USULAN TUGAS AKHIR

ANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN PENERAPAN PPKM DI MEDIA SOSIAL TWITTER DENGAN MENGGUNAKAN METODE *XGBOOST*

Oleh:

I Putu Angga Purnama Widiarta
F1D018024

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK
UNIVERSITAS MATARAM
April 2022

HALAMAN PENGESAHAN

Bagian ini ditimpa dengan lembar pengesahan yang dihasilkan dari system https://ta.if.unram.ac.id/

ABSTRAK

Dokumen ini merupakan format panduan bagi penulis untuk menulis skripsi yang siap disahkan oleh pembimbing maupun Program Studi.. Para penulis harus mengikuti petunjuk yang diberikan dalam *template* ini. Anda dapat menggunakan dokumen ini baik sebagai petunjuk penulisan dan sebagai *template* di mana Anda dapat mengetik teks Anda sendiri. Tuliskan abstrak dalam bahasa Indonesia dengan jumlah kata 200-250 kata, yang memuat **permasalahan**, **tujuan**, **metode**, **hasil** dan **kesimpulan** tugas akhir.

Kata kunci -- Letaskkan kata kunci Anda di sini, kata kunci dipisahkan dengan koma. Istilah dengan bahasa Indonesia sebanyak 5 (lima) kata kunci

DAFTAR ISI

HALAMAN PENGESAHAN ii
ABSTRAK iii
DAFTAR ISI iv
DAFTAR GAMBAR v
DAFTAR TABEL vi
DAFTAR KODE SUMBER vii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 4
1.3 Batasan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 5
1.6 Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA 6
2.1 Penelitian Terkait 6
2.2 Teori Penunjang 11
2.2.1 Text Mining 11
2.2.2 Sentimen Analisis 11
2.2.3 *Twitter API* .....**Error! Bookmark not defined.**
2.2.4 *Web Crawling* 12
2.2.5 Preprocessing 12
2.2.6 Term Frequency – Relevance Frequency (TF-RF) 14
2.2.7 *XGBoost* 15
2.2.8 *Confusion Matrix* 16
BAB III METODOLOGI PENELITIAN 19
3.1 Alat dan Bahan 19
3.1.1 Alat Penelitian 19
3.1.2 Bahan Penelitian 19
3.2 Studi Literatur 20
3.3 Alur Penelitian 20
3.4 Kebutuhan Sistem 22
3.4.1 Analisis Pengguna 22
3.4.2 Analisis Perangkat Keras 22
3.4.3 Analisis perangkat lunak 23
3.5 Perancangan Sistem 23
3.5.1 *Web Crawling Twitter* 24
3.5.2 *Input Dataset Tweet Training* dan *Testing* 25
3.5.3 *Text Preprocessing Tweets Dataset* 27
3.5.4 *Feature Selection* 33
3.5.5 Klasifikasi Dengan XGBoost 37
3.6 Pengujian 38
3.7 Jadwal Penelitian 41
DAFTAR PUSTAKA 42

DAFTAR GAMBAR

Gambar 2.1 Proses *text mining* 11
Gambar 3.1 Alur penelitian..... 21
Gambar 3.2 Perancangan sistem 24
Gambar 3.3 Ilustrasi *cross validation* 5 *fold* 39

DAFTAR TABEL

Tabel 2.1 Penelitian sebelumnya 8
Tabel 2.2 Tabel *confusion matrix* 17
Tabel 3.1 Kebutuhan perangkat keras 22
Tabel 3.2 Kebutuhan perangkat lunak 23
Tabel 3.3 *Tweet training* 26
Tabel 3.4 *Tweet casefolding* 27
Tabel 3.5 *Tweet tokenization* 29
Tabel 3.6 *Tweet stopword removal* 31
Tabel 3.7 *Tweet stemming* 32
Tabel 3.8 Nilai TF 34
Tabel 3.9 Nilai RF 35
Tabel 3.10 Nilai TF-RF Kategori Tweet Positif 36
Tabel 3.11 Nilai TF-RF Kategori Tweet Negatif 36
Tabel 3.12 *hyperparameter XGBoost* 37
Tabel 3.13 *Confusion matrix* yang digunakan pada penelitian 40
Tabel 3.14 Jadwal penelitian 41

DAFTAR KODE SUMBER

Kode Sumber 3.1 Contoh Penulisan Kode Sumber menggunakan add-ons Easy Code Formatter pada Ms. Word.....**Error! Bookmark not defined.**
Kode Sumber 3.2 Contoh Penulisan Kode Sumber via http://www.planetb.ca/syntax-highlight-word**Error! Bookmark not defined.**

BAB I
PENDAHULUAN

1.1 Latar Belakang

Corona Virus Disease (Covid-19) merupakan bagian dari keluarga besar virus yang dapat menyebabkan penyakit baik pada hewan maupun manusia. Ditemukan pada akhir tahun 2019 [1]. *Corona* telah dikaitkan dengan infeksi saluran pernafasan pada manusia, mulai dari flu biasa hingga penyakit yang lebih serius seperti *Severe Acute Respiratory Syndrome* (SARS) dan *Middle Easy Respiratory Syndrome* (MERS), menurut *World Health Organizatio*n (WHO). Wabah ini bermula di Wuhan, Provinsi Hubei, China. Seperti diketahui, masyarakat Tionghoa sering mengomsumsi makanan "aneh" seperti kelelawar, babi, anjing, tikus dan hewan lainnya [2].

Pada *quartal* awal tahun ini, berdasarkan pada data yang diperoleh dari halaman *website* worldometers, kasus harian *covid-19* dalam skala global mengalami penurunan yang sangat signifikan, tercatat pada tanggal 1 januari 2022 jumlah kasus harian yang tercatat sebanyak 1.858.097 kemudian per tanggal 15 mei 2022 jumlah kasus harian yang tercatat menunjukkan angka sebanyak 657.158, menurunnya kasus harian ini sangat dipengaruhi oleh faktor dari pemerataan vaksinasi yang sudah dilakukan, kemudian untuk skala di Indonesia, kasus harian yang tercatat pada tanggal 15 januari 2022 menunjukkan jumlah sebanyak 1054 kasus, tidak seperti kasus harian yang terjadi di lingkup global pada umumnya, kasus harian di Indonesia malah mengalami kenaikan dari bulan Januari sampai bulan Februari yang dimana puncaknya terjadi pada tanggal 17 Februari dimana kasus harian yang tercatat berjumlah 63.956, lalu kurva tersebut tiba – tiba mengalami penurunan dari bulan tersebut hinggal per tanggal 27 April yang menunjukkan jumlah sebanyak 617. Indonesia memang menjadi salah satu negara yang ikut terjangkit virus corona. Oleh karena itu, pemberlakuan pembatasan kegiatan masyarakat (PPKM) diterapkan oleh pemerintah sebagai suatu kebijakan baru yang dimulai pada tanggal 11 Januari 2021 demi menekan angka persebaran dari penyakit *covid-19* yang disebabkan oleh virus corona dengan cara membatasi

pergerakan beserta aktivitas masyarakat. Kebijakan ini terdiri dari beberapa tingkatan, dimana tingkatannya ditentukan oleh seberapa banyak kasus yang telah terjadi di suatu daerah dimana kebijakan ini diterapkan, dimulai dari level 1 (kasus rendah), level 2 (kasus sedang), level 3 (kasus tinggi), level 4 (kasus sangat tinggi).

Kebijakan pemerintah pusat dalam melaksanakan PPKM berdampak signifikan terhadap berbagai sektor kehidupan masyarakat. Kurangnya kerjasama antar pihak, terutama antara pemerintah pusat dan pemerintah daerah menyebabkan pengendalian virus corona menjadi terombang-ambing akibat dari ketidakselarasan koordinasi [3]. Pengaruhnya terhadap sektor ekonomi adalah yang paling terlihat. Pembatasan kemampuan untuk melakukan kegiatan skala besar pasti akan mengakibatkan perekonomian menjadi semakin sulit, dengan beberapa kegiatan ekonomi berhenti. Secara alami, itu memiliki dampak signifikan pada struktur kekuasaan masyarakat. Orang akan memprioritaskan makanan dan kebutuhan penting lainnya, sementara menunda-nunda permintaan sekunder dan tersier, yang mengakibatkan pengurangan substansial dalam tabungan [4].

Pro dan kontra bermunculan di kalangan masyarakat, hal ini dapat dilihat di berbagai lini khususnya media sosial. Media sosial adalah jenis media yang menghubungkan pengguna dan memungkinkan mereka untuk berkomunikasi satu sama lain. Salah satu platform yang sering digunakan selama periode PPKM adalah twitter. Di Twitter, opini publik memiliki sifat yang tidak dibatasi dan bebas [5]. Artinya, opini yang dibuat bisa bersifat baik, negatif, atau netral. Opini di ranah politik memiliki pengaruh besar terhadap seberapa baik kinerja pemerintah [6]. Dalam PPKM, opini publik dinyatakan sebagai reaksi yang positif, negatif, atau netral terhadap pemerintah. Namun, agar opini dapat digunakan sebagai informasi yang bermakna, diperlukan prosedur analisis sentimen yang dapat menangani semua opini publik untuk memperoleh inferensi tekstual dari isi benak seluruh masyarakat Indonesia.

Analisis sentimen merupakan salah satu bidang studi dari bidang studi dengan lingkup lebih besar yang disebut dengan pemrosesan bahasa alami (*natural language processing*) atau biasa disingkat dengan nama NLP. NLP merupakan serangkaian teknik komputasi yang termotivasi secara teoritis untuk menganalisis

dan mewakili teks yang terjadi secara alami pada satu atau lebih tingkat analisis linguistik untuk tujuan mencapai pemrosesan bahasa mirip manusia untuk berbagai tugas atau aplikasi [7]. Sementara analisis sentimen adalah metode untuk memahami, menganalisis, dan memproses *input* tekstual secara otomatis untuk memperoleh informasi sentimen dari suatu opini [8]. Analisis sentimen dilakukan dengan cara mengekstrak kemudian mengolah suatu teks atau kalimat dari sumber tertentu seperti berita dan media sosial untuk memperoleh sentimen yang terkandung pada teks atau kalimat, sentimen tersebut terdari dari 3 jenis opini, yaitu opini positif, opini negatif, dan opini netral, sehingga dengan dilakukannya sentimen analisis, perusahaan atau instansi memperoleh manfaat yaitu dapat mengetahui respon masyarakat terhadap suatu pelayanan, kebijakan atau produk, melalui *feedback* yang diberikan oleh masyarakat maupun para ahli [9]. Pada sentimen analisis, *input* yang digunakan meliputi suatu kalimat atau teks yang ingin digali emosi atau sentimen yang tersirat didalamnya, sementara *output* yang dihasilkan adalah sentimen atau emosi yang ada pada teks atau kalimat yang digunakan sebagai *input*. Metode yang menonjol untuk memproses sentimen adalah metode yang menggunakan pendekatan *machine learning*. Penelitian mengenai sentimen analisis dengan menggunakan pendekatan *machine learning* terkait dengan kebijakan PPKM sebelumnya dengan metode *Support Vector Machine* (SVM) sudah dilakukan oleh Putra, dkk. Dimana pada penelitian tersebut, nilai akurasi yang diperoleh sebesar 64% [10]. Kemudian pada penelitian yang dilakukan oleh Krisdiyanto, T dkk. Proses analisis opini diklasifikasian menjadi 2 sentimen yaitu positf atau negatif, proses klasifikasi menggunakan metode *Naïve Bayes Clasifiers*, diperoleh akurasi sebesar 99% yang termasuk kedalam polaritas positif dan 1% pada polaritas negatif [11]. Pada penelitian ini, penulis akan mengimplementasikan penggunaan dari metode *XGBoost* sebagai algoritma klasifikasi, dan mengimplementasikan metode *TF-RF* (*Term Frequency – Relevance Frequency*) sebagai metode untuk menentukan bobot dari suatu *term* pada teks. *XGBoost* menghemat waktu, mengoptimalkan sumber daya memori, dan dapat diterapkan secara paralel selama proses implementasi untuk mengelola sentimen.

1.2 Rumusan Masalah

Berdasarkan dari latar belakang yang telah diuraikan, dapat dirumuskan permasalahan penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan metode *eXtreme Gradient Boosting* (*XGBoost*) dalam melakukan analisis sentimen masyarakat terhadap penerapan kebijakan PPKM di media sosial *Twitter?*
2. Bagaimana performa pengujian pada analisis sentimen masyarakat terhadap penerapakan kebijakan PPKM di media sosial *Twitter*?
3. Bagaimana tanggapan mayoritas masyarakat Indonesia terhadap penerapan kebijakan PPKM di media sosial *Twitter*?

1.3 Batasan Masalah

Batasan masalah dalam melakukan proses pada penelitian ini yaitu:

1. Dataset komentar hanya menggunakan komentar berbahasa Indonesia.
2. Dataset yang dikumpulkan dari *twitter* hanya dalam bentuk teks.
3. *Tweet* yang digunakan sebagai data diambil dari *platform Twitter* dengan menggunakan *hashtag* "#ppkm".

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini yaitu:

1. Mengimplementasikan metode *eXtreme Gradient Boosting* (*XGBoost*) dalam melakukan analisis sentimen masyarakat terhadap penerapan kebijakan PPKM di media sosial *Twitter*.
2. Mengetahui performa pengujian pada analisis sentimen masyarakat terhadap penerapakan kebijakan PPKM di media sosial *Twitter*.
3. Mengetahui tanggapan mayoritas masyarakat Indonesia terhadap penerapan kebijkan PPKM di media sosial *Twitter*.

1.5 Manfaat Penelitian

Manfaat yang diberikan dari penelitian ini yaitu:

1. Menganalisa dan mengklasifikasikan sentimen masyarakat Indonesia di *Twitter* terhadap penerapan kebijakan PPKM yang dibuat oleh pemerintah ke dalam kategori positif dan negatif.
2. Menjadi acuan ataupun referensi Pemerintah Republik Indonesia untuk mengukur seberapa efektif dan efisien kebijakan yang telah mereka terapkan berdasarkan dari data yang diperoleh melalui *Twitter*.
3. Menjadi referensi mahasiswa lain untuk memahami analisis sentimen dan metode *eXtreme Gradient Boost* (*XGBoost*).

1.6 Sistematika Penulisan

Penyusunan tugas akhir ini berdasar dari sistematika penulisan berikut yaitu:

1. Bab I. Pendahuluan

   Bab ini membahas tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, dan sistematika penulisan.
2. Bab II. Tinjauan Pustaka dan Dasar Teori

   Bab ini memuat tentang tinjauan Pustaka yang menjabarkan hasil penelitian yang berkaitan dengan penelitian ini dan dasar teori yang menjabarkan teori-teori penunjang yang berhubungan dengan penelitian ini.
3. Bab III. Metode Perancangan

   Memuat tentang metode perancangan, mulai dari pelaksanaan penelitian, diagram alir penelitian, menentukan alat dan bahan, lokasi penelitian, dan Langkah-langkah penelitian.
4. Bab IV. Hasil dan Pembahasan

   Memuat tentang hasil dan pembahasan yang diperoleh berdasarkan hasil pengukuran dan pelaksanaan.
5. Bab V. Penutup

   Memuat tentang kesimpulan dan saran berdasarkan hasil pembahasan yang telah diperoleh.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Bryan Pratama, dkk (2019) melakukan studi analisis dengan judul "Sentiment Analysis Of The Indonesian Police Mobile Brigade Corps Based On Twitter Posts Using The SVM And NB Methods" pada studi tersebut dilakukan analisa pada t*weet – tweet* dengan kata kunci "Brimob" dimana total *tweet* yang digunakan sebanyak 1000 *tweets*. Studi ini menggunakan *text mining* dengan didukung oleh *support vector machine* (SVM) untuk mengklasifikasikan sentimen publik terhadap brimob di *twitter*. Akurasi yang diperoleh dengan SVM mencapai 86,96% sedangkan dengan *Naive Bayes* diperoleh akurasi sebesar 86,48% [12].

Tahun 2019, Eka dkk. melakukan studi analisis sentimen pada contoh Gojek dan Grab, menggunakan algoritma *Naive Bayes Classifier*, dan menemukan bahwa akurasi, *recall*, dan presisi metode *Naive Bayes Classifier* masing-masing adalah 72,33%, 73,95%, dan 73,24%. Penelitian tersebut kemudian dilanjutkan oleh (D. A. Al-Qudah et al., 2020) melakukan penelitian analitik sentimen terhadap penyedia layanan *e-payment* menggunakan algoritma yang disebut *XGBoost* dan membandingkan hasilnya dengan J84, *Nave Bayes*, dan KNN. Akurasi maksimum didapatkan oleh KNN dan *XGBoost* yang masing-masing memiliki nilai *recall* 85,2 persen dan 82,8 persen. Sedangkan dengan menggunakan nilai presisi *Naive Bayes* didapatkan akurasi tertinggi sebesar 72 persen.

Dana A. Al- Qudah, dkk (2020) dengan penelitian mereka berjudul "Sentiment Analysis for e-Payment Service Providers Using Evolutionary eXtreme Gradient Boosting" melakukan analisa pada pendapat pelanggan dari servis pembayaran elektronik melalui media sosial Arab. Dataset diperoleh dari *twitter* dan *facebook*, kemudian teknik ekstraksi fitur yang digunakan yaitu TF-IDF, dan akurasi yang diperoleh darri penggunaan metode *XGBoost* disini adalah 66,8%, lebih tinggi apabila dibandingkan dari tiga metode lainnya yang coba digunakan juga oleh penulis yaitu K-NN, J48, dan NB [13].

Terkait dengan sentimen analisis, beberapa penelitian telah dilakukan sebelumnya. Fajar Fathur Rachman (2020), dalam penelitiannya yang berjudul "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada media sosial *Twitter*" melakukan penelitian sentimen analisis dengan menggunakan algoritma *Latent Dirichlet Allocation* (LDA) untuk mengelompokkan opini masyarakat dengan tujuan mengetahui topik pembicaraan yang sering dibahas masyarakat terkait dengan wacana vaksinasi, hasil analisis menunjukkan bahwa masyarakat lebih banyak memberikan respon positif terhadap wacana tersebut (30%) dibandingkan dengan respon negatifnya (26%).

Angelina Puput Giovani, dkk. (2020) dalam penelitian dengan judul "Analisis Sentimen Aplikasi Ruang Guru di Twitter Menggunakan Algoritma Klasifikasi" melakukan komparasi beberapa algoritma yaitu *Naive Bayes*, *Support Vector Machine*, dan K-*Nearest Neighbour* yang menggunakan *feature selection* dengan yang tidak menggunakan *feature selection*, serta juga membandingkan nilai Area *Under Curve* dari metode – metode tersebut untuk mengetahui algoritma mana yang paling optimal, hasil pengujian menunjukkan bahwa algoritma SVM dengan *feature selection* menjadi algoritma terbaik dengan nilai akurasi 78,55% dan AUC 0,853.

Sulaiman Ainin, dkk. (2020) dengan penelitian berjudul "Sentiment Analyses Of Multilingual Tweets On Halal Tourism" menuliskan tentang penelitian yang mereka lakukan pada *tweet* – *tweet* dari rentang waktu 2008 - 2018 yang berkaitan dengan multilingual halal *tourism* dimana konten dan sentimen dari *tweet* – *tweet* tersebut dianalisa, mereka menggunakan 19 kata kunci untuk mengesktrak data dari *tweet* dimana 5 kata kunci tersebut adalah bahasa Malaysia, dan sisanya bahasa Inggris. Setelah dilakukan analisa diperoleh kesimpulan bahwa *tweet* terkait pariwisata halal pada negara non muslim melebihi jumlah *tweet* pada negara muslim, penelitian ini menunjukkan bahwa pariwisata halal mulai populer di negara seperti Inggris, Kanada, dan Spanyol.

Elena, Podasca (2021) dengan peneltian berjudul "Predicting The Movement Direction Of OMXS30 Stock Index Using *XGBoost* and Sentiment Analysis" melakukan prediksi pada indeks pasar saham Swedia menggunakan

metode *XGBoost* yang disertakan dengan sentimen analisis dari berita keuangan guna membantu meningkatkan kinerja klasifikasi ketika memprediksi tren harga harian dari indeks pasar saham Swedia yaitu OMXS30. Hasil pada penelitian ini menunjukkan bahwa *XGBoost* memiliki kinerja yang baik dalam mengklasifikasikan tren hari OMXS30 dimana akurasi yang diperoleh mencapai 73%.

Aldiansyah Putra, dkk. (2021) dalam penelitiannya berjudul "Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Algoritma SVM" melakukan penelitian terhadap respons masyarakat di *Twitter* berupa pro dan kontra mereka kepada kebijakan pemerintah dalam pemberlakuan pembatasan kegiatan masyarakat (PPKM), metode yang digunakan pada penelitian tersebut adalah *Support Vector Machine*, dengan memanfaatkan 3000 dataset yang kemudian diperoleh hasil akurasi sebesar 64%. Dari penelitian tersebut, algoritma SVM dapat mengenali *tweet* yang berisikan penolakan PPKM sebagai *tweet* bertendensi negatif dan juga kata – kata yang memiliki hubungan terhadap tendensi negatif tersebut.

Tabel 2.1 Penelitian sebelumnya

No.	Peneliti	Judul	Keterangan
1.	Bryan Pratama et al	Sentiment Analysis Of The Indonesian Police Mobile Brigade Corps Based On Twitter Posts Using The SVM And NB Methods	Menggunakan 1000 *tweets* dengan kata kunci "brimob", akurasi yang diperoleh untuk masing – masing metode klasifikasi yaitu SVM senilai 86.96% dan *naïve baiyes* senilai 86,48%
2.	Eka et al	Analisis Sentimen Pada Contoh Gojek dan Grab	Melakukan studi analisis sentimen terhadap gojek dan grab menggunakan algoritma *naïve baiyes.* Akurasi, *recall*, dan presisi

			yang diperoleh yaitu 72,33%, 73,95%, dan 73,24%.
3.	D. A. Al-Qudah et al.	Sentiment Analysis for e-Payment Service Providers Using Evolutionary eXtreme Gradient Boosting	Analisa pada pendapat pelanggan dari servis pembayaran elektronik melalui media sosial arab, dataset diperoleh dari *facebook* dan *twitter*, menggunakan ekstraksi fitur TF-IDF dan memperoleh akurasi senilai 66,8% dengan menggunakan algoritma *XGBoost*.
4.	Fajar Fathur Rachman	Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial *Twitter*	Analisis dilakukan dengan menggunakan algoritma Latent Dirichlet Allocation (LDA) dan dataset berupa *tweets*, hasil menunjukkan respon masyarakat terhadap wacana tersebut (30%) positf dibandingkan dengan respon negatifnya senilai (26%).
5.	Angelina Puput Giovani et al	Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi	Penelitian ini menekankan komparasi antara metode Naive Bayes, Support Vector Machine, dan K-

			Nearest Neighbour dengan dan tanpa *feature selection*, SVM dengan *feature selection* menghasilkan akurasi terbaik senilai 78,55%
6.	Sulaiman Ainin et al.	Sentiment Analyses Of Multilingual Tweets On Halal Tourism	Dataset merupakan *tweet* terkait dengan penelitian yang diambil dari rentang tahun 2008-2018, kesimpulan yang diperoleh bahwa tweet terkait pariwisata halal pada negara non muslim melebihi jumlah tweet pada negara muslim.
7.	Elena, Podasca	Predicting The Movement Direction Of OMXS30 Stock Index Using XGBoost and Sentiment Analysis	Melakukan prediksi indeks harga pasar saham Swedia, disertai dengan sentimen analisis berita keuangan, dengan menggunakan algoritma XGBoost, akurasi yang diperoleh mencapai 73%.
8.	Aldiansyah Putra et al.	Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Algoritma SVM	Penelitian dilakukan dengan menggunakan algoritma SVM, dan *dataset* sebanyak 3000 *tweet* dimana akurasi yang diperoleh senilai 64%.

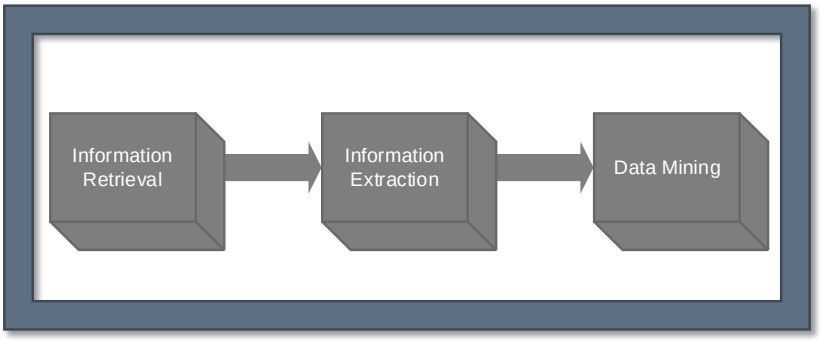
2.2 Teori Penunjang

Teori penunjang berisikan tentang konsep – konsep yang digunakan pada pembuatan dan perancangan sistem akan dibahas pada sub bab berikut :

2.2.1 Text Mining

Text mining merupakan proses penambangan teks yang menggunakan *computer* untuk mengestrak informasi secara otomatis dari berbagai sumber tertulis untuk menemukan informasi baru yang sebelumnya belum pernah ditemukan. Elemen kuncinya adalah dengan menghubungkan informasi yang telah dikumpulkan untuk menciptakan fakta baru atau hipotesis baru yang dapat diuji lebih lanjut dengan menggunakan algoritma komputasional [14].

Text mining merupakan bagian dari data *mining*, perbedaan mendasar dari *text mining* dan data *mining* adalah penambangan teks mengesktrak pola dari teks bahasa alami daripada dari *database* terstruktur yang berisi informasi *factual*. Teks ditulis untuk dibaca orang, sementara *database* dirancang agar program dapat diproses secara otomatis [14]. Untuk mengembangkan model yang belajar dari data pelatihan dan dapat mengantisipasi hasil pada informasi baru berdasarkan pengalaman dalam model pelatihan proses, penambangan teks menggabungkan teknik statistik, linguistik, dan pembelajaran mesin. Berikut adalah langkah – langkah yang terjadi pada *text mining*.



Gambar 2.1 Proses *text mining*

2.2.2 Sentimen Analisis

Sentimen analisis memiliki banyak sebutan, beberapa diantaranya merujuk pada nama – nama seperti subjektif analisis, penggalian opini, dan ekstraksi penilaian dengan beberapa koneksi ke komputasi afektif [15]. Sentimen analisis adalah studi tentang opini dan sentimen serta evaluasi sikap, penilaian, dan perasaan

yang dimiliki orang tentang hal-hal seperti produk, organisasi, isu, tema, dan fitur entitas.

Pada dasarnya sentimen analisis digunakan untuk menentukan opini yang ada pada teks dari suatu kalimat, apakah opini tersebut bersifat positif, negatif, atau netral [16]. Opini berada di pusat hampir semua aktivitas manusia karena mereka memiliki kekuatan untuk mengubah cara orang berperilaku. Berlawanan dengan pengetahuan faktual, opini dan sentimen sama-sama memiliki kualitas atau sifat yang unik karena keduanya subjektif. Karena sudut pandang satu orang hanya mewakili sudut pandang pribadi orang itu, yang seringkali tidak cukup untuk dijadikan dasar pengambilan keputusan, maka penting untuk mempertimbangkan pendapat banyak orang daripada hanya satu itu.

2.2.3 *Web Crawling*

Istilah *web crawling* atau *web scraping* sering digunakan untuk merujuk pada metode atau teknologi untuk mengumpulkan data yang dapat diakses *public* dari internet untuk fungsi tertentu. Meskipun informasi yang dikumpulkan dari internet seringkali beragam, namun jika dikompilasi dalam satu paket menggunakan metode ini, akan sangat membantu. Analisis sentimen adalah salah satu pengaplikasian dari *web crawling* yang mengidentifikasi perasaan orang tentang topik tertentu [17].

2.2.4 Preprocessing

Teks *preprocessing* digunakan dalam penelitian ini untuk mempersiapkan data untuk analisis sentimen. Data yang diproses akan dikumpulkan dari teks – teks yang memberikan informasi tentang sentimen penulis, apakah itu positif atau negatif. Analisis sentimen terlebih dahulu harus dilakukan secara manual untuk menentukan apakah sebuah sentimen baik atau negatif dengan menganalisis maksud dari garis – garis dalam sentimen tersebut untuk mempermudah pengelolaan data [18]. Teks adalah data tidak terstruktur yang mungkin tidak tersedia dalam bentuk paling mentahnya untuk digunakan oleh program *computer* secara langsung. Selain itu, data teks tidak dapat dikenai operasi numerik. Akibatnya, teks harus diproses terlebih dahulu untuk menghasilkan data yang dapat

digunakan dengan komputer. Terdapat beberapa langkah dasar yang dilakukan pada *text preprocessing*, berikut adalah:

2.2.4.1 Cleaning

Cleaning dilakukan untuk menghilangkan karakter, simbol, dan tanda baca yang tidak diperlukan dalam melakukan analisis sentimen, proses ini dilakukan karena data awal yang diperoleh merupakan data mentah yang memiliki banyak *noise* [19]. Proses ini nantinya dapat digabungkan pada saat proses *tokenization* dilakukan.

2.2.4.2 Casefolding

Untuk mempermudah sistem dalam mengenali setiap kata kemudian dalam proses pelatihan, *casefolding* mengubah semua karakter huruf besar dalam teks menjadi huruf kecil. Contoh kasus pada langkah *casefolding* yaitu ada pada proses untuk menghilangkan delimiter, delimiter dapat dianggap sebagai karakter selain huruf, dimana delimiter merupakan urutan satu karakter atau lebih yang dipakai untuk membatasi atau memisahkan data yang disajikan dalam *plain text* [20].

2.2.4.3 Tokenization

Tokenization adalah proses membagi aliran teks menjadi token, yang dapat berupa kata, frasa, simbol, atau komponen bermakna lainnya, kata – kata pada kalimat yang dipisahkan oleh spasi akan diubah ke dalam bentuk *array* atau susunan kata[21]. Pada *tokenization*, setiap kata dapat ditentukan seberapa sering kata tersebut muncul, penentuan kemunculan frekuensi dari kata – kata tersebut dapat dilakukan dengan menggunakan penghitung frekuensi kemunculan kata [22].

2.2.4.4 Stopword Removal

Stopwords, juga dikenal sebagai *noise words*, *stopwords* adalah kata-kata yang mengandung sedikit informasi yang biasanya tidak diperlukan [23]. Agar algoritma dapat fokus menemukan setiap kalimat, konsep, dan kata apa pun yang tidak terkait dengan nilai emosional, maka kata tersebut akan dihilangkan dari proses analisis sentimen. Untuk membuat proses pelatihan lebih efektif di kemudian hari, penghapusan *stopwords* melibatkan penghapusan konjungsi dan kata lain dari kalimat yang tidak memiliki arti yang sama dengan frasa.

2.2.4.5 *Stemming*

Stemming adalah metode memproleh kata dasar dengan menghilangkan imbuhan seperti awalan, akhiran, dan awalan serta akhiran kalimat. *Stemming* merupakan salah satu fungsi krusial pada sistem dengan basis *Natural Language Processing* (NLP), tujuan utama dari fitur ini yakni untuk meningkatkan *recall* dari suatu algoritma yang digunakan dengan memproses akhiran kata secara otomatis dengan memecah kata menjadi akar kata. Peningkatan nilai *recall* dicapai tanpa mengorbankan akurasi pengambilan dokumen. Sebelum istilah indeks benar-benar ditetapkan ke indeks, *stemming* biasanya dilakukan dengan menghilangkan semua sufiks dan awalan (imbuhan) yang melekat [24].

2.2.5 Term Frequency – Relevance Frequency (TF-RF)

Term weighting merupakan metode yang digunakan untuk melakukan proses penghitungan bobot pada setiap *term* yang dicari pada setiap dokumen sehingga ketersediaan dan kemiripan dari suatu *term* di dalam dokumen dapat diketahui [25]. Pada penelitian ini, metode yang akan diterapkan yaitu metode TF-RF (*Term Frequency – Relevance Frequency*), metode ini diciptakan sebagai usaha dalam memerbaiki beberapa metode yang sudah ada.

Term Frequency (TF) adalah faktor yang menentukan bobot istilah dalam sebuah teks yang tergantung pada seberapa sering teks (*term*) tersebut muncul. Saat mengekspresikan suatu kata(*term*) maka frekuensi dari *term* tersebut akan dinilai. Bobot *term* pada dokumen atau nilai kesesuaian akan meningkat seiring dengan banyaknya kemunculan *term* tersebut pada dokumen. Persamaan dari metode ini adalah [21].

$$TF(d,t) = f(d,t) \tag{2.1}$$

Yaitu $f(d,t)$ merupakan frekuensi kemunculan *term t* pada dokumen d.

Pada *Relevance Frequency* (RF) yang merupakan metode yang diusulkan oleh Man Lan, frekuensi terhadap kemunculan *term* di kategori yang berkaitan dilihat sebagai pertimbangan relevansi dokumen [26]. Jadi pada TF-RF, bobot dari suatu *term* dihitung dengan menggunakan persamaan [26].

$$tf_{td}rf = tf_{td} * \log\left(2 + \frac{b}{\max(1, c)}\right) \quad (2.2)$$

Keterangan:

$tf_{td}rf$ = Pembobotan dokumen ke dalam model ruang vector

tf_{td} = Jumlah kemunculan kata t dalam dokumen

b = Jumlah dokumen yang mengandung kata t

c = Jumlah dokumen yang tidak mengandung kata t

2.2.6 *XGBoost*

eXtreme gradient boosting, disebut sebagai *XGBoost*, adalah algoritma berbasis *tree* yang termasuk ke dalam golongan algoritma *tree* yang sama dengan *decision tree* dan *random forest* [27]. Dengan bantuan prinsip *ensemble*, algoritma *supervised tree XGBoost* mengubah sejumlah set pembelajar yang lemah (pohon) menjadi model yang kuat sehingga dapat membuat prediksi yang akurat [28]. Dikarenakan fakta bahwa *XGBoost* dapat bekerja 10 kali lebih cepat dibanding dengan implementasi dari *gradient boosting* lainnya, banyak akademisi ataupun peneliti yang menerapkan algoritma ini untuk melakukan klasifikasi dan regresi dalam berbagai situasi, termasuk prediksi penjualan, prediksi perilaku pelanggan, prediksi iklan, dan prediksi teks web [29].

Metode menambahkan model baru ke pendekatan *ensemble* disebut *boosting*, hal ini dilakukan untuk mengoreksi kesalahan dari model sebelumnya. Model akan ditambahkan satu per satu sampai sampai tidak ada lagi peningkatan yang mungkin dilakukan. Teknik *ensemble* menggunakan model pohon klasifikasi dan regresi yang disebut *tree ensemble models*. Strategi yang dikenal dengan teknik *ensemble* menggabungkan prediksi dari berbagai *tree* menjadi satu [29]. Ini berusaha untuk secara berurutan memodelkan setiap *predictor* menggunakan kesalahan residual dari model sebelumnya. Ketika *dataset* dimasukkan, langkah pertama adalah menggunakan dataset yang dipilih untuk membangun model awal. Persamaan 2.3 dan 2.5 kemudian digunakan untuk menentukan nilai prediksi awal dan kesalahan residual dari model asli. Model pertama dibuat menggunakan persamaan nomor 2.3, sedangkan model berikutnya dibuat menggunakan persamaan nomor 5.

$$h_0(x) = mean(Y) \tag{2.3}$$

$$\hat{Y} = Y - h_0(x) \tag{2.4}$$

Dimana Y merepresentasikan nilai residual *error* model awal dan $h_0(x)$ merepresentasikan nilai prediksi awal dari model pertama. Model kedua kemudian akan dibuat menggunakan residual *error* dari model pertama untuk menentukan nilai prediksinya. Kesalahan residual dari model pertama dan kedua kemudian akan digunakan untuk membuat model ketiga untuk menentukan nilai prediksinya. Sebanyak n_enstimator ditetapkan, maka proses ini akan terus berulang [29].

XGBoost menghasilkan satu set *decision tree* yang mana setiap model pohon bergantung pada pohon sebelumnya. Nilai prediksi awal untuk model pertama di *XGBoost* akan lemah, tetapi karena lebih banyak model dibangun, bobot diperbarui untuk menghasilkan prediksi yang lebih kuat. Untuk meminimalkan fungsi tujuan, nilai proyeksi dari masing – masing model akan dijumlahkan kemudia dimasukkan ke dalam persamaan nomor 2.5 [29].

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.5}$$

Dimana n adalah jumlah model yang akan digunakan, I adalah fungsi untuk mengukur selisih antara target y_i dan $\hat{y}_i$ yang diprediksi. $f_t(x_i)$ adalah model baru yang dibangun. Sedangkan Ω adalah fungsi untuk membuat model terhindar dari *overfitting* . Persamaan nomor 2.5 digunakan untuk mencari nilai keseluruhan.

2.2.7 *Confusion Matrix*

Confusion Matrix adalah metode untuk membandingkan nilai nyata dan yang diantisipasi untuk mengevaluasi keefektifan model pembelajaran mesin dalam prediksi label. Sebuah tabel yang disebut *confusion matrix* memiliki empat set terpisah dari kombinasi nilai yang diharapkan dan yang sebenarnya. Empat istilah — *true positive*, *false negative*, *true negative*, dan *false negative* digunakan dalam *confusion matrix* untuk menunjukkan hasil operasi kategorisasi. Selanjutnya penulis akan merancang metode *XGBoost* dengan memanfaatkan skor keempat item tersebut sebagai input analitis untuk menentukan nilai akurasi, presisi, *recall*,

dan f1*score*. Menghitung akurasi, *recall*, dan presisi merupakan salah satu metode yang digunakan untuk menilai klasifikasi. Dalam metode ini, *confusion matrix* berfungsi sebagai panduan perhitungan.

Tabel 2.2 Tabel *confusion matrix*

Kelas		Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Pada *matrix* diatas dapat dinyatakan sebagai berikut :

a. *True Positive* (TP), adalah jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1.
b. *False Positive* (FP), adalah jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1.
c. *False Negative* (FN), adalah jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0.
d. *True Negative* (TN), adalah jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0.

Penghitungan akurasi dilakukan dengan menghitung jumlah prediksi benar yang kemudian dibagi dengan jumlah prediksi, berikut pada 2.6 adalah persamaan dari penghitungan akurasi :

$$accuracy = \frac{TP\ +\ TN}{TP + TN + FP + FN} * 100\% \quad (2.6)$$

$$precision = \frac{TP}{TP + FP} * 100\% \quad (2.7)$$

$$recall = \frac{TP}{TP + FN} * 100\% \quad (2.8)$$

Pada persamaan 2.7 dinyatakan sebagai *precision* yang dimana *precision* adalah tingkat keakuratan antara data yang diminta dengan hasil prediksi yang

diberikan oleh model. Sedangkan pada persamaan 2.8 ditunjukkan *recall* yang merupakan tingkat keberhasilan model dalam menemukan Kembali sebuah informasi.

BAB III
METODOLOGI PENELITIAN

3.1 Alat dan Bahan

Berisi alat dan bahan yang akan digunakan untuk penelitian.

3.1.1 Alat Penelitian

Dalam penelitian tentang analisis sentimen masyarakat Indonesia di media sosial *twitter* terhadap kebijakan pemerintah dalam penerapan PPKM di Indonesia, digunakan beberapa alat yang terdiri dari perangkat keras dan perangkat lunak. Alat – alat tersebut adalah sebagai berikut.

a. Perangkat Keras

Perangkat keras yang digunakan dalam penelitian adalah satu unit komputer dengan spesifikasi berikut :

1. *Processor* Intel® Core™ i5-7400 3,50 GHz
2. Memori RAM DDR4 32GB
3. Kartu grafis nvidia RTX 2070 8GB VRAM

b. Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Sistem operasi *Windows* 10 *Pro*
2. *Jupyter Notebook*
3. *Visual Studio Code*
4. Bahasa Pemrograman *Python* versi 3.9
5. *Microsoft Office*

3.1.2 Bahan Penelitian

Bahan penelitian yang digunakan dalam penelitian tentang analisis sentimen masyarakat Indonesia pada media sosial *twitter* terhadap kebijakan pemerintah dalam melaksanakan penerapan PPKM di Indonesia menggunakan *XGBoost* ini adalah *tweet* masyarakat Indonesia yang berisikan opini terhadap kebijakan pemerintah dalam penerapan PPKM. Pada penelitian ini, *tweet* yang diperoleh merupakan *tweet* dari rentang waktu bulan April 2020 hingga April

2022 dengan total 20.000 *tweet*, dari total *tweet* - *tweet* tersebut, mengacu pada penelitian sebelumnya yang menggunakan 3000 *tweet* dari total 5000 *tweet* yang diperoleh [10], maka pada penelitian ini penulis mencoba dengan menggunakan *tweet* sejumlah 10.000 *tweet* yang nantinya akan diberi label positif dan negatif secara manual oleh 2 orang mahasiswa dan 1 alumni Universitas Mataram yang dengan sukarela membantu proses pelabelan *tweet* ini untuk menghindari bias apabila pelabelan dilakukan oleh penulis itu sendiri, orang – orang tersebut yaitu Muhammad Khaidar Rahman, Umbara Diki Pratama, dan I Nengah Suardika. *Tweet* – *tweet* tersebut nantinya digunakan sebagai *training* dan *testing* dataset.

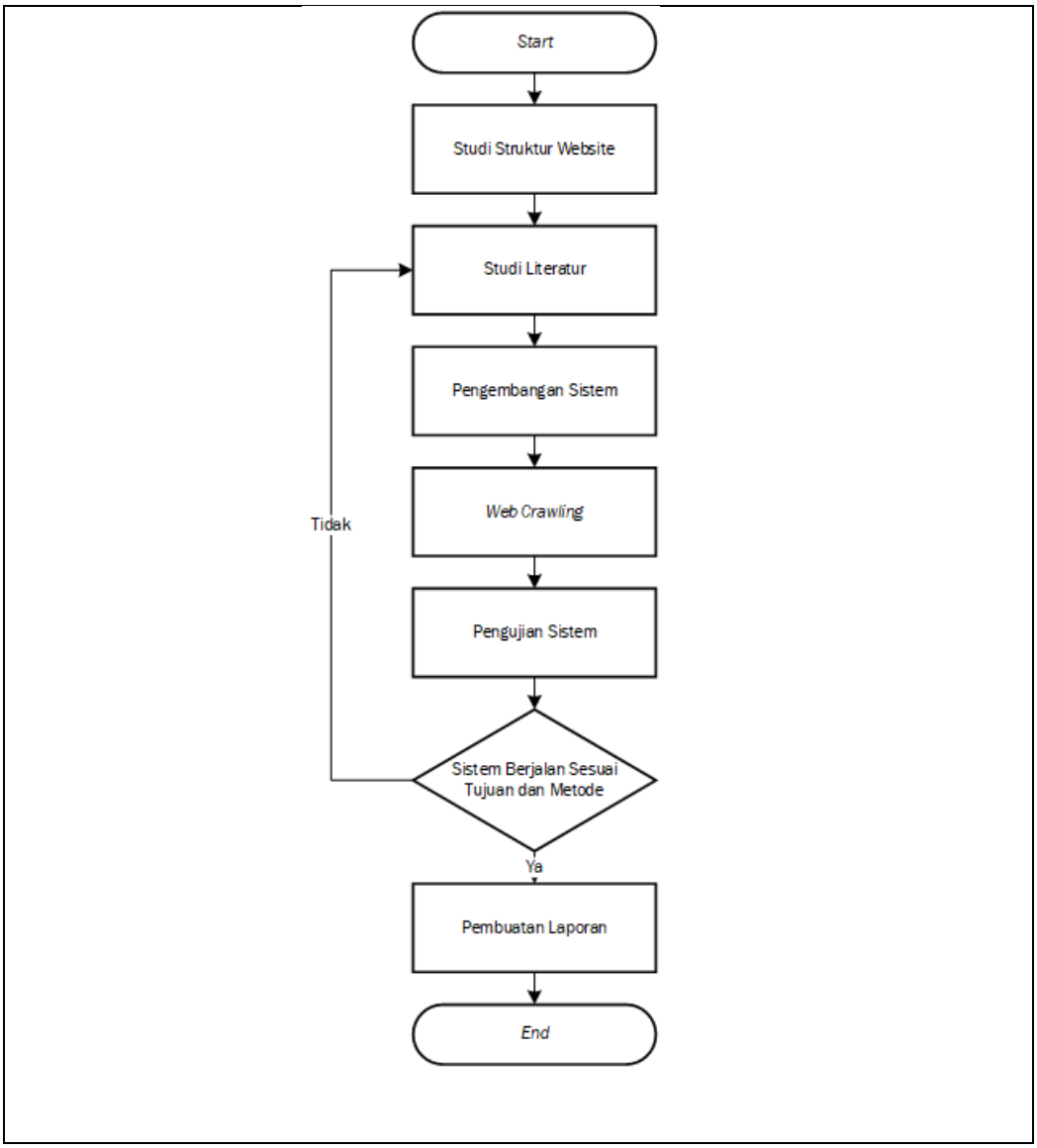
3.2 Studi Literatur

Studi literatur yang dilakukan dengan tujuan untuk mendukung penelitian yaitu mempelajari buku elektronik, jurnal – jurnal penelitian, serta berbagai sumber lainnya yang berkaitan dengan topik penelitian, yaitu analisis sentimen dan *web crawling*. Lebih spesifik, materi yang dipelajari adalah *text preprocessing*, *natural language processing*, analisis sentimen serta pemanfaatan metode *XGBoost* dalam melakukan analisis sentimen. Jurnal – jurnal yang dipelajari membahas berbagai studi kasus tentang analisis sentimen dengan metode *XGBoost*.

3.3 Alur Penelitian

Analisis sentimen masyarakat Indonesia pada media sosial *twitter* terhadap kebijakan pemerintah dalam melaksanakan penerapan PPKM di Indonesia menggunakan *XGBoost* dilakukan dalam beberapa tahapan. Tahap pertama yang dilakukan yakni melakukan studi terhadap struktur *website* yang akan dilakukan *crawling*, *website* tersebut adalah laman *Twitter.com*. Kemudian dilakukan studi literatur untuk mendapatkan pengetahuan serta gambaran akan penelitian yang dilakukan. Literatur yang dipelajari berupa jurnal penelitian serta buku yang membahas tentang *web crawling* dan analisis sentimen menggunakan metode *XGBoost*. Kemudian dilakukan pengembangan sistem yang dilandaskan pada literatur yang telah dipelajari sebelumnya. Setelah

sistem telah selesai dikembangkan, dilakukan *crawling dataset* dari *website www.twitter.com.* Setelah itu dilakukan pengujian pada sistem. Apabila sistem berjalan sesuai dengan tujuan dan metode yang digunakan, penelitian dilanjutkan dengan pembuatan laporan. Apabila sistem tidak berjalan sesuai dengan yang diharapkan, maka akan dilakukan kembali studi literatur untuk memperbaiki kesalahan – kesalahan yang menyebabkan sistem yang dibangun kurang optimal. Diagram alir penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Alur penelitian

3.4 Kebutuhan Sistem

Dalam penelitian tentang analisis sentimen masyarakat terhadap kebijakan pemerintah dalam penerapan PPKM di Indonesia pada media sosial *Twitter* menggunakan metode *XGBoost*, analisis kebutuhan sistem dibagi menjadi 3 jenis yaitu analisis pengguna, analisis perangkat keras dan analisis perangkat lunakyang digunakan dalam penelitian.

3.4.1 Analisis Pengguna

Pengguna dari sistem ini adalah orang – orang atau peneliti yang akan melakukan penelitian terkait dengan analisis sentimen di masa yang akan datang, khususnya mereka yang mengangkat topik serupa dengan penelitian ini ataupun mereka yang menggunakan metode serupa sehingga pada penelitian ini, orang – orang tersebut dapat menggunakan penelitian ini sebagai landasan teori pada penelitian mereka selanjutnya, maupun sebagai sumber referensi pustaka. Selain itu juga, hasil dari penelitian ini dapat digunakan oleh orang – orang seperti *developer* suatu aplikasi, apabila mereka membutuhkan suatu model klasifikasi sentimen untuk membangun aplikasi yang mereka buat.

3.4.2 Analisis Perangkat Keras

Perangkat keras yang digunakan dalam pembangunan sistem, pelatihan data, serta pengujian sistem merupakan elemen penting dalam penelitian ini. Perangkat keras yang mumpuni dapat membantu mempercepat proses-proses yang dilakukan seperti pelatihan data yang membutuhkan sumber daya cukup tinggi. Perangkat keras yang digunakan dalam penelitian ini memiliki spesifikasi seperti yang terdapat pada Tabel 3.1.

Tabel 3.1 Kebutuhan perangkat keras

No.	Nama Perangkat	Spesifikasi
1.	*Processor*	Intel® Core™ i5-7400 3,50 GHz
2.	Memori	Memori RAM 32GB DDR4
3.	GPU	NVIDIA RTX 2070
4.	*Storage*	256GB SSD, 1TB HDD, 500GB M.2

3.4.3 Analisis perangkat lunak

Selain perangkat keras, perangkat lunak juga memiliki peranan penting dalam proses pengembangan sistem. Penggunaan perangkat lunak yang tepat dapat membantu mempercepat proses penelitian. Perangkat lunak yang digunakan dalam penelitian ini dapat dilihat pada Tabel 3.2.

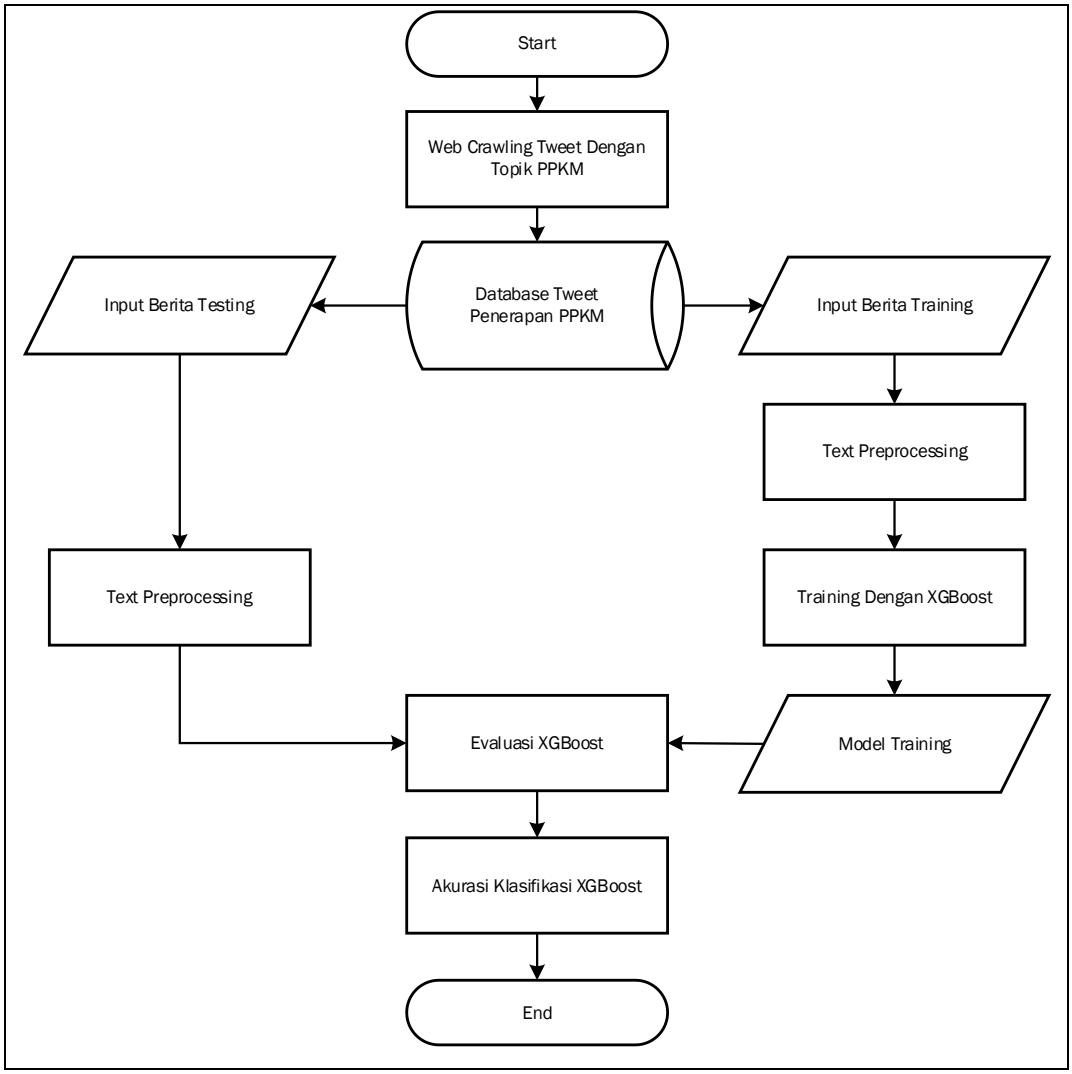
Tabel 3.2 Kebutuhan perangkat lunak

No.	Nama Perangkat	Spesifikasi
1.	Sistem Operasi	Windows 10
2.	*Text Editor*	*Jupyter Notebook*
3.	*Microsoft Office*	*Ms.Office Professional Plus 2019*
4.	Bahasa pemrograman *python*	*Python* 3.9.6
5.	*Library* NLTK	*Python* nltk 1.1.2
6.	*Library scikit learn*	*Python scikit learn* 0.23.2
7.	*Library* Sastrawi	*Python* Sastrawi 1.0.1
8.	*Library Scraping*	*Python snscrape* 0.4.3.20220106
9.	*Web Browser*	*Opera Mini*
10.	*Library Pandas*	*Python pandas* 1.4.4
11.	*Library Regex*	*Python regex* 2022.8.17

3.5 Perancangan Sistem

Rancangan dari sistem analisis sentimen komentar masyarakat Indonesia pada media sosial *twitter* terhadap kebijakan pemerintah Indonesia dalam penerapan PPKM dengan menggunakan algoritma *eXtreme Gradient Boost* (*XGBoost*) yang terdiri dari beberapa tahapan, yang dapat dilihat pada Gambar 3.2. Alur dari perancangan sistem tersebut dimulai dari tahapan *crawling tweet – tweet* yang memiliki tagar #ppkm dimana nantinya *tweet – tweet* tersebut digunakan sebagai dataset pada sistem, *tweet – tweet* yang berisi opini *netizen* Indonesia terkait dengan penerapan kebijakan PPKM yang dilakukan oleh pemerintah diambil dari *database twitter* dan *tweet – tweet* tersebut memiliki beberapa kriteria yakni *tweet* memuat tagar #ppkm, kemudian *tweet* haruslah menggunakan bahasa Indonesia, dan *tweet – tweet* tersebut adalah *tweet* yang dibuat dalam rentang waktu yang

dimulai dari tanggal 1 April 2020 hingga 1 April 2022, kemudian *tweet* dibagi menjadi 2 jenis dataset, yaitu *tweet* yang digunakan sebagai data *training* dan *tweet* yang digunakan sebagai data *testing*, selanjutnya pada *tweet* – *tweet* tersebut dilakukan *preprocessing* agar *tweet* – *tweet* nantinya menjadi lebih relevan pada saat memasuki proses *training* oleh model, setelah dilakukan *training* model, maka dilanjutkan dengan mengevaluasi model tersebut dengan *testing* data untuk memperoleh tingkat akurasi dari model yang telah dilatih.



Gambar 3.2 Perancangan sistem

3.5.1 *Web Crawling Twitter*

Pada tahap ini, *tweet* dikumpulkan melalui jejaring media sosial *Twitter*. *Tweet* yang dikumpulkan adalah *tweet* yang menggunakan *hashtag* "#ppkm". "#ppkm" kemudian dimasukkan pada *query* pencarian *tweets* yang digunakan oleh

library snscrape untuk melakukan *crawling* data pada *tweet – tweet* berbahasa Indonesia yang memuat tagar "ppkm" di dalamnya. Pada hasil pencarian yang telah dilakukan, didapati sebanyak 20000 *tweet* berbahasa Indonesia, yang membahas terkait kebijakan pemerintah Indonesia dalam penerapan PPKM. *Tweet – tweet* tersebut berisi berbagai macam jenis sentimen yang terkandung di dalamnya, dari sentimen positif, netral, dan negatif, namun pada penelitian ini kategori sentimen yang diambil hanya berupa sentimen positif dan negatifnya saja. *Tweet – tweet* yang telah dikumpulkan tersebut nantinya digunakan sebagai data latih dan data uji pada program untuk memperoleh akurasi terhadap bagaimana sentimen sentimen masyarakat Indonesia secara keseluruhan terkait dengan kebijakan pemerintah Indonesia dalam melakukan penerapan PPKM melalui pengujian yang dilakukan dengan menerapkan algoritma *XGBoost*.

3.5.2 *Input Dataset Tweet Training* dan *Testing*

Pada tahap ini, *tweet* yang telah diperoleh dari laman media sosial *twitter* yang telah dimuat sebagai dataset akan dibagi menjadi 2 kategori yaitu *tweet – tweet* yang digunakan sebagai data *training* dan data *testing. Tweet* yang digolongkan sebagai data *training* digunakan untuk membuat model klasifikasi sedangkan *tweet* yang digunakan sebagai data *testing* digunakan untuk menguji model yang telah dibuat.

a. *Input Tweet Training*

Tweets training yang sebelumnya telah diperoleh melalui proses *scraping* dari media sosial *Twitter* yang kemudian dimuat ke dalam dataset dengan ekstensi *.csv* akan dimasukan ke dalam sistem untuk diproses. *Tweet* yang diperoleh merupakan *tweet – tweet* berbahasa Indonesia, dimana *tweet* tersebut memuat opini masyarakat Indonesia tentang kebijakan penerapan PPKM yang dilakukan pemerintah, *tweet – tweet* tersebut ditandai sedemikian dikarenakan memuat tagar "#ppkm" pada penulisannya. Selanjutnya dilakukan *preprocessing* pada seluruh *tweet – tweet* yang telah dimuat sebagai data *training*, yang kemudian di-*training* menggunakan algoritma *XGBoost*.

Contoh *tweet* yang digunakan sebagai *tweet training* pada sistem dapat dilihat pada Tabel 3.3.

Tabel 3.3 *Tweet training*

Date	*Username*	*Tweets*
2021-10-25 07:34:41+00:00	nuchillinaris	"Sebab, program penanggulangan #Covid19 dirasakan oleh masyarakat bawah. Pun dg program pemulihan ekonomi sangat membumi & dirasakan benar oleh masyarakat yg perekonomiannya sangat terdampak akibat kebijakan #PPKM. #7ThJokowiLuarBiasa Jokowi diakui dunia! https://t.co/ATrYbGU7px"
2022-03-22 17:25:29+00:00	ViantAntony	Ruwet Ruwet Ruwet inilah Negeri RuwetNesia. Hebatnya Virus itu adalah dia tau Ramadhan akn datang meraka akn meperbanyak bhkn #PPKM kemungkinan di perpanjang. Yakan pak @KemenkesRI ??? https://t.co/m0wm0fkHUW
2022-03-22 15:23:09+00:00	LaNyallaMM1	"Saya berharap pelonggaran aktivitas bukan hanya untuk menggenjot perekonomian. Tetapi juga dimanfaatkan sektor pendidikan untuk meningkatkan Sumber Daya Manusia yang sedikit mundur karena pandemi.

		@JatimPemprov #LaNyalla #ketuadpdri #dpdri #daridaerahuntukindonesia #ppkm"

3.5.3 *Text Preprocessing Tweets Dataset*

Text preprocessing yang dilakukan pada penelitian ini dibagi menjadi 3 tahap, yaitu tahap *casefolding, tokenization, stemming*, dan *stop-word removal*.

a. *Casefolding*

Casefolding merupakan tahapan pertama yang dilakukan pada *preprocessing text*, pada tahap ini *dataset* yang ada akan disamaratakan penggunaan huruf kapitalnya, yang dimana pada *dataset* ini, seluruh *tweet* akan diubah hurufnya menjadi huruf kecil, ini bertujuan agar *tweets* menjadi konsisten pada penggunaan hurufnya, dan mencegah sistem mengalami kebingungan dikarenakan kata yang sama apabila penulisan hurufnya berbeda, maka kata tersebut akan dianggap sebagai kata yang berbeda oleh sistem.

Tabel 3.4 *Tweet casefolding*

Date	*Username*	*Tweets*
2021-10-25 07:34:41+00:00	nuchillinaris	"sebab, program penanggulangan #covid19 dirasakan oleh masyarakat bawah. pun dg program pemulihan ekonomi sangat membumi & dirasakan benar oleh masyarakat yg perekonomiannya sangat terdampak akibat kebijakan #ppkm.

		#7thjokowiluarbiasa jokowi diakui dunia! https://t.co/atrybgu7px"
2022-03-22 17:25:29+00:00	ViantAntony	ruwet ruwet ruwet inilah negeri ruwetnesia. hebatnya virus itu adalah dia tau ramadhan akn datang meraka akn meperbanyak bhkn #ppkm kemungkinan di perpanjang. yakan pak @kemenkesri ??? https://t.co/m0wm0fkhuw
2022-03-22 15:23:09+00:00	LaNyallaMM1	"saya berharap pelonggaran aktivitas bukan hanya untuk menggenjot perekonomian. tetapi juga dimanfaatkan sektor pendidikan untuk meningkatkan sumber daya manusia yang sedikit mundur karena pandemi. @jatimpemprov #lanyalla #ketuadpdri #dpdri #daridaerahuntukindonesia #ppkm"

b. *Tokenization*

Tokenization merupakan untuk mentransformasikan *tweets* menjadi kumpulan kata yang disebut *terms*. Pada *tokenization* juga dilakukan penghilangan tanda baca. Hal ini dilakukan karena tanda baca tidak dapat

digunakan sebagai *terms* karena terdapat pada hampir seluruh dokumen. Sebelum proses *tokenization*, terlebih dahulu dilakukan proses *case folding* atau mengubah setiap kata menjadi huruf kecil. Karakter selain huruf dihilangkan dan dianggap *delimiter*. Tujuannya adalah agar tidak terjadi kesalahan interpretasi oleh komputer ketika ada dua kata yang sama tapi dianggap berbeda karena perbedaan huruf besar dan huruf kecil. Contoh *tweets* yang telah melewati proses *tokenization* dapat dilihat pada Tabel 3.5.

Tabel 3.5 *Tweet tokenization*

Date	*Username*	*Tweet_Tokens*
2021-10-25 07:34:41+00:00	nuchillinaris	'sebab', 'program', 'penanggulangan', 'dirasakan', 'oleh', 'masyarakat', 'bawah', 'pun', 'dg', 'program', 'pemulihan', 'ekonomi', 'sangat', 'membumi', 'amp', 'dirasakan', 'benar', 'oleh', 'masyarakat', 'yg', 'perekonomiannya', 'sangat', 'terdampak', 'akibat', 'kebijakan', 'jokowi', 'diakui', 'dunia'
2022-03-22 17:25:29+00:00	ViantAntony	'ruwet', 'ruwet', 'ruwet', 'inilah', 'negeri', 'ruwetnesia', 'hebatnya', 'virus', 'itu', 'adalah', 'dia', 'tau', 'ramadhan', 'akn', 'datang', 'meraka', 'akn', 'meperbanyak', 'bhkn', 'kemungkinan', 'di', 'perpanjang', 'yakan', 'pak'
2022-03-22 15:23:09+00:00	LaNyallaMM1	'saya', 'berharap', 'pelonggaran', 'aktivitas', 'bukan', 'hanya', 'untuk', 'menggenjot', 'perekonomian', 'tetapi', 'juga', 'dimanfaatkan', 'sektor', 'pendidikan', 'untuk',

		'meningkatkan', 'sumber', 'daya', 'manusia', 'yang', 'sedikit', 'mundur', 'karena', 'pandemi'

c. *Stopword Removal*

Pada tahap ini *stopwords* pada *tweet* akan dihapus guna meningkatkan keefektifan proses *training* di kemudian hari, *stopwords* adalah kata – kata pada bidang NLP (*Natural Language Processing*) yang dinyatakan memiliki sedikit makna, bahkan hampir tidak bermakna, kata – kata tersebut seperti ‘yang’, ‘yaitu’, ‘di’, ‘tempat’, ‘terus’, ‘walau’, dan masih banyak lainnya. Pada *machine learning* maupun *deep learning*, *stopword* biasanya dihapus terlebih dahulu sebelum proses pelatihan dilakukan dikarenakan *stopword* cenderung muncul dalam jumlah banyak, dimana hal tersebut berdampak pada tidak adanya informasi unik yang diberikan oleh *stopword – stopword* untuk dapat digunakan pada proses klasifikasi atau *clustering*. Pada penelitian ini *sample stopwords* diperoleh dari yang sudah disediakan pada *library* NLTK, *stopwords – stopwords* yang ada pada *library* tersebut kemudian penulis coba gabungkan dengan beberapa *stopwords* yang penulis cenderung temukan pada *tweet – tweet* yang digunakan. Beberapa contoh dari *stopwords* tersebut tersaji di tabel ini.

```
['ada', 'adalah', 'adanya', 'adapun', 'agak', 'agaknya', 'agar',
'akan', 'akankah', 'akhir', 'akhiri', 'akhirnya', 'aku',
'akulah', 'amat', 'amatlah', 'anda', 'andalah', 'antar',
'antara', 'antaranya', 'apa', 'apaan', 'apabila', 'apakah',
'apalagi', 'apatah', 'artinya', 'asal', 'asalkan', 'atas',
'atau', 'ataukah', 'ataupun', 'awal', 'awalnya',...-n]
```

Sedangkan *tweet – tweet* yang sudah dihilangkan *stopwordnya* dapat dilihat pada tabel berikut,, dapat dibandingkan pada tahap sebelumnya, kata seperti ‘sebab’, ‘pun’, ‘benar’ dihilangkan pada *tweet* karena kata – kata tersebut tergolong ke dalam *stopwords*.

Tabel 3.6 *Tweet stopword removal*

Date	*Username*	*Tweet_WSW*
2021-10-25 07:34:41+00:00	nuchillinaris	'program', 'penanggulangan', 'dirasakan', 'masyarakat', 'dg', 'program', 'pemulihan', 'ekonomi', 'membumi', 'amp', 'dirasakan', 'masyarakat', 'yg', 'perekonomiannya', 'terdampak', 'akibat', 'kebijakan', 'jokowi', 'diakui', 'dunia'
2022-03-22 17:25:29+00:00	ViantAntony	ruwet', 'ruwet', 'ruwet', 'negeri', 'ruwetnesia', 'hebatnya', 'virus', 'tau', 'ramadhan', 'akn', 'meraka', 'akn', 'meperbanyak', 'bhkn', 'perpanjang', 'yakan'
2022-03-22 15:23:09+00:00	LaNyallaMM1	'berharap', 'pelonggaran', 'aktivitas', 'menggenjot', 'perekonomian', 'dimanfaatkan', 'sektor', 'pendidikan', 'meningkatkan', 'sumber', 'daya', 'manusia', 'mundur', 'pandemi'

d. *Stemming*

Teknik dalam memperoleh kata dasar atau dalam artian lain *stem* dari suatu kata pada suatu kalimat disebut dengan nama *stemming*. Pada proses tersebut, dilakukan pemotongan pada imbuhan (*affix*) kata, baik itu *prefix* maupun *suffix* [30]. Proses *stemming* dilakukan dengan menggunakan algoritma Nazief dan Adriani karena *tweet* yang digunakan pada penelitian merupakan *tweet* berbahasa Indonesia, selain itu juga algoritma ini memiliki tingkat presisi yang lebih baik dibandingkan algoritma lainnya seperti algoritma Porter. Algoritma *stemming* antara suatu bahasa dengan bahasa yang lain memiliki perbedaan, ini dikarenakan morfologi yang berbeda antara suatu bahasa, seperti morfologi bahasa Indonesia apabila disandingkan dengan bahasa Inggris, contoh kasusnya

adalah pada teks berbahasa Inggris, hanya diperlukan menghilangkan *sufiks* pada suatu teks untuk memperoleh *root word* dari sebuah kata. Sementara pada bahasa Indonesia, proses yang terjadi lebih kompleks, ini dikarenakan terdapat beberapa variasi imbuhan yang harus dibuang agar memperoleh *root word* dari sebuah kata. [30]. Algoritma Nazief dan Adriani melakukan *stemming* dengan menghilangkan *inflection suffixes* (“-lah”, “-kah”, “-ku”, “-mu”, atau “- nya”), *possessive pronouns* (“-ku”, “-mu”, atau “-nya”), *derivation suffixes* (“-i”, “-an” atau “-kan”) dan *derivation prefixes* (“di-”,“ke-”,“se-”,“te-”, “be-”, “me-”, atau “pe-”), kemudian mencocokan kata dengan kata yang ada di kamus. Proses *stemming* dilakukan untuk menyelaraskan suatu kata yang memiliki imbuhan berbeda agar kata tersebut dapat diartikan atau dimaknai sebagai kata yang sama. Contoh *tweets* yang telah melewati proses *stemming* dapat dilihat pada Tabel 3.7.

Tabel 3.7 *Tweet stemming*

Date	*Username*	*Tweet_Stemming*
2021-10-25 07:34:41+00:00	nuchillinaris	'program', 'tanggulang', 'rasa', 'masyarakat', 'bawah', 'dg', 'program', 'pulih', 'ekonomi', 'sangat', 'bumi', 'amp', 'rasa', 'benar', 'masyarakat', 'yg', 'ekonomi', 'sangat', 'dampak', 'akibat', 'bijak', 'jokowi', 'aku', 'dunia'
2022-03-22 17:25:29+00:00	ViantAntony	'ruwet', 'ruwet', 'ruwet', 'ini', 'negeri', 'ruwetnesia', 'hebat', 'virus', 'tau', 'ramadhan', 'akn', 'datang', 'raka', 'akn', 'meperbanyak', 'bhkn', 'mungkin', 'panjang', 'yakan', 'pak'
2022-03-22 15:23:09+00:00	LaNyallaMM1	'harap', 'longgar', 'aktivitas', 'bukan', 'genjot', 'ekonomi', 'manfaat', 'sektor', 'didik', 'tingkat', 'sumber', 'daya', 'manusia', 'sedikit', 'mundur', 'pandemi'

3.5.4 *Feature Selection*

Pada tahap ini, dilakukan seleksi fitur yang dianggap relevan dalam mewakili suatu kelas, dalam penelitian ini fitur tersebut adalah kata. Kata yang digunakan adalah *unigram* dari hasil *preprocessing*. Dalam penelitian ini metode *feature selection* yang digunakan adalah metode *term weighting TF-RF* yang dimana akan dibagi ke dalam dua tahap terlebih dahulu pada pemrosesannya:

3.5.4.1 *Term Frequency – Relevance Frequency (TF-RF)*

a. *Term Frequency* (TF)

Pada proses berikut ini, semua kata yang ada pada *tweets* akan dijadikan sebagai *feature* pada masing – masing *tweets* untuk proses *training* dan *test*. Melalui proses ini, akan terbentuk *vector* berdasarkan *term* atau kata yang ada pada seluruh teks. Berdasarkan pada jumlah kata yang muncul pada *tweets* tersebut sesuai dengan kata acuannya maka *tweet training* dan *test* dapat diberikan nilai numerik pada *vector*-nya yang sesuai dengan jumlah kemunculan kata acuan dibagi jumlah kata pada kalimat dimana kata acuan tersebut berada. Bobot *term frequency* bersumber dari hasil perhitungan nilai – nilai numerik tersebut.

Setelah tahapan *preprocessing* dilakukan pada *tweet* – *tweet* yang ada, maka selanjutnya *tweet* – *tweet* tersebut akan dikonversikan atau diubah ke dalam bentuk angka sehingga *tweet* – *tweet* tersebut memiliki bobot agar dapat diproses oleh sistem. Penerapan dari kalkulasi *term frequency* dapat dilihat pada 4 contoh dokumen *tweet* yang telah mengalami *preprocessing* sebagai berikut ini:

Negative Tweet 1 : **program tanggulang rasa masyarakat bawah dg program pulih ekonomi sangat bumi amp rasa benar masyarakat yg ekonomi sangat dampak akibat bijak jokowi aku dunia.**

Negative Tweet 2 : **ruwet ruwet ruwet ini negeri ruwetnesia hebat virus tau ramadhan akn datang raka akn meperbanyak bhkn mungkin panjang yakan pak.**

Positive Tweet 3 : **harap longgar aktivitas genjot ekonomi manfaat sektor didik tingkat sumber daya manusia maju pandemi.**

Positive Tweet 4 : **presiden doa masyarakat hidup sehat umur tingkat ekonomi.**

Masing – masing *term* atau kata akan dihitung frekuensi kemunculan nya dalam sebuah dokumen seperti yang terdapat pada Tabel 3.8, dengan persamaan yang digunakan yaitu Persamaan (2.1).

Tabel 3.8 Nilai TF

Term	**Term Frequency (TF)**			
	D1	**D2**	**D3**	**D4**
program	$\frac{2}{24}$	0	0	0
tanggulang	$\frac{1}{24}$	0	0	0
rasa	$\frac{2}{24}$	0	0	0
masyarakat	$\frac{2}{24}$	0	0	$\frac{1}{8}$
ruwet	0	$\frac{3}{20}$	0	0
virus	0	$\frac{1}{20}$	0	0
pulih	$\frac{1}{24}$	0	0	0
ekonomi	$\frac{2}{24}$	0	$\frac{1}{16}$	$\frac{1}{8}$
sehat	0	0	0	$\frac{1}{8}$
hidup	0	0	0	$\frac{1}{8}$
…	…	…	…	…

b. *Relevance Frequency* (RF)

Apabila jumlah nilai dari *term frequency* (TF) telah ditemukan pada setiap dokumen, maka prosedur selanjutnya adalah mencari nilai *relevance frequency* (RF) pada setiap kata. Mengacu pada frekuensi kemunculan *term* di kategori yang berkaitan, maka dipertimbangkan relevansi dokumen pada nilai *relevance*

frequency. Nilai RF dari suatu *term* atau kata menjadi tinggi ketika *term frequency* dari kata tersebut memiliki nilai yang tinggi pada suatu dokumen yang mencangkup kata tersebut dan pada kelas dokumen lainnya [21]. Berikut disajikan perhitungan nilai RF pada beberapa kata yang dapat dilihat di Tabel 3.9 berikut :

Tabel 3.9 Nilai RF

Term	**Relevance Frequency (RF) (Negative Tweet)** $\boldsymbol{log(2+\frac{b}{max(1,c)})}$	**Relevance Frequency (RF) (Positive Tweet)** $\boldsymbol{log(2+\frac{b}{max(1,c)})}$
program	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
tanggulang	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
rasa	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
masyarakat	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$
ruwet	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
virus	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
pulih	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$
ekonomi	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$	$log\left(2+\frac{2}{\max(1,0)}\right)=1.38629$
sehat	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$
hidup	$log\left(2+\frac{0}{\max(1,0)}\right)=0.69314$	$log\left(2+\frac{1}{\max(1,1)}\right)=1.09861$
…	…	…

c. *Term Frequency-Relevance Frequency* (TF-RF)

Pada proses ini, pemberian bobot pada kata dilakukan dengan cara mengalikan nilai *term frequency* dengan nilai *relevance frequency*, hal ini dapat dilihat pada Persamaan (2.2). Pada Tabel 3.10 dan Tabel 3.11 ditampilkan beberapa kata yang telah dihitung nilai akhirnya [21].

Tabel 3.10 Nilai TF-RF Kategori Tweet Positif

Term	**TF.RF = TF(d,t) * RF(t)**			
	D1	**D2**	**D3**	**D4**
program	0.0577616	0	0	0
tanggulang	0.0288808	0	0	0
rasa	0.0577616	0	0	0
masyarakat	0.0915508	0	0	0.1373262
ruwet	0	0.103971	0	0
virus	0	0.034657	0	0
pulih	0.0288808	0	0	0
ekonomi	0.1155241	0	0.086643	0.1732862
sehat	0	0	0	0.1373262
hidup	0	0	0	0.1373262
…	…	…	…	

Tabel 3.11 Nilai TF-RF Kategori Tweet Negatif

Term	**TF.RF = TF(d,t) * RF(t)**			
	D1	**D2**	**D3**	**D4**
program	0.0915508	0	0	0
tanggulang	0.0457754	0	0	0
rasa	0.0915508	0	0	0
masyarakat	0.0915508	0	0	0.1373262
ruwet	0	0.164791	0	0
virus	0	0.0549305	0	0
pulih	0.0457754	0	0	0

ekonomi	0.0915508	0	0.0686631	0.1373262
sehat	0	0	0	0.0866425
hidup	0	0	0	0.0866425
…	…	…	…	

3.5.5 Klasifikasi Dengan XGBoost

Setelah *term – term* dari *tweets* pada dataset diperoleh bobotnya yang telah diubah menjadi bentuk vektor melalui perhitungan pada metode pembobotan TF-RF, maka tahapan selanjutnya adalah melakukan klasifikasi XGBoost. Dataset [X Y] pertama kali akan dicari rata – rata nilai target (Y), ini dilakukan untuk memperoleh nilai prediksi awal ($h0$) dan nilai residual error ($\hat{Y}$) awal, dimana proses ini dinyatakan pada Persamaan (2.3) dan (2.4). Untuk memperoleh model pertama (M1) yang merupakan sebuah *decision tree* yang dilatih dengan variabel independen dan residual error [X$\hat{Y}$] sebagai data untuk mendapatkan prediksi dari model M1, maka dilakukanlah *training* data pada model pertama tersebut terlebih dahulu. Ketika proses *training*, model dapat dimaksimalkan dengan menggunakan beberapa *hyperparameter* yang tersedia pada model yang digunakan.

Pada algoritma *XGBoost* tingkatan seberapa berpengaruh *hyperparameter* pada kinerja model dapat bervariasi, faktor yang memengaruhi kinerja model selain dari penggunaan *hyperparameter* yaitu *dataset*, dan masalah yang ingin diselesaikan. Pada penelitian ini penulis mencoba menerapkan beberapa *hyperparameter* yang biasanya paling krusial dalam menentukan kinerja model yang dimana diperlihatkan sebagai berikut pada Tabel 3.12 di bawah ini.

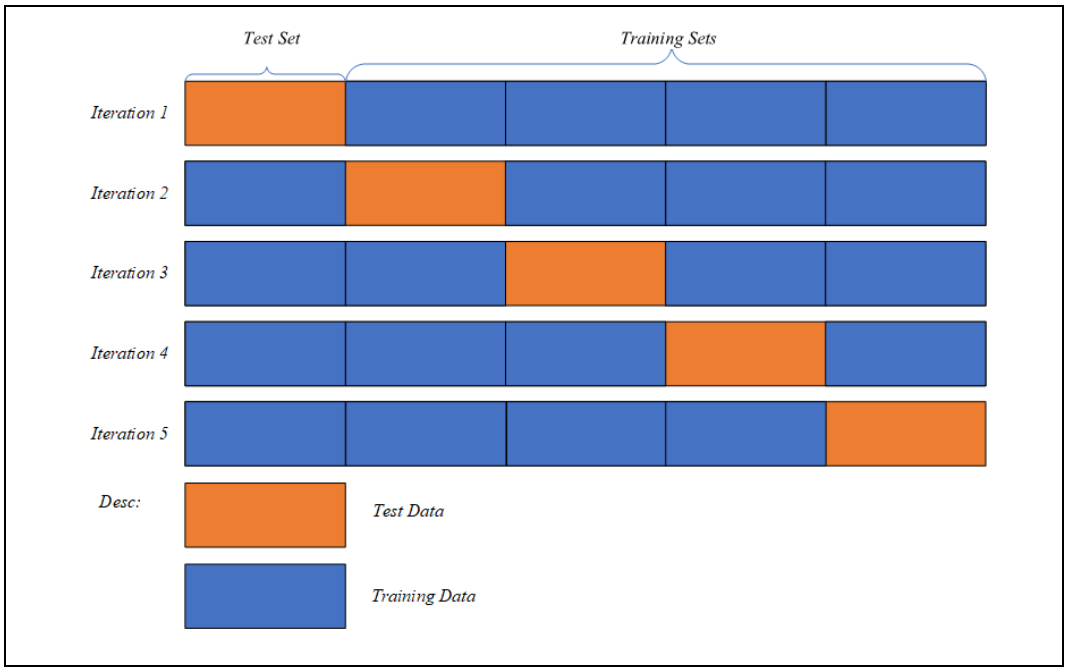
Tabel 3.12 *hyperparameter XGBoost*

Hyperparameter	**Score**
gamma	1
learning_rate	1
n_estimators	100
max_depth	1
subsample	1

Penulis memperoleh nilai *score* pada Tabel 3.11 yang mengacu pada penelitian sebelumnya yang juga menggunakan algoritma *XGBoost* [19]. '*n_estimators*' merupakan *hypeparameter* yang menentukan jumlah *decision tree* yang dibangun pada model, '*learning_rate*' merupakan *hyperparameter* yang mengontrol tingkatan cepat lambatnya model belajar, '*max_depth*' mengontrol seberapa maximum kedalaman masing – masing *decision tree* pada model, '*subsample*' menentukan persentase jumlah sampel yang diambil secara *random* dari *training data* untuk membangun setiap *decision tree* pada model, '*gamma*' menentukan ambang batas pada saat model akan berhenti membuat pemisahan *node*. *Score* pada *hyperparameter* dapat berubah dikarenakan penulis melakukan eksperimen untuk menemukan *set* variasi *hyperparameter* yang optimal dengan tujuan untuk meningkatkan kinerja model. Penulis nantinya mencoba menggunakan masing – masing 3 nilai yang bervariasi pada 5 *hyperparameter* yang telah disajikan pada Tabel 3.11, ini digunakan dengan tujuan untuk memperoleh nilai akurasi yang lebih akurat pada kinerja model.

3.6 Pengujian

Pengujian yang dilakukan dalam penelitian terkait analisis sentimen masyarakat pada media sosial *twitter* menggunakan metode *XGBoost* adalah dengan menggunakan teknik *k-fold cross validation*, dan *Dataset* yang terkait pada penelitian ini diperoleh melalui crawling *tweets* pada *twitter*. Teknik *k-fold cross validation* merupakan teknik yang digunakan untuk mengevaluasi hasil klasifikasi. Dalam skenario pengujian pada penelitian ini, *dataset* dibagi menjadi beberapa bagian yang disebut *fold*. Pada setiap iterasi yang dilakukan, salah satu *fold* digunakan sebagai *testing data*, dan sisa *fold* digunakan sebagai *training data*. Proses ini dilakukan sebanyak nilai K yang ditetapkan hingga seluruh *fold* digunakan sebagai *testing data* [21].



Gambar 3.3 Ilustrasi *cross validation* 5 *fold*

Berdasarkan pada Gambar 3.3 yang merupakan ilustrasi dari *cross validation*, maka prosesnya dapat dijabarkan sebagai berikut ini [21]:

1. Jumlah *instance* dibagi sebanyak K bagian atau disebut *fold*.
2. Pada iterasi ke-1 adalah saat bagian ke-1 dijadikan sebagai data uji dan empat bagian sisanya dijadikan sebagai data latih. Kemudian dilakukan penghitungan akurasi atau kesamaan atau kedekatan pada hasil pengukuran dengan menggunakan angka atau data yang sebenarnya berdasarkan porsi dari data tersebut. Persamaan yang digunakan pada perhitungan akurasi adalah sebagai berikut:

$$Akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum jumlah\ data\ uji} \times 100 \quad (3.1)$$

3. Pada iterasi ke-2 yaitu saat bagian ke-2 dijadikan sebagai data uji dan bagian lain sisanya dijadikan sebagai data latih. Kemudian, dilakukan penghitungan akurasi berdasarkan porsi data tersebut.
4. Lalu seterusnya penghitungan dilakukan hingga mencapai iterasi atau *fold* ke-K. Kemudian rata – rata akurasi yang diperoleh dari K buah akurasi akan dijadikan sebagai akurasi final.

Dua kelas atau kategori yang berbeda akan digunakan pada penelitian ini untuk membedakan sentimen dari *tweets* yaitu, kategori *tweets* positif dan *tweets* negatif. Terdapat beberapa parameter yang diuji dalam penelitian ini antara lain :

1. Pengaruh nilai K pada metode *K-fold cross validation* yang diterapkan terhadap nilai akurasi model yang diperoleh.
2. Pengaruh himpunan variasi *hyperparameter* pada model *eXtreme Gradient Boosting* terhadap akurasi kinerja model. *XGBoost* memiliki banyak *hyperparameter* yang sangat berperan penting dalam memengaruhi kinerja model.
3. Pengujian akurasi menggunakan metode *eXtreme Gradient Boosting* (*XGBoost*).

Pada setiap percobaan evaluasi dilakukan dengan menghitung *accuracy recall* dan *precision* dari model.

Untuk menghitung nilai-nilai tersebut diperlukan *confussion matrix* untuk menyajikan hasil klasifikasi dalam bentuk tabel. Tabel 3.13 *confussion matrix* dalam penelitian ini dapat dilihat pada.

Tabel 3.13 *Confusion matrix* yang digunakan pada penelitian

Kebenaran	Hasil Klasifikasi		Total
	Positif	Negatif	
Positif	*True Positive*	*False Negative*	Total Kelas Positif
Negatif	*False Positive*	*True Negative*	Total Kelas Negatif
	Prediksi Kelas Positif	Prediksi Kelas Negatif	

Pada Tabel 3.12 *confusion matrix*, Ketika hasil klasifikasi dari kelas memberikan hasil positif sementara pada kebenarannya juga berkategori positif, maka hal tersebut dapat digolongkan sebagai *true positive*, sementara apabila hasil klasifikasi menunjukkan kelas negatif dan kebenarannya juga menunjukkan negatif maka hal tersebut dinamakan dengan *true* negative. Namun apabila hasil klasifikasi dan kebenarannya berlawanan, seperti didapati hasil klasifikasinya berkategori positif sementara kebenarannya negatif maka hal tersebut dinamakan *false positive*, sementara untuk sebaliknya disebut dengan *false negative*.

Recall dan *precision* untuk tiap sentimen dihitung dengan menggunakan Persamaan (2.8) dan Persamaan (2.7). Sedangkan untuk *accuracy* dihitung menggunakan Persamaan (2.6). Nilai *recall* dan *precision* untuk tiap percobaan didapatkan dengan mencari nilai rata-rata dari *recall* dan *precision* per sentimen. Performa model secara keseluruhan didapatkan dengan menghitung nilai *accuracy* serta nilai rata-rata *recall* dan *precision* dari seluruh percobaan.

3.7 Jadwal Penelitian

Waktu yang digunakan dalam proses pengembangan sistem analisis sentimen masyarakat Indonesia pada media sosial *Twitter* terhadap penerapan kebijakan PPKM di Indonesia yaitu selama enam bulan. Jadwal kegiatan dapat dilihatpada Tabel 3.14.

Tabel 3.14 Jadwal penelitian

No.	Kegiatan	Waktu (Bulan)						Keterangan
		I	II	III	IV	V	VI	
1	Analisis							Analisis kebutuhan
2	Pengumpulan Data							Pengumpulan *tweet* sebagai *dataset*
3	Pembangunan Sistem							Pengkodean sistem
4	*Testing*							Pengujian sistem
5	Implementasi							Penerapan sistem
6	Dokumentasi							Dokumentasi Sistem

DAFTAR PUSTAKA

[1] O. Walsyukurniat, Z. Stkip, and N. Selatan, "GERAKAN MENCEGAH DARIPADA MENGOBATI TERHADAP PANDEMI COVID-19." [Online]. Available: https://www.sehatq.com/artikel/bahaya-virus-

[2] S. Seti Indriani, S. K. Universitas Padjadjaran Jl Raya Jatinangor -Bandung, and D. Prasanti, "Analisis konvergensi simbolik dalam media sosial youth group terkait kasus COVID-19 di Indonesia," *Jurnal Kajian Komunikasi*, vol. 8, no. 2, pp. 179–193, 2020.

[3] L. Agustino, "Analisis Kebijakan Penanganan Wabah Covid-19: Pengalaman Indonesia," *Jurnal Borneo Administrator*, vol. 16, no. 2, pp. 253–270, Aug. 2020, doi: 10.24258/jba.v16i2.685.

[4] "Seminar Nasional Penelitian LPPM UMJ Website: http://jurnal.umj.ac.id/index.php/semnaslit E-ISSN:2745-6080." [Online]. Available: http://jurnal.umj.ac.id/index.php/semnaslit

[5] N. D. Asih and M. Rosit, "Opini Publik di Media Sosial: Analisis Isi Opini Kandidat Ahok-Djarot dan Anies-Sandi di Twitter," vol. 8, no. 2, Mar. 2018.

[6] E. Tungadi, Z. Saharuna, M. Nur Yasir Utomo, T. Elektro, and P. Negeri Ujung Pandang, *Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar*. [Online]. Available: https://dev.twitter.com

[7] E. D. Liddy, "Natural Language Processing Natural Language Processing Natural Language Processing 1," 2001. [Online]. Available: https://surface.syr.edu/istpub

[8] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *Jurnal Sistem dan Teknologi Informasi (Justin)*, vol. 8, no. 2, p. 183, Apr. 2020, doi: 10.26418/justin.v8i2.36776.

[9] E. M. Sipayung, H. Maharani, and I. Zefanya, "PERANCANGAN SISTEM ANALISIS SENTIMEN KOMENTAR PELANGGAN MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER," 2016. [Online]. Available: http://ejournal.unsri.ac.id/index.php/jsi/index

[10] A. Putra, D. Haeirudin, H. Khairunnisa, and R. Latifah, "Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Algoritma Svm," 2021.

[11] T. Krisdiyanto, E. Maricha, and O. Nurharyanto, “Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers,” *Jurnal CoreIT*, vol. 7, no. 1, 2021.

[12] B. Pratama *et al.*, “Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods,” in *Journal of Physics: Conference Series*, May 2019, vol. 1201, no. 1. doi: 10.1088/1742-6596/1201/1/012038.

[13] D. A. Al-Qudah, A. M. Al-Zoubi, P. A. Castillo-Valdivieso, and H. Faris, “Sentiment analysis for e-payment service providers using evolutionary extreme gradient boosting,” *IEEE Access*, vol. 8, pp. 189930–189944, 2020, doi: 10.1109/ACCESS.2020.3032216.

[14] M. Hearst, “What Is Text Mining?,” 2003.

[15] Y. Mejova, V. Shirsat, and R. S. Jagdale, “Sentiment Analysis: An Overview Hybrid Sent iment Analysis Framework for a Morphologically Rich Language Jelena Mit rović, Miljana Mladenovic Subgroup det ect ion in ideological discussions Mona Diab Sent iment Analysis of Event s from T wit t er Using Open Source Tool Sentiment Analysis: An Overview Comprehensive Exam Paper,” 2009.

[16] D. Rustiana Program Studi Sistem Komputer Perguruan Tinggi Raharja and N. Rahayu Magister Teknologi Informatika Perguruan Tinggi Raharja, “ANALISIS SENTIMEN PASAR OTOMOTIF MOBIL: TWEET TWITTER MENGGUNAKAN NAÏVE BAYES,” *Jurnal SIMETRIS*, vol. 8, 2017.

[17] v. Smith, *Go Web Scraping Quick Start Guide: Implement the power of Go to scrape and crawl data from the web*. Packt Publishing Ltd, 2019.

[18] A. H. Tri Jaka, “Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining.”

[19] “ANALISIS SENTIMEN PADA LAYANAN GOJEK INDONESIA.”

[20] S. Informatika and A. Polinema, “IMPLEMENTASI ANALISIS SENTIMEN TWITTER MENGENAI OPINI MASYARAKAT TERHADAP RKUHP TAHUN 2019,” *SIAP)*, p. 2020.

[21] R. Dwiyansaputra, G. Satya Nugraha, F. Bimantoro, and A. Aranta, “DETEKSI SMS SPAM BERBAHASA INDONESIA MENGGUNAKAN TF-IDF DAN STOCHASTIC GRADIENT DESCENT CLASSIFIER (Indonesian SMS Spam

Detection using TF-IDF and Stochastic Gradient Descent Classifier).” [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/

[22] G. Gupta, “Text Document Tokenization for Word Frequency Count using Rapid Miner (Taking Resume as an Example).” [Online]. Available: www.ijcaonline.org

[23] P. Buttar, J. Kaur, and P. Kaur Buttar, “A Systematic Review on Stopword Removal Algorithms,” 2018, [Online]. Available: http://www.ijfrcsce.org

[24] M. Anjali and G. Jivani, “A Comparative Study of Stemming Algorithms.” [Online]. Available: www.ijcta.com

[25] A. T. Ni’mah and A. Z. Arifin, “Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis,” *Rekayasa*, vol. 13, no. 2, pp. 172–180, Aug. 2020, doi: 10.21107/rekayasa.v13i2.6412.

[26] A. N. Assidyk, E. B. Setiawan, S. Si, I. Kurniawan, S. Pd, and M. Si, “Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada Trending Topic di Twitter dengan Menggunakan Klasifikasi K-Nearest Neighbor.”

[27] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] W. F. Mustika, H. Murfi, and Y. Widyaningsih, “Analysis Accuracy of XGBoost Model for Multiclass Classification - A Case Study of Applicant Level Risk Prediction for Life Insurance,” in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 71–77. doi: 10.1109/ICSITech46713.2019.8987474.

[29] M. Riza Kurniawanda, F. Adline, and T. Tobing, “Analysis Sentiment Cyberbullying in Instagram Comments with XGBoost Method,” *International Journal of New Media Technology)*, vol. 9, no. 1, p. 28, 2022.

[30] D. Wahyudi, T. Susyanto, D. Nugroho, P. Studi Teknik Informatika, S. Sinar Nusantara Surakarta, and P. Studi Sistem Informasi, “IMPLEMENTASI DAN ANALISIS ALGORITMA STEMMING NAZIEF & ADRIANI DAN PORTER PADA DOKUMEN BERBAHASA INDONESIA”.