-processing phase to properly arrange and process imperfect data. Next, the data were processed into numbers by weighting words through the method of Term Frequency-Inverse Document Frequency and then classifying them using Naive Bayes. The researcher carried out five trials and gained an average ratio comparison between testing data and training data of 78.54%, at the highest accuracy value of 80.45% in the comparison of 90:10 (90% training data and 10% testing data).

Keywords: *sentiment analysis, mining text, Term Frequency – Inverse Document Frequency, Naive Bayes Classifier, Sinovac vaccine*

ABSTRAK

Vaksin *Sinovac* atau *CoronaVac* adalah vaksin virus *corona* yang dikembangkan oleh perusahaan swasta China. penggunaan vaksin sinovac menjadi pro kontra di Indonesia. Meskipun Badan Pengawas Obat dan

Makanan (BPOM) telah mengeluarkan izin penggunaan darurat, banyak masyarakat yang masih meragukan tingkat efektivitas vaksin *sinovac* dan mengeluh tentang gejala yang ditimbulkan setelah vaksinasi. Masyarakat banyak memberikan tanggapan tentang vaksin *sinovac* di media sosial, salah satunya media sosial *twitter*. Dari tanggapan tersebut dapat diolah dengan analisis sentimen sehingga bisa menjadi informasi yang bermanfaat bagi beberapa pihak. Dalam penelitian ini pengumpulan data menggunakan *software orange data mining* dengan kata pencarian vaksin *sinovac* di *twitter* lalu data yang sudah didapatkan di label secara manual. Selanjutnya data akan diolah menggunakan proses *preprocessing* yang bertujuan untuk menata dan mengolah dengan baik sebuah teks yang belum sempurna. setelah itu data akan diolah menjadi angka dengan cara membobotkan kata menggunakan metode *Term Frequency – Inverse Document Frequency* dan diklasifikasi menggunakan *naïve bayes*. uji coba dilakukan sebanyak lima kali dengan perbandingan rasio data latih dan data uji. dari kelima perbandingan rasio data uji dan latih rata – rata yang diperoleh dari kelima perbandingan tersebut adalah 78,54% dengan nilai akurasi tertinggi 80,45% pada perbandingan 90:10(90% data latih dan 10%data uji).

Kata kunci: Analisa Sentimen, Teks Mining, *Term Frequency – Inverse Document Frequency*, *Naive Bayes Classifier*, Vaksin *Sinovac*.

PENDAHULUAN

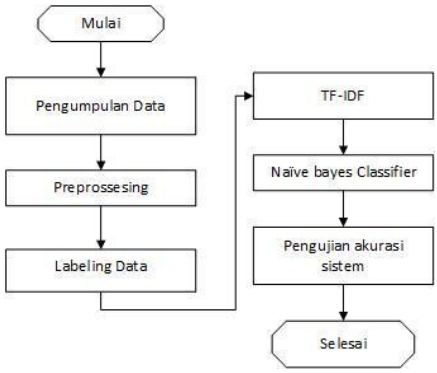
Ditengah Penyebaran Virus COVID-19 ini, muncul istilah herd immunity atau kekebalan kelompok yang dipercaya dapat membantu menekan penyebaran virus COVID-19. herd immunity atau kekebalan kelompok adalah kondisi ketika sebagian besar orang dalam suatu kelompok telah memiliki kekebalan terhadap penyakit infeksi tertentu. Salah satu cara untuk mendapatkan kekebalan kelompok pada COVID-19 ini adalah dengan vaksinasi. salah satu vaksin COVID-19 yang digunakan di Indonesia adalah vaksin sinovac.[1] Penggunaan vaksin sinovac menjadi pro kontra di Indonesia. Meskipun Badan Pengawas Obat dan Makanan telah mengeluarkan izin penggunaan darurat vaksin sinovac, banyak masyarakat yang masih meragukan tingkat efektivitas vaksin sinovac dan mengeluh tentang gejala yang ditimbulkan setelah vaksinasi. Masyarakat banyak memberikan tanggapan tentang vaksin sinovac di media sosial, salah satunya media sosial twitter.

Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes[2][3]. Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan oleh Xhemali, Hinde Stone dalam jurnalnya (*Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages)* mengatakan bahwa (*Naïve Bayes Classifier* memiliki tingkat akurasi yg lebih baik dibanding model *classifier* lainnya).[4] Penelitian terkait tentang Analisis sentimen menggunakan metode naïve bayes di twitter dengan judul (Sentimen Analisis Terkait ―Lockdown pada Sosial Media Twitter) yang di teliti oleh Adilah, T., Alkhalifi, Y., Mayangky, N. A., & Gata, W. pada penelitian tersebut meneliti tentang analisis sentiment terkait lockdown pada media sosial twitter menggunakan labeling vader dan menggunakan pembobotan kata tf-idf dengan metode *Naïve Bayes Classifier* dan *Support Vector Machine*.[5] hasil akurasi yang didapatkan dari kedua algoritma tersebut adalah lebih dari 80%.

METODE

Tahapan Penelitian

Untuk melakukan penelitian supaya jelas diperlukan tahapan secara urut dan benar agar proses penelitian dapat tersusun dengan baik sehingga bisa diketahui alur tujuan penelitian yang diharapkan. Berikut alur dan tahapan metodologi penelitian pada gambar 1 dibawah ini:



Gambar 1. Tahapan Penelitian

Pengumpulan data

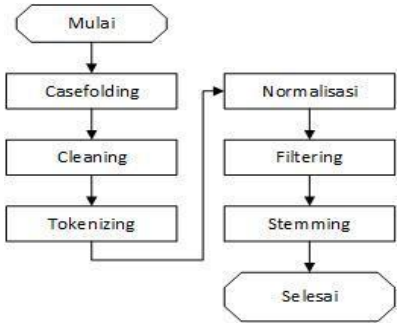
Pada proses ini peneliti akan melakukan pengumpulan data dan labeling manual nilai sentimennya. Contoh pada tabel 1 dimana yang positif diberikan angka 1 dan negative angka -1.

Tabel 1. Contoh data

Data	Tweet	label
1	Vaksin CoronaVac produksi *Sinovac* halal dan aman untuk digunakan oleh lansia.	1
2	mau vaksin *sinovac* takut sama efek samping nya	-1

Preprocessing

Preprocessing merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya, aturan-aturan yang ada di data tekstual semi terstruktur atau tidak terstruktur[6]. *Preprocessing* merupakan salah satu langkah penting dalam Analisis sentiment. Proses ini bertujuan untuk menata dan mengolah dengan baik sebuah teks yang belum sempurna[7], sehingga hasil dari proses ini adalah dokumen yang baik. Berikut tahapan preprocessing gambar 2:



Gambar 2. Tahapan *preprocessing*

Case Folding
Merupakan tahapan untuk proses mengubah seluruh huruf menjadi huruf kecil.
Cleaning
Merupakan proses untuk membersihkan data dengan penghilangan tanda baca serta karakter simbol (!@#$%^&*() +{}:"?><./;'[]=-). Atribut atau simbol yang tidak berpengaruh akan dihapus dan diganti dengan ruang kosong atau karakter spasi.

Tokenizing
Merupakan proses memisahkan setiap kata dari kalimat.
Normalisasi
tahapan *normalisasi* adalah pengembalian kata-kata yang tidak baku menjadi baku. menjadi kata baku atau sesuai dengan Kamus Besar Bahasa Indonesia (KBBI).
Filtering
Tahapan *filtering* adalah adalah tahapan menghilangkan tanggapan/tweet yang duplikat menjadi satu, dan membuang kata yang tidak perlu dengan menggunakan *stopword*.
Stemming
Tahapan *stemming* adalah tahapan pengubahan kata menjadi kata dasarnya. menghilangkan semua imbuhan (affixes) baik yang terdiri dari awalan (prefiks), sisipan (infiks), akhiran (suffixes) dan suffixes (kombinasi dari awalan dan akhiran) pada kata turunan.

Labeling Data
Setelah proses preprocessing data, proses selanjutnya yaitu memberikan label kepada semua data komentar[8]. Dengan cara nilai angka minus sebagai label negatif dan nilai angka positif sebagai label positif.

Pembobotan TF-IDF
Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus.[9].

Data		*preprocessing*	label
1	Vaksin CoronaVac produksi *Sinovac* halal dan aman untuk digunakan oleh lansia.	'produksi', 'halal', 'aman', 'lansia'	Positif
2	Polri sudah memberi contoh bahwa vaksin *sinovac* aman buat kita.	'polri', 'contoh', 'aman'	Positif
3	Kabar baik untuk kita semua 10 juta bahan baku vaksin *sinovac* tahap 5 sudah tiba di bandara Soekarno Hatta	'kabar', 'juta', 'bahan', 'baku', 'tahap'	Positif
4	Gue pribadi tidak sudi dan najis di vaksin *sinovac*... Haram dan najisss	'pribadi', 'sudi', 'najis', 'haram', 'najis	Negatif
5	mau vaksin *sinovac* takut sama efek samping nya	'takut', 'samping'	Negatif
6	Ribka mengisyaratkan masih meragukan keamanan dari vaksin *sinovac* tersebut.	'isyarat', 'ragu', 'aman'	Negatif

Gambar 3. Contoh data atau dokumen yang sudah di preprocessing

langkah pertama yang dilakukan menghitung TF dengan cara menghitung berapa banyak term (kata) dibagi total jumlah kata dalam teks. Kemudian menghitung DF dengan cara

menjumlahkan dokumen yang memiliki nilai. Langkah selanjutnya adalah dengan menghitung IDF sesuai dengan rumus berikut :

$$idf(t) = \log log \frac{n+1}{1+df} + 1$$

$idf(t)$ = Invers dokumen frekuensi
n = Total jumlah pada dokumen
df = Frekuensi dokumen dari term

No	kosakata	positif			negatif			DF	IDF = LOG(D/DF)
		D1	D2	D3	D4	D5	D6		
1	aman	1	1	0	0	0	1	3	1,5596
2	bahan	0	0	1	0	0	0	1	2,2528
3	baku	0	0	1	0	0	0	1	2,2528
4	contoh	0	1	0	0	0	0	1	2,2528
5	halal	1	0	0	0	0	0	1	2,2528
6	haram	0	0	0	1	0	0	1	2,2528
7	isyarat	0	0	0	0	0	1	1	2,2528
8	juta	0	0	1	0	0	0	1	2,2528
9	kabar	0	0	1	0	0	0	1	2,2528
10	lansia	1	0	0	0	0	0	1	2,2528
11	najis	0	0	0	2	0	0	1	2,2528
12	polri	0	1	0	0	0	0	1	2,2528
13	pribadi	0	0	0	1	0	0	1	2,2528
14	produksi	1	0	0	0	0	0	1	2,2528
15	ragu	0	0	0	0	0	1	1	2,2528
16	samping	0	0	0	0	1	0	1	2,2528
17	sudi	0	0	0	1	0	0	1	2,2528
18	tahap	0	0	1	0	0	0	1	2,2528
19	takut	0	0	0	0	1	0	1	2,2528

Gambar 4. Contoh Perhitungan IDF

Untuk mencari nilai TF-IDF menggunakan rumus :

$$Wdt = tfdt \times IDft$$

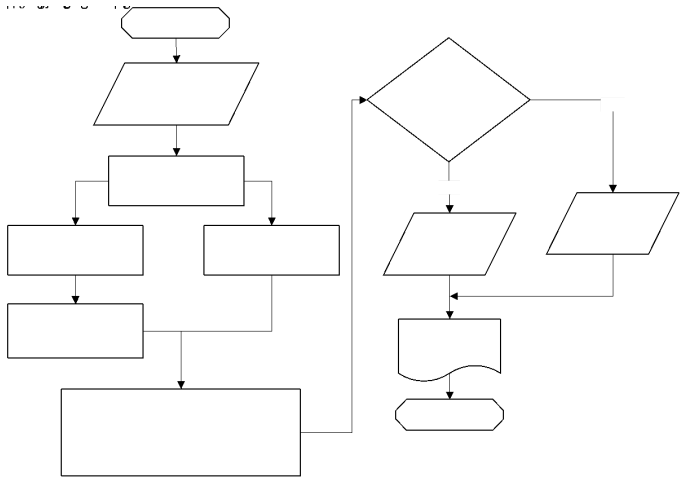
W = Bobot Dokumen
Tf = Term frekuensi
Idf = Invers dokumen frekuensi
d = Dokumen ke -d

tf-idf sckitlearnd							
No	kosakata	positif			negatif		
		TF-IDF 1	TF-IDF 2	TF-IDF 3	TF-IDF 5	TF-IDF6	TF-IDF7
1	aman	0,371156	0,439681	0	0	0	0,439681
2	bahan	0	0	0,447214	0	0	0
3	baku	0	0	0,447214	0	0	0
4	contoh	0	0,635091	0	0	0	0
5	halal	0,53611	0	0	0	0	0
6	haram	0	0	0	0,377964	0	0
7	isyarat	0	0	0	0	0	0,635091
8	juta	0	0	0,447214	0	0	0
9	kabar	0	0	0,447214	0	0	0
10	lansia	0,53611	0	0	0	0	0
11	najis	0	0	0	0,755929	0	0
12	polri	0	0,635091	0	0	0	0
13	pribadi	0	0	0	0,377964	0	0
14	produksi	0,53611	0	0	0	0	0
15	ragu	0	0	0	0	0	0,635091
16	samping	0	0	0	0	0,707107	0
17	sudi	0	0	0	0,377964	0	0
18	tahap	0	0	0,447214	0	0	0
19	takut	0	0	0	0	0,707107	0

Gambar 5. Contoh Perhitungan IDF

Naïve Bayes Classifier

Berikut alur dan tahapan *Naïve Bayes Classifier* pada gambar 6 flowchart dibawah ini :



Gambar 6. Flowchart *Naïve bayes Classifier*

Setelah mencari hasil nilai bobot TF-IDF data akan dibagi menjadi 2 yaitu data latih dan data uji, kemudian mencari nilai probabilitas kategori dan probabilitas masing-masing dokumen dari data latih.

Tahap data latih

Mencari nilai probabilitas kategori dan probabilitas masing-masing kata dari data latih. Menghitung nilai probabilitas kelas dan probabilitas masing-masing kata seperti pada Gambar 7

No	kosakata	Nk		Probailitas kata nv	
		positif	negatif	positif	negatif
1	aman	0,8108	0,4397	0,0584	0,0496
2	bahan	0,4472	0,0000	0,0467	0,0345
3	baku	0,4472	0,0000	0,0467	0,0345
4	contoh	0,6351	0,0000	0,0527	0,0345
5	halal	0,5361	0,0000	0,0496	0,0345
6	haram	0,0000	0,3780	0,0323	0,0475
7	isyarat	0,0000	0,6351	0,0323	0,0564
8	juta	0,4472	0,0000	0,0467	0,0345
9	kabar	0,4472	0,0000	0,0467	0,0345
10	lansia	0,5361	0,0000	0,0496	0,0345
11	najis	0,0000	0,7559	0,0323	0,0605
12	polri	0,6351	0,0000	0,0527	0,0345
13	pribadi	0,0000	0,3780	0,0323	0,0475
14	produksi	0,5361	0,0000	0,0496	0,0345
15	ragu	0,0000	0,6351	0,0323	0,0564
16	samping	0,0000	0,7071	0,0323	0,0589
17	sudi	0,0000	0,3780	0,0323	0,0475
18	tahap	0,4472	0,0000	0,0467	0,0345
19	takut	0,0000	0,7071	0,0323	0,0589

Gambar 7. Perhitungan probabilitas kata Naive Bayes

Tahap Data Uji

Setelah melakukan proses data latih selanjutnya yaitu menguji data yang akan di uji, untuk mengetahui data yang diuji ini termasuk dalam kategori probabilitas pada sentimen negatif dan sentimen positif. Contoh menggunakan 2 data uji seperti pada tabel 1.

Tabel 1. Contoh data uji yang sudah dilakukan *preprocessing* dan labeling

data	Data awal	*Preprocessing*	Sentimen
Data uji 1	vaksin *sinovac* sudah teruji bpom bahan aman dan halal	'uji', 'bahan', 'aman', 'halal'	positif
Data uji 2	Tidak perlu takut dan ragu lagi untuk vaksin covid-19, karena vaksin *sinovac* aman dan halal	'perlu', 'takut', 'ragu', 'aman', 'halal'	Positif

Selanjutnya menghitung nilai probabilitas pada data uji dengan mencari kata yang sama pada data latih.

Tabel 2. Contoh nilai Probabilitasnya data 1

Data 10	positif	Negatif
uji		
bahan	0,0467	0,0345
aman	0,0584	0,0496
halal	0,0496	0,0345

Pada tabel 1 kata "uji" nilai probabilitasnya kosong karena tidak ada terdapat kata yang sama pada data latih. Dan seterusnya pada uji 2.

Berikut ini hasil prediksi sentimen pada data uji dengan menggunakan *naïve bayes* terdapat pada tabel 3.

Tabel 3. Hasil **Prediksi Menggunakan *Naïve Bayes***

No	Data	Sentimen awal	Sentimen Prediksi
1	vaksin *sinovac* sudah teruji bpom bahan aman dan halal	Positif	Positif
2	Tidak perlu takut dan ragu lagi untuk vaksin covid-19, karena vaksin *sinovac* aman dan halal	Positif	negatif

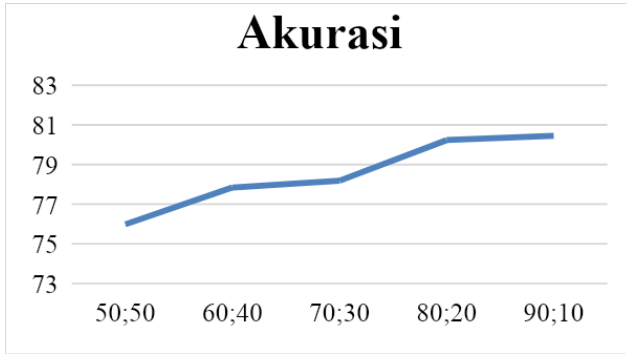
HASIL DAN PEMBAHASAN

Data yang diambil dari aplikasi orange berdasarkan tanggapan masyarakat di *twitter* tentang vaksin *sinovac* yang diambil dari 17 juli sampai dengan 23 juli 2021 dengan total jumlah sebanyak 3569 tweet. setelah dilakukan proses *preprocessing* data menjadi 2198 tweet yang berisi 1095 sentimen positif dan 1103 sentimen negatif.

Pada penelitian menghasilkan jumlah 5 kata yang sering muncul di sentiment negatif dan positif. hasil dari jumlah 5 kata yang sering muncul di sentimen negatif dan positif dapat dilihat pada gambar 10 di bawah ini gambar 8.

Gambar 8. Frekuensi kata negatif(kanan) dan kata positif(kiri)

Pada gambar 4 pada bagian negatif memperlihatkan kata yang sering muncul yaitu pegal dengan kemunculan sekitar 159 kali, diikuti dengan kata demam, gejala, pusing, sakit. sementara di sentiment positif kata yang sering muncul yaitu sehat dengan kemunculan sekitar kurang lebih 180, diikuti dengan kata alhamdulillah, terima, aman, pegal. Dari hasil pengujian performa akurasi pada rasio perbandingan data latih dan data uji dibuat grafik yang bisa dilihat pada gambar 9 berikut :



Gambar 5. Grafik akurasi dari 5 rasio data latih dan data uji

Gambar 9 menunjukan grafik berupa akurasi yang mengalami kenaikan bersamaan dengan naik nya perbandingan data latih. grafik akurasi mengalami kenaikan yang bermula dari perbandingan 50:50 dengan akurasi 75,98% sampai menjadi 80,45% pada perbandingan 90:10. rata – rata yang diperoleh dari kelima perbandingan tersebut adalah 78,54% dengan nilai akurasi tertinggi 80,45% pada perbandingan 90:10.

KESIMPULAN

1. Dari data yang diambil sebanyak 3569, setelah dilakukan proses preprocessing data menjadi 2198 tweet yang berisi 1095 sentimen positif dan 1103 sentimen negatif. di sentiment positif kata yang sering muncul yaitu sehat dengan kemunculan sekitar kurang lebih 180, diikuti dengan kata alhamdulillah, terima, aman, pegal. Pada sentimen negatif menunjukan tanggapan negatif masyarakat tentang vaksin *sinovac* setelah melakukan vaksinasi yaitu pegal, demam, pusing. pada sentimen positif menunjukan tanggapan positif masyarakat tentang vaksin *sinovac* setelah melakukan vaksinasi yaitu sehat, alhamdulilah, aman.
2. Implementasi *Naïve Bayes Classifier* pada Analisis sentiment di *twitter* tentang vaksin *COVID-19 sinovac* di Indonesia pada perbandingan 90:10 (90% Data Latih dan 10% Data uji) dengan jumlah data latih 1978 dan data uji 220 menghasilkan accuracy sebesar 80,45%, presisi sebesar 76,57%, dan recall sebesar 83,33%

DAFTAR PUSTAKA

[1]Ratriani, V. (2021). *SEHAT KONTAN*. Dipetik maret 27, 2021, dari https://kesehatan.kontan.co.id/news/kenali-6-cara-kerja-vaksin-sinovac-lawan-virus-corona-1?page=all

[2] Adilah, T., Alkhalifi, Y., Mayangky, N. A., & Gata, W. (2020). Analisa Sentimen Opini Publik Mengenai Larangan Mudik Pada Twitter Menggunakan Naive Bayes. *Jurnal CoreIT, Vol.6,No.2, Desember 2020, 6*, 2020.

[3] Arni , U. D. (2021). *GARUDA CYBER INDONESIA*. Dipetik Maret 27, 2021, dari https://garudacyber.co.id/artikel/1254-apa-itu-text-mining

[4] Hidayat, J. S., Priatna, W., & Wiyanto. (2019). IMPLEMENTASI TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN VECTOR SPACE MODEL (VSM) UNTUK PENCARIAN BERITA BAHASA INDONESIA. *Pelita Teknologi: Jurnal Ilmiah Informatika, Arsitektur dan Lingkungan 14 (2) 2019 119-133*, 119-132.

[5]Adilah, T., Alkhalifi, Y., Mayangky, N. A., & Gata, W. (2020). Analisa Sentimen Opini Publik Mengenai Larangan Mudik Pada Twitter Menggunakan Naive Bayes. *Jurnal CoreIT, Vol.6,No.2, Desember 2020, 6*, 2020.

[6]Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *JURNAL INOVTEK POLBENG - SERI INFORMATIKA, VOL. 3, NO. 1 , JUNI 2018*, 50-59.

[7]Suhartono, D. (2018). Dipetik November 27, 2021, dari School of Computer Science Binus University: https://socs.binus.ac.id/2018/08/09/menggunakan-nltk-untuk-pemrosesan-teks/

[8]BD Meilani, RK Hapsari, IF Novian (2021). Classification of community opinion on the use of the Transjakarta bus based on twitter social network using naïve bayes method, IOP Conference Series: Materials Science and Engineering

[9]Grafi Maulana, Budanis Dwi Meilani (2021), Analisis Sentimen Komentar Masyarakat Terhadap Tempat Digital Printing Menggunakan Metode Naïve Bayes, Prosiding Seminar Nasional Sains dan Teknologi Terapan