PENERAPAN *COMBINE SAMPLING* PADA ANALISIS SENTIMEN MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* (STUDI KASUS: SENTIMEN KEBIJAKAN LEPAS MASKER DI INDONESIA)

SKRIPSI

EVLYN PRICILIA KONDY

H051191027

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2023

PENERAPAN *COMBINE SAMPLING* PADA ANALISIS SENTIMEN MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* (STUDI KASUS: SENTIMEN KEBIJAKAN LEPAS MASKER DI INDONESIA)

SKRIPSI

Diajukan sebagai salah satu syarat memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

EVLYN PRICILIA KONDY

H051191027

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

NOVEMBER 2023

LEMBAR PERNYATAAN KEOTENTIKAN

Saya yang bertanda tangan di bawah ini menyatakan dengan sungguh-sungguh bahwa skripsi yang saya buat dengan judul:

Penerapan *Combine Sampling* Pada Analisis Sentimen Menggunakan Algoritma *K-Nearest Neighbor* (Studi Kasus: Sentimen Kebijakan Lepas Masker di Indonesia)

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun

Makassar, 27 November 2023

Evlyn Pricilia Kondy

NIM H051191027

PENERAPAN *COMBINE SAMPLING* PADA ANALISIS SENTIMEN MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* (STUDI KASUS: SENTIMEN KEBIJAKAN LEPAS MASKER DI INDONESIA)

Disetujui Oleh:

Pembimbing Utama	Pembimbing Pertama
Siswanto, S.Si., M.Si.	**Dr. Nirwan, M.Si.**
NIP. 19920107 201903 1 012	**NIP. 19630306 198702 002**

Ketua Program Studi

Dr. Anna Islamiyati, S.Si., M.Si.
NIP. 19740808 200501 2 002

Pada 27 November 2023

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama	: Evlyn Pricilia Kondy
NIM	: H051191027
Program Studi	: Statistika
Judul Skripsi	: Penerapan *Combine Sampling* Pada Analisis Sentimen Menggunakan Algoritma *K-Nearest Neighbor* (Studi Kasus: Sentimen Kebijakan Lepas Masker di Indonesia)

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

DEWAN PENGUJI

1. Ketua : Siswanto, S.Si., M.Si. (.....)
2. Sekretaris : Dr. Nirwan, M.Si. (.....)
3. Anggota : Drs. Raupong, M.Si. (.....)
4. Anggota : Anisa, S.Si., M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 27 November 2023

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya sehingga penulis bisa sampai di titik ini dan mampu menyelesaikan penyusunan tugas akhir ini dengan judul "**Penerapan *Combine Sampling* Pada Analisis Sentimen Menggunakan Algoritma K-Nearest Neighbor (Studi Kasus: Sentimen Kebijakan Lepas Masker di Indonesia)**" yang disusun sebagai salah satu syarat akademik untuk memperoleh gelar sarjana di Program Studi Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

Penulis menyadari bahwa dengan segala keterbatasan pengetahuan dan kemampuan yang dimiliki penyelesaian tugas akhir ini tidak terlepas dari dukungan dan dorongan dari berbagai pihak yang senantiasa turut memberikan bantuan baik secara moril maupun materil sehingga penulis dapat menyelesaikan tugas akhir ini. Oleh karena itu, penulis senantiasa mengucapkan terima kasih yang setulus-tulusnya serta penghargaan yang setinggi-tingginya untuk orang tua penulis, Ayah **Cenry Kondy** dan Ibu **Melina Chandra** yang telah membesarkan dan mendidik penulis, memberikan dukungan penuh, pengorbanan, limpahan cinta dan kasih sayang tanpa batas, serta dengan ikhlas telah mengiringi setiap langkah penulis dengan doa dan restu mulianya. Ucapan terima kasih juga penulis haturkan kepada adik-adik penulis **Yesly Pricilia Kondy** dan **Helena Pricilia Kondy** karena telah menjadi adik-adik yang sangat baik dan senantiasa memberikan semangat maupun doa terbaiknya untuk penulis dalam menyelesaikan tugas akhir ini, serta keluarga besar penulis, terima kasih atas dukungan dan doa mulianya selama ini.

Penghargaan yang tulus dan ucapan terima kasih dengan penuh keikhlasan dan ketulusan juga penulis ucapkan kepada:

1. **Bapak Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.**, selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya.
2. **Bapak Dr. Eng. Amiruddin**, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh jajarannya.
3. **Ibu Dr. Anna Islamiyati S.Si., M.Si.**, selaku Ketua Departemen Statistika, **segenap Dosen Pengajar dan Staf** yang telah memberikan ilmu dan

kemudahan kepada penulis dalam berbagai hal selama menempuh pendidikan sarjana di Departemen Statistika.

4. **Bapak Siswanto, S.Si., M.Si.**, selaku Pembimbing Utama yang dengan penuh kesabaran telah meluangkan waktu dan pemikirannya untuk senantiasa memberikan bimbingan, saran, dukungan, dan motivasi kepada penulis dari awal hingga selesainya penulisan tugas akhir ini.
5. **Bapak Dr. Nirwan, M.Si.,** selaku Pembimbing Pertama sekaligus Dosen Pembimbing Akademik penulis yang dengan kesabaran telah meluangkan waktu dan pemikirannya untuk senantiasi memberikan arahan, dorongan semangat, saran, dan motivasi kepada penulis dari awal hingga selesainya penulisan tugas akhir ini.
6. **Bapak Drs. Raupong, M.Si.** dan **Ibu Anisa, S.Si., M.Si.,** selaku Tim Penguji yang telah meluangkan waktu dalam memberikan motivasi serta kritikan yang membangun kepada penulis dalam penyempurnaan tugas akhir ini.
7. **William Suciangto** yang selalu membantu, memberi semangat, dan menjadi pendengar bagi penulis selama ini.
8. **Dian Ayu Permata Sari Rusdy** dan **Refa Joyce Semida** yang senantiasa memberi semangat, dukungan, bantuan selama perkuliahan hingga penulisan tugas akhir. Bagi Dian terima kasih sudah sering mengajarkan materi kepada penulis apabila ada materi yang tidak dipahami. Bagi Refa terima kasih banyak selalu cepat dan sigap memberikan informasi dan membantu urusan terkait perkuliahan dan selalu menghibur penulis.
9. Kelompok 38 dan 39 KKN 108 Unhas**, Jafir Ramadhan, Azkiah Azisah Jufri, Putri Aqidah Setiawan, Fourensius Edison Junianto, Nur Aisyah, Nurfa Nurul Utami, Audy Alifia Rudy, Aldi Musa, Alif Syahrani.** Terima kasih atas kebersamaan dan kehangatan selama KKN serta selalu membantu penulis.
10. **Agus Hermawan, Muhammad Syamsul Bahri, Muhammad Yusran, Yasmin Pratiwi, Diah Lestari, Wahyu Dwi Rahmawati, Marhama, Wa Ode Siti Amni, dan Rahma** yang telah membantu penulis selama pengerjaan tugas akhir.

11. Teman - teman **Statistika 2019**. Terima kasih atas ilmu, kebersamaan, suka dan duka dalam menjalani perkuliahan di Departemen Statistika. Terima kasih sudah menerima dan selalu membantu penulis selama perkuliahan.
12. Tentor–tentor Bimbel CIF**, Ce Tizia Thilma, Ce Alexandra Thelzya, Ce Lenny Chen, Kak Oktaviana Anu Samon, dan Kak Fitrianah.** Terima kasih atas semangat, bantuan, dukungan, saran, dan telah menjadi keluarga bagi penulis.
13. **Keluarga Besar Himastat FMIPA Unhas**, terima kasih atas ilmu, pengalaman, dan telah menjadi tempat belajar bagi penulis.
14. **Keluarga Besar UKM KPI Unhas**, terkhusus untuk **Divisi Penelitian Kabinet Cendekia**. Terima kasih atas ilmu, pengalaman, dan telah menjadi tempat belajar bagi penulis.
15. **Pak Yunus dan Ibu Aminah, adik-adik kelas 5 SDN 31, serta masyarakat pulau Samatellu Lompo** yang dengan tulus dan penuh kasih telah menyambut, membantu penulis selama melaksanakan KKN.
16. Kepada seluruh pihak yang tidak dapat penulis sebutkan satu persatu, terima kasih setinggi-tingginya untuk segala dukungan, partisipasi, dan apresiasi yang diberikan kepada penulis.

Penulis menyadari bahwa masih banyak kekurangan dalam penyusunan skripsi ini, untuk itu dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini dapat memberikan manfaat untuk berbagai pihak.

Makassar, 27 November 2023

EVLYN PRICILIA KONDY
NIM. H051191027

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama	: Evlyn Pricilia Kondy
NIM	: H051191027
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Non-eksklusif** (***Non-exclusive Royalty- Free Right***) atas tugas akhir saya yang berjudul:

"Penerapan *Combine Sampling* Pada Analisis Sentimen Menggunakan Algoritma *K-Nearest Neighbor* (Studi Kasus: Sentimen Kebijakan Lepas Masker di Indonesia)"

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar, 27 November 2023
Yang menyatakan,

EVLYN PRICILIA KONDY
NIM. H051191027

ABSTRAK

Kebijakan lepas masker merupakan salah satu topik yang pernah dibahas di *twitter*. Oleh sebab itu, dapat dilakukan analisis sentimen mengenai topik kebijakan lepas masker. Namun pada data *real* yang diperoleh dari *twitter* dapat mengandung kelas data yang tidak seimbang. Ketidakseimbangan jumlah data dapat mengganggu proses klasifikasi. *Combine sampling* merupakan pendekatan penyeimbangan data dengan menggabungkan metode *oversampling* dan metode *undersampling*. Data pada penelitian ini adalah *tweet* berbahasa Indonesia dengan kata kunci "Kebijakan Lepas Masker". Metode *oversampling* dan metode *undersampling* yang digunakan pada penelitian ini adalah SMOTE dan Tomek *Links* sedangkan untuk metode klasifikasi yang digunakan adalah *K-Nearest Neighbor*. Penelitian ini bertujuan untuk menyeimbangkan jumlah data latih pada kedua kelas yang belum seimbang menggunakan metode *Combine Sampling* serta melakukan klasifikasi sentimen terkait kebijakan lepas masker di Indonesia menggunakan algoritma *K-Nearest Neighbor*. Setelah tahapan pembagian data latih dan data uji diperoleh 234 data latih bersentimen positif dan 652 data latih bersentimen negatif. Jumlah data latih pada kedua kelas tidak seimbang sehingga data pada kelas negatif merupakan data mayor dan data pada kelas positif merupakan data minor. Diperoleh jumlah data latih setelah tahapan *combine sampling* yaitu 613 data pada kelas positif dan 613 data pada kelas negatif. Setelah jumlah data pada kedua kelas telah seimbang dilanjutkan dengan klasifikasi sentimen dan diperoleh akurasi sebesar 60,4%, presisi sebesar 78,5%, dan *recall* sebesar 65%. Penyebab nilai akurasi sebesar 60,4% adalah makna yang ingin disampaikan oleh setiap *tweet* terhadap kebijakan lepas masker di Indonesia tidak terbaca dengan baik oleh pembelajaran mesin yang mengakibatkan terjadinya kesalahan klasifikasi.

Kata Kunci: *Combine Sampling*, SMOTE, Tomek *Links*, *K-Nearest Neighbor*, Analisis Sentimen, Kebijakan Lepas Masker.

ABSTRACT

The policy of removing masks is one of the topics that has been discussed on twitter. Therefore, sentiment analysis can be done on the topic of unmasked policies. But on the data obtained from twitter may contain unbalanced data classes. An imbalance in the amount of data can disrupt the classification process. Combine sampling is a data balancing approach by combining methods oversampling and methods undersampling*. The data in this research are tweets in Indonesian with the keywords "The Policy of Removing Masks". Oversampling method and undersampling method used in this research were SMOTE and Tomek Links while the classification method used is K-Nearest Neighbor. This research aims to balance the amount of training data in the two classes which are not yet balanced using the method* of *Combine Sampling as well as classifying sentiment related to the mask removal policy in Indonesia using an algorithm K-Nearest Neighbor. After dividing the training data and test data, 234 training data with a positive sentiment and 652 training data with a negative sentiment were obtained. The amount of training data in the two classes is not balanced so the data in the negative class is major data and the data in the positive class is minor data. The amount of training data obtained after the stages*combine sampling *namely 613 data in the positive class and 613 data in the negative class. After the amount of data in both classes had been balanced, it continued with sentiment classification and obtained an accuracy of* 60,4%*, precision of* 78,5%*, and recall of* 65%*. The cause of the accuracy value* 60,4% *is the meaning that each person wants to convey the mask removal policy in Indonesia was not read well by machine learning which resulted in misclassification.*

Keywords: *Combine Sampling,* SMOTE, Tomek *Links, K-Nearest Neighbor, Sentiment Analysis, The Policy of Removing Masks.*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi internet yang semakin berkembang memudahkan seseorang untuk mengakses berbagai berita serta memberikan opini. Salah satu layanan media sosial yang memungkinkan pengguna mempublikasikan opini, berita, komentar dalam bentuk pesan pendek adalah *microblogging* (Tresnawati, 2017). Salah satu layanan *microblogging* yang populer adalah *twitter*. Pengguna *twitter* dapat mengemukakan pendapatnya melalui *tweet*. Semakin banyak *tweet* yang membahas suatu topik maka dapat menjadi *trending topic* di *twitter*.

Salah satu *trending topic* yang pernah dibahas di *twitter* adalah "Kebijakan Lepas Masker". Kebijakan lepas masker di Indonesia menimbulkan pro dan kontra dalam masyarakat. Kebijakan ini memberikan kelonggaran bagi masyarakat untuk dapat melepas masker di luar ruangan atau area terbuka yang tidak padat orang. Untuk mengolah dan menganalisis opini-opini pada topik kebijakan lepas masker yang berbentuk teks diperlukan analisis sentimen. Analisis sentimen mengekstraksi informasi dari suatu sentimen (Dang *et al*., 2020). Tujuan dari analisis sentimen yaitu mengklasifikasi sebuah teks yang bersifat positif, netral, dan negatif berdasarkan hasil sentimen (Isnain *et al*., 2021). Sentimen positif merupakan pendapat yang meningkatkan nilai seseorang atau sesuatu sedangkan sentimen negatif merupakan pendapat yang menurunkan nilai seseorang atau sesuatu Terdapat beberapa metode klasifikasi yang digunakan pada analisis sentimen yaitu: *K-Nearest Neighbors* (KNN), *decision tree*, *naïve bayes*, *support vector machine*, dan lain-lain (Mulyani, 2022).

Algoritma KNN merupakan *supervised algorithm* dan teknik klasifikasi yang paling mudah dan sederhana di antara semua algoritma pembelajaran mesin (Prasetio, 2020). Algoritma KNN memiliki performa yang baik untuk analisis sentimen berbahasa Indonesia. Algoritma KNN dapat diterapkan pada data dengan jumlah yang besar. Selain itu algoritma ini juga bersifat nonparametrik sehingga tidak memiliki asumsi terhadap data dan menggunakan sejumlah parameter yang fleksibel (Permanna & Holle, 2021). Namun disamping kelebihan yang dimiliki oleh algoritma KNN terdapat juga kekurangan yaitu pada performa KNN. Performa

KNN bergantung pada pemilihan parameter *k*. Apabila nilai parameter *k* terlalu kecil maka akan mengarah ke kondisi *overfitting* dan apabila nilai parameter *k* terlalu besar maka akan menyertakan terlalu banyak poin dari kelas yang lain (Prasetio, 2020). *Overfitting* merupakan kondisi saat *machine learning* memiliki kinerja yang baik hanya saat pelatihan dilakukan namun kurang baik saat pengujian dilakukan (Widhiyasana *et al*., 2021). Pada penelitian yang dilakukan oleh Baita, dkk. (2021) berjudul "Analisis Sentimen Mengenai Vaksin *Sinovac* Menggunakan Algoritma *Support Vector Machine* (SVM) dan KNN" diperoleh nilai akurasi menggunakan algoritma KNN sebesar 56%. Performa akurasi yang dihasilkan algoritma KNN pada penelitian ini dapat dikatakan cukup rendah dibandingkan dengan algoritma SVM.

Berdasarkan paparan diatas, salah satu cara untuk meningkatkan performa algoritma KNN adalah dengan menyeimbangkan jumlah data yang telah diperoleh. Data *real* yang diperoleh dari suatu sumber dapat mengandung jumlah data yang tidak seimbang. Ketidakseimbangan data terjadi apabila jumlah objek suatu kelas data yang telah dilabeli lebih banyak dibandingkan kelas data lainnya. Kelas data yang tidak seimbang terbagi menjadi dua yakni kelas data mayor dan kelas data minor. Kelas data mayor adalah kelas data yang telah terlabeli dengan jumlah objek lebih banyak dibandingkan kelas data minor (Barro & Afendi, 2013). Ketidakseimbangan data mengakibatkan kesalahan klasifikasi kelas minoritas karena data cenderung mendukung kelas mayoritas (Utami, 2022). Kelas data yang seimbang akan mempengaruhi nilai akurasi yang tinggi dan hasil klasifikasi yang optimal (Fajri & Astuti, 2022). Dalam mengatasi ketidakseimbangan data dapat dilakukan dengan tiga pendekatan yaitu pendekatan tingkat data, pendekatan menggunakan algoritma, dan hibrid (gabungan pendekatan tingkat data dan algoritma). Salah satu pendekatan yang banyak digunakan adalah pendekatan tingkat data. Pendekatan ini mudah dilakukan karena tidak terikat pada metode analisis yang digunakan. Pendekatan ini terbagi menjadi tiga yakni *oversampling*, *undersampling* dan *combine sampling*. Metode *combine sampling* merupakan gabungan dari *oversampling* dan *undersampling*. Secara umum metode ini mampu memberikan hasil yang lebih baik dibandingkan dengan metode *oversampling* maupun *undersampling* (Sain & Purnami, 2015). Hal ini disebabkan pada metode

oversampling dapat menimbulkan masalah *overfitting*. Beberapa metode *oversampling* yaitu SMOTE (*Synthetic Minority Oversampling Technique*), *random over sampling*, *adaptive synthetic* (ADASYN), dan sebagainya. Metode SMOTE merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah *overfitting* yang ditimbulkan dari *oversampling*. SMOTE juga mampu meningkatkan akurasi prediksi bagi kelas minoritas. Untuk meningkatkan performa metode *oversampling* dapat digabungkan dengan metode *undersampling* sebagai metode pembersihan. Beberapa metode *undersampling* yaitu Tomek *Links*, *random under sampling*, *edited nearest neighbors*, dan sebagainya (Indrawati, 2021). Salah satu metode *undersampling* yang dapat digunakan yaitu Tomek *Links*. Metode ini bertujuan untuk mengeliminasi data *noise* yang berada di *borderline* yang dapat menganggu proses klasifikasi.

Beberapa penelitian yang telah menggunakan kombinasi SMOTE dan Tomek *Links* yaitu penelitian yang dilakukan oleh Utami (2022) dengan judul "Analisis Sentimen dari Aplikasi *Shopee* Indonesia Menggunakan Metode *Recurrent Neural Network*" diperoleh nilai akurasi prediksi sebesar 80%, presisi sebesar 84,4%, dan sensitivitas sebesar 92,5%. Pada penelitian lain yang dilakukan oleh Sain dan Purnami (2015) dengan judul "*Combine Sampling Support Vector Machine for Imbalanced Data Classification*" diperoleh nilai akurasi tertinggi sebesar 99,06% untuk dataset *Abalone*.

Berdasarkan penelitian-penelitian sebelumnya mengenai penggunaan kombinasi SMOTE dan Tomek *Links*, peneliti ingin melakukan penelitian mengenai penerapan *combine sampling* terhadap analisis sentimen mengenai kebijakan lepas masker menggunakan algoritma KNN untuk mengklasifikasi sentimen positif dan negatif.

1.2 Rumusan Masalah

Rumusan masalah yang dibahas pada penelitian ini adalah sebagai berikut:

1. Bagaimana penyeimbangan jumlah data latih dengan menerapkan *combine sampling* pada analisis sentimen kebijakan lepas masker di Indonesia?
2. Bagaimana persentase kinerja algoritma KNN dalam klasifikasi sentimen kebijakan lepas masker di Indonesia?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. *Tweet* yang diambil dan dianalisis hanya yang berbahasa Indonesia pada tanggal 1 Mei 2022 sampai 31 Oktober 2022 dengan kata kunci kebijakan lepas masker.
2. Data akan dibagi menjadi data latih dan data uji dengan perbandingan 70 : 30.
3. Pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF).
4. Metode *combine sampling* yang digunakan pada penelitian ini yaitu SMOTE dan Tomek *Links*.
5. Nilai *k* yang digunakan pada algoritma KNN adalah 2 hingga 10.
6. Perhitungan jarak antara data uji dan data latih menggunakan metode jarak *euclidean*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Menyeimbangkan jumlah data latih dengan menerapkan *combine sampling* pada analisis sentimen kebijakan lepas masker di Indonesia.
2. Mendapatkan persentase kinerja algoritma KNN dalam klasifikasi sentimen kebijakan lepas masker di Indonesia.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan informasi mengenai penanganan kelas data tidak seimbang menggunakan kombinasi metode SMOTE dan Tomek *Links*.
2. Memberikan informasi mengenai optimasi klasifikasi algoritma KNN melalui penyeimbangan kelas data menggunakan *combine sampling* pada tahap praproses data.
3. Memberikan gambaran kepada pemerintah dan masyarakat umum mengenai kebijakan lepas masker di Indonesia.

BAB II
TINJAUAN PUSTAKA

2.1 Kebijakan Lepas Masker

Presiden Joko Widodo secara resmi mengeluarkan kebijakan pelonggaran penggunaan masker di luar ruangan pada tanggal 17 Mei 2022. Kebijakan ini dibuat dengan mempertimbangkan jumlah kasus COVID-19 di Indonesia semakin terkendali sehingga masyarakat dapat beraktivitas tanpa masker diluar ruangan. Kebijakan ini sebagai salah satu langkah transisi COVID-19 di Indonesia dari pandemi menjadi endemi (Kementerian Kesehatan Direktorat Promosi Kesehatan dan Pemberdayaan Masyarakat, 2022). Terdapat beberapa syarat dan ketentuan mengenai pelonggaran masker di tempat umum yaitu (Satuan Tugas Penanganan COVID-19, 2022):

1) Diperbolehkan untuk tidak pakai masker jika di luar ruangan atau area terbuka yang tidak padat orang. Apabila di ruangan tertutup dan transportasi publik tetap wajib pakai masker.
2) Masyarakat telah mendapatkan vaksinasi lengkap sehingga dapat melindungi diri sendiri maupun orang lain
3) Tidak memiliki penyakit komorbid dan sedang menderita tuberkulosis (TBC)

2.2 *Twitter Crawling*

Twitter merupakan salah satu *media social* yang dimiliki dan dioperasikan oleh *Twitter*, Inc. berjenis *microblogging* yang digunakan untuk menyebarkan pesan secara singkat dan padat dengan 140 karakter kepada pembacanya di seluruh dunia (Mulyani, 2022). Di Indonesia pengguna *twitter* mencapai 59% dan sebagai peringkat 5 media sosial yang paling banyak digunakan pada tahun 2020 (Krisdiyanto *et al*., 2021). Pengguna *twitter* dapat mengirimkan pesan singkat, *emoticon*, gambar, video dan dapat dilihat secara publik maupun privat (Nugroho, 2018).

Pesan-pesan yang ada pada *twitter* dapat dijadikan sebuah data. *Twitter* telah menyediakan *Application Programming Interface* (API) untuk memperoleh data dan mengolahnya. Proses pengambilan *tweet* dari API *twitter* dapat dilakukan dengan menggunakan *web engine crawel* (Aditya *et al*., 2015). *Crawling*

merupakan teknik untuk mengumpulkan data yang terdapat pada sebuah *website*. Informasi yang diperoleh menyesuaikan dengan kata kunci. Melalui *twitter crawling* akan ditemukan informasi akun, *mention*, *retweet*, favorit, pengikut, teman dari suatu akun (Dikiyanti *et al*., 2021).

2.3 Analisis Sentimen *Twitter*

Analisis sentimen merupakan studi komputasi yang mempelajari mengenai pendapat, penilaian, emosi tiap individu beserta atributnya. Analisis sentimen diperlukan untuk membuat keputusan berdasarkan pendapat baik untuk individu maupun organisasi (Liu, 2010). Tujuan dari analisis sentimen adalah untuk memahami, mengekstrak, mengolah data berupa data teks. Hasil dari analisis sentimen dapat melihat opini mengenai suatu masalah, kecenderungan hal di pasar, persepsi terhadap suatu kualitas pelayanan, melakukan pemantauan terhadap suatu produk, prediksi penjualan, dan pengambilan keputusan (Joang *et al*., 2017).

Salah satu percabangan dari analisis sentimen adalah analisis sentimen *twitter*. *Twitter* sebagai salah satu media sosial yang memiliki pertumbuhan pengguna yang meningkat hari demi hari dan tiap pendapat yang terdapat di *twitter* sangat dibutuhkan baik dari peneliti maupun perusahaan. Analisis sentimen *twitter* mencakup klasifikasi sentimen berupa sentimen positif, netral, dan negatif; mengidentifikasi topik sentimen; dan identifikasi pemegang topik (Zimbra *et al*., 2018). Berdasarkan sumber datanya, analisis sentimen terbagi menjadi dua yaitu *coarse-grained sentiment* dan *fined-grained sentiment*. *Coarse-grained sentiment* berfokus pada level dokumen dan menghasilkan sentimen positif dan negatif sedangkan *fined-grained sentiment* berfokus pada level kalimat dan menghasilkan sentimen positif, netral, negatif (Ardiani *et al*., 2020).

Suatu sentimen dapat dikelompokkan ke dalam suatu kelas yang bersifat positif, negatif (Ardiani *et al*., 2020):

1) Menurut KBBI sentimen positif merupakan reaksi atau sikap yang meningkatkan nilai seseorang atau sesuatu. Contoh sentimen positif "Senangnya bisa beraktivitas di luar ruangan tanpa masker".
2) Menurut KBBI sentimen negatif merupakan reaksi atau sikap yang menurunkan nilai seseorang atau sesuatu. Sentimen negatif umumnya

menggunakan kata negasi. Contoh sentimen negatif "Covid belum selesai tapi sudah ada kebijakan lepas masker parah".

2.4 Praproses Data Teks

Praproses data teks merupakan tahapan awal untuk mengubah suatu dokumen berbentuk teks menjadi data yang berstruktur supaya data dapat diolah lebih lanjut dalam proses *text mining*. Tujuan praproses data teks adalah untuk meningkatkan akurasi dari data. Praproses data teks berbahasa Indonesia perlu memperhatikan aturan penulisan kata serta imbuhan pada kata. Luaran dari praproses data teks yaitu kata-kata dasar (*term*) yang sesuai dengan KBBI. Langkah-langkah praproses data teks sebagai berikut (Kurniawan, 2017):

1) *Data Cleansing*

   Data cleansing adalah proses untuk membersihkan data atau karakter yang dapat mengganggu pengolahan data seperti *hashtag*, *username*, URL, dan tanda baca sehingga hasil yang diperoleh yaitu hanya huruf "a" sampai "z"

2) *Case Folding*

   Case Folding merupakan proses untuk mengubah semua karakter menjadi huruf kecil.

3) *Tokenizing*

   Tahap ini akan dilakukan *tokenize* atau memisahkan isi kalimat menjadi kata-kata individu (*term*).

4) *Stopword removal*

   Metode ini bertujuan untuk menghilangkan kata-kata yang tidak memiliki kontribusi dalam pengklasifikasian teks atau tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat.

5) *Stemming*

   Metode *stemming* bertujuan untuk menghilangkan imbuhan baik awalan, akhiran, sisipan, dan *confixes* (gabungan awalan dan akhiran).

2.5 Data *Splitting*

Data *splitting* merupakan pembagian data menjadi dua atau lebih sub data. Data *splitting* memiliki peran penting dalam pembelajaran mesin terutama dalam

pembuatan model data. Pada pembagian dua data umumnya akan terbentuk data latih dan data uji (Adinugroho, 2022). Data latih merupakan data yang digunakan sebagai acuan untuk membangun model klasifikasi. Data uji merupakan data yang digunakan untuk menguji peforma dari model klasifikasi (Darwis *et al*., 2021).

Tidak terdapat aturan baku yang menyatakan rasio terbaik antara data latih dengan data uji, tapi beberapa analisa empiris menunjukkan bahwa hasil terbaik dicapai jika digunakan sekitar 70-80% untuk data latih dan 20-30% untuk data uji (Putra *et al*., 2021). Beberapa penelitian yang menggunakan proporsi data latih dan data uji 70:30 diantaranya penelitian oleh Rofiqoh dkk (2017) pada analisis sentimen tingkat kepuasan pelanggan telekomunikasi seluler Indonesia pada *twitter* menggunakan metode *support vector machine* dan *lexicon based features*, penelitian oleh Jananto dkk (2021) pada klasifikasi data induk mahasiswa sebagai prediktor ketepatan waktu lulus menggunakan algoritma CART, dan penelitian oleh Lasijan dkk (2023) pada prediksi harga emas dunia menggunakan metode *long-short term memory* sehingga dalam penelitian ini digunakan proporsi data latih dan data uji 70:30.

2.6 Pembobotan Kata

Pembobotan kata merupakan suatu tahapan untuk memberikan bobot (*term*) yang terdapat pada dokumen teks yang akan diolah. Adapun tahapan pada pembobotan kata sebagai berikut (Nurjanah *et al*., 2017):

1) *Term Frequency* (TF) merupakan frekuensi munculnya sebuah kata dalam dokumen.
2) *Document Frequency* (DF) merupakan banyaknya dokumen yang mengandung kata-kata tertentu.
3) *Inverse Document Frequency* (IDF) merupakan frekuensi munculnya kata-kata individu (*term*) pada keseluruhan dokumen teks. Persamaan dari IDF sebagai berikut:

$$idf_t = \ln\left(\frac{N}{df_{(t)}}\right) + 1 \qquad (2.1)$$

Keterangan:

idf_t : *Inverse Document Frequency* dari kata *t*

N : Jumlah dokumen teks

$df_{(t)}$: Jumlah dokumen yang mengandung *term* t

4) *Term Frequency-Inverse Document* (TF-IDF) merupakan perkalian dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) untuk menghitung bobot dan mengetahui seberapa penting suatu *term* dari dokumen. Apabila *term* sering muncul dalam suatu dokumen maka nilai bobot akan semakin besar sedangkan apabila *term* sering muncul dalam banyak dokumen maka akan menghasilkan nilai bobot yang semakin kecil. Persamaan dari TF-IDF ditunjukkan pada Persamaan (2.2).

$$Tf.idf_{t,d} = tf_{td} \times idf_{(t)} \quad (2.2)$$

Keterangan:

$Tf.idf_{t,d}$: *Term Frequency-Inverse Document* (TF-IDF)

tf_{td} : *Term frequency* (TF)

$idf_{(t)}$: *Inverse Document Frequency* (IDF)

2.7 *Combine Sampling*

Sampling merupakan bagian dari ilmu statistik yang berfokus terhadap pemilihan data yang dihasilkan dari satu kumpulan populasi data. Metode *sampling* dapat digunakan untuk mengatasi ketidakseimbangan kelas data. Kelas data yang tidak seimbang terbagi menjadi dua yaitu kelas data mayor dan kelas data minor. Kelas data mayor merupakan kelas data dengan jumlah objek lebih banyak dibandingkan kelas data minor (Barro & Afendi, 2013). Keseimbangan kelas data diperlukan untuk meminimalisir *error* saat proses klasifikasi (Liu *et al*., 2018). Beberapa pendekatan untuk mengatasi ketidakseimbangan data yaitu pendekatan tingkat data, pendekatan menggunakan algoritma, dan hibrid (gabungan pendekatan tingkat data dan algoritma). Pendekatan tingkat data merupakan salah satu pendekatan untuk menyeimbangkan kelas data secara eksternal dengan cara mengurangi sampel pada kelas mayor atau mereplikasi data pada kelas minor. Pendekatan menggunakan algoritma adalah merupakan pendekatan penyeimbangan data secara internal menggunakan sebuah algoritma klasifikasi baru atau memperbaiki suatu algoritma supaya lebih konduktif terhadap data kelas minor (Saifudin & Wahono, 2015). Kemudian terdapat gabungan penyeimbangan

kelas data antara pendekatan tingkat data dan pendekatan menggunakan algoritma yaitu pendekatan hibrid. Pendekatan ini digunakan untuk mengatasi masalah pada pendekatan tingkat data dan algoritma. Pendekatan tingkat data banyak digunakan untuk mengatasi ketidakseimbangan kelas data. Hal ini disebabkan pendekatan ini tidak terikat pada metode analisis yang digunakan (Spelmen & Porkodi, 2018). Pendekatan ini bekerja dengan memodifikasi distribusi data pada tahap praproses (Buda *et al*., 2018). Pendekatan tingkat data terdiri atas tiga teknik yaitu *oversampling*, *undersampling*, dan *combine sampling*.

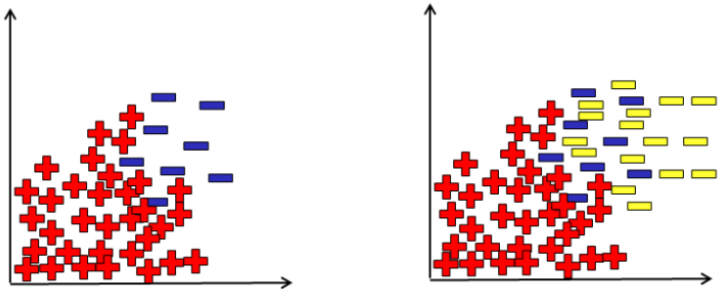
Teknik *oversampling* merupakan teknik penyeimbangan kelas data yang bekerja dengan cara menambah jumlah dataset yang kurang pada kelas minoritas. Beberapa metode *oversampling* yaitu SMOTE (*Synthetic Minority Oversampling Technique*), *random over sampling*, *adaptive synthetic* (ADASYN), dan sebagainya (Indrawati, 2021). Salah satu metode *oversampling* yang dapat digunakan adalah SMOTE. Metode ini banyak digunakan karena mampu menghasilkan akurasi yang baik dan efektif dalam menangani kelas data yang tidak seimbang. SMOTE memiliki kelemahan yaitu dapat menimbulkan masalah *overfitting*. *Overfitting* merupakan suatu kondisi ketika jumlah parameter yang masuk ke dalam model terlalu besar dibandingkan ukuran data yang digunakan untuk membangun sebuah model (Nikmatul Kasanah *et al*., 2017). Selain teknik *oversampling* terdapat juga teknik *undersampling*. Teknik *undersampling* bekerja dengan mengurangi jumlah data pada kelas mayoritas sehingga jumlah data antara kelas mayoritas dan minoritas seimbang. Beberapa metode *undersampling* yaitu *neighborhood cleaning rule* (NCL), Tomek *Links, random under sampling*, *edited nearest neighbors*, dan sebagainya (Indrawati, 2021). Salah satu metode *undersampling* yang dapat digunakan yaitu Tomek *Links*. Metode ini bertujuan untuk mengeliminasi data *noise* yang berada di *borderline* yang dapat menganggu proses klasifikasi. Kelemahan dari metode *undersampling* yaitu berkurangnya informasi penting yang bisa membantu dalam proses klasifikasi (Choirunnisa, 2019).

Cara mengatasi kedua kelemahan yang dimiliki masing-masing teknik dapat digunakan gabungan dari kedua teknik ini. Teknik *combine sampling* menggunakan SMOTE untuk mereplikasi data pada kelas minoritas menggunakan algoritma KNN. Tahap selanjutnya setelah tahap replikasi data pada kelas minoritas akan

diidentifikasi *noise* menggunakan Tomek *Links*. Tahapan *combine sampling* sebagai berikut:

2.7.1 *Synthetic Minority Over-Sampling Technique*

Synthetic Minority Over-Sampling Technique (SMOTE) merupakan sebuah metode *oversampling* dan telah digunakan untuk mengatasi ketidakseimbangan kelas data (Siringoringo, 2018). SMOTE bekerja dengan menyeimbangkan jumlah distribusi data mayoritas dan minoritas. Data kelas minoritas akan diduplikasi hingga jumlah nya sama dengan kelas mayoritas. Ilustrasi cara kerja SMOTE dapat dilihat pada **Gambar 2.1**:



Gambar 2.1 Ilustrasi keadaan data sebelum dan sesudah proses SMOTE

Gambar 2.1 merupakan ilustrasi keadaan data sebelum dan sesudah proses SMOTE. Tahapan SMOTE dalam pembentukan data sintetis dimulai dengan memilih secara acak data observasi yang akan disintesis pada kelas minoritas (x_i). Data sintetis merupakan data baru yang dibangkitkan berdasarkan *k* tetangga terdekat. Kemudian dipilih *k* kelas data yang memiliki jarak terdekat dengan data yang akan di replikasi (x_{knn}). Tahap terakhir data sintesis dibentuk melalui interpolasi secara acak menggunakan dua data sampel dengan persamaan (Douzas *et al*., 2018):

$$x_{syn(ij)} = x_{ij} + \left(x_{knn(ij)} - x_{ij}\right) \times w \qquad (2.3)$$

dengan,

$x_{syn(ij)}$: data sintesis ke-*i term-j* yang akan dibuat

x_{ij} : data ke-*i term-j* yang akan di replikasi

$x_{knn(ij)}$: data ke-*i term-j* yang memiliki jarak terdekat dari x_{ij}

w : nilai acak antara 0 dan 1

Prosedur ini dapat membuat sebanyak mungkin data sintentik untuk kelas

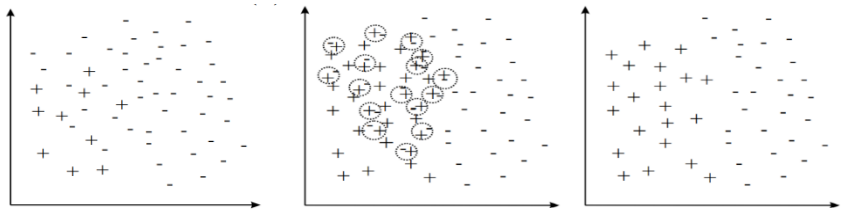
minoritas. Pendekatan ini cukup efektif karena data sintesis yang dibentuk relatif dekat dalam ruang fitur dengan data yang ada dari kelas minoritas. Namun untuk menghindari terjadinya tumpang tindih data diperlukan penggunaan *undersampling* salah satunya Tomek *Links* (Sabilla & Vista, 2021).

2.7.2 Tomek *Links*

Tomek *Links* merupakan sebuah metode *undersampling* yang dibuat pertama kali oleh Ivan Tomek pada tahun 1976. Metode ini bekerja dengan menghapus data pada kelas mayoritas yang dianggap memiliki kesamaan karakteristik (*borderline*). Tomek *Links* juga berperan dalam pembersihan data dari *noise* atau data yang dianggap *misclassify* (Khaulasari, 2016). Tomek *Links* dapat didefinisikan sebagai berikut (Swana *et al.*, 2022):

1) Diberikan dua sampel x dan y milik kelas yang berbeda dan $d(x, y)$ adalah jarak x dan y.
2) Sepasang (x, y) disebut Tomek *Links* jika tidak ada sampel z sehingga $d(x, z) < d(x, y)$ atau $d(y, z) < d(y, x)$

Tomek *Links* merupakan teknik *undersampling* yang bekerja sebagai metode pembersihan yaitu kedua sampel dari kedua kelas yang berbeda akan dihapus. Sistem kerja Tomek *Links* akan berhenti ketika tidak ada lagi data *noise* atau data *borderline* pada data kelas mayoritas (Shintia, 2018). Ilustrasi penerapan Tomek *Links* dapat dilihat pada **Gambar 2.2**:



(1) (2) (3)

Gambar 2.2 Ilustrasi data: (1) data semula, (2) deteksi Tomek *Links*, (3) penghapusan Tomek *Links*

Gambar 2.2 merupakan ilustrasi data mengenai penerapan Tomek *Links* sebagai langkah dalam pembersihan data *noise* pada *borderline*. Setelah jumlah data seimbang maka akan dilakukan pemeriksaan data yang memiliki jarak terdekat namun memiliki kelas yang berbeda. Data tersebut merupakan data *noise* dan

merupakan data yang misklasifikasi. Data *noise* dapat mengganggu proses klasifikasi. Beberapa penelitian yang menggunakan Tomek *Links* untuk membersihkan data *noise* diantaranya penelitian oleh penelitian oleh Sain dan Purnami (2015) yang menggunakan kombinasi SMOTE dan Tomek *Links* dan digabungkan dengan *support vector machine*. Penelitian ini memperoleh nilai akurasi tertinggi sebesar 99,06% pada dataset *abalone*. Selanjutnya penelitian oleh Nugraha dkk (2022) yang membandingkan kinerja tanpa dan menggunakan SMOTE serta kombinasi SMOTE dan Tomek *Links* pada algoritma C5.0 untuk mengatasi ketidakseimbangan kelas pada *credit card fraud* dan diperoleh peningkatan nilai *specificity* dan AUC secara signifikan namun terdapat penurunan nilai akurasi dan *sensitivity* pada penerapan SMOTE dan kombinasi SMOTE dan Tomek *Links*. Penelitian oleh Hairani dkk (2023) yang juga membandingkan kinerja tanpa dan menggunakan SMOTE serta kombinasi SMOTE dan Tomek *Links* pada data klasifikasi penderita diabetes menggunakan *improvement random forest* dan diperoleh peningkatan nilai akurasi, *sensitivity*, *precision*, dan F1-*score* dengan menggunakan kombinasi SMOTE dan Tomek *Links*.

2.8 *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* merupakan algoritma pengklasifikasian berdasarkan *k* buah data latih yang memiliki jarak paling dekat dengan suatu objek. Algoritma ini banyak digunakan untuk klasifikasi teks dan data. Syarat nilai *k* adalah tidak lebih besar dari jumlah data latih, harus bernilai ganjil dan lebih dari 1 (Rivki & Bachtiar, 2017). Proses klasifikasi akan berhenti apabila setiap objek memiliki *class*. Algoritma ini bekerja dengan menghitung jarak antara objek di data uji dan objek di data latih (Damarta *et al*., 2021). Secara umum terdapat tiga metode pengukuran jarak yang sering digunakan untuk mengukur jarak antara data latih yang paling dekat dengan objek yang akan diklasifikasi yaitu (Rivki & Bachtiar, 2017):

1. Jarak *Euclidean*

   Jarak *euclidean* digunakan untuk mengukur kedekatan antara dua buah objek yang digambarkan sebagai garis lurus dalam *euclidean space*. Metode ini merupakan fungsi jarak dari algoritma dasar KNN. Persamaan jarak *euclidean*

untuk mengukur kemiripan data sebagai berikut (Habibi & Santika, 2020):

$$d(x_k, y_k) = \sqrt{\sum_{i=1}^{n} (x_{ki} - y_{ki})^2} \tag{2.4}$$

Keterangan:

$d(x_k, y_k)$: Jarak *euclidean* data latih dokumen *k* dan data uji ke-*k*

x_{ki} : Titik data latih dokumen *k term i*

y_{ki} : Titik data uji ke-*k term i*

n : Banyak kata-kata individu (*term*)

2. Jarak *Manhattan*

Jarak *manhattan* digunakan untuk menghitung perbedaan absolut antara koordinat sepasang objek. Persamaan jarak *manhattan* sebagai berikut:

$$d(x_k, y_k) = \sum_{i=1}^{n} |x_{ki} - y_{ki}| \tag{2.5}$$

Keterangan:

$d(x_k, y_k)$: Jarak *manhattan* data latih dokumen *k* dan data uji ke-*k*

x_{ki} : Titik data latih dokumen *k term i*

y_{ki} : Titik data uji ke-*k term i*

n : Banyak kata-kata individu (*term*)

3. *Cosine Similarity*

Cosine similarity merupakan metode similaritas yang digunakan untuk menghitung similaritas dari dua buah dokumen. Persamaan *cosine similarity* dapat dituliskan sebagai berikut:

$$\cos(d_j, q_k) = \frac{\sum_{i=1}^{n} (td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2} \times \sqrt{\sum_{i=1}^{n} tq_{ik}^2}} \tag{2.6}$$

Keterangan:

$\cos(d_j, q_k)$: Tingkat kesamaan dokumen dengan *query* tertentu

td_{ij} : *Term* ke-i dalam vektor untuk dokumen ke-j

tq_{ik} : *Term* ke-i dalam vektor untuk dokumen ke-k

n : Banyak kata-kata individu (*term*)

Metode jarak *euclidean* memberikan jarak terpendek antara dua titik (jarak lurus) sedangkan jarak *manhattan* memberikan jarak terjauh antara dua data sehingga metode jarak *euclidean* lebih cocok digunakan untuk mengukur kedekatan jarak antara titik data latih dan titik data uji pada algoritma KNN (Simanjuntak *et al*., 2019). Metode *cosine similarity* bekerja dengan mengukur jarak sudut antara dua buah vektor. Metode ini bekerja tanpa memperhatikan besaran vektor dan hanya arahnya saja. Oleh sebab itu metode jarak *euclidean* merupakan metode yang dapat digunakan untuk menghitung jarak terdekat antara data latih dan data uji.

Algoritma *K-Nearest Neighbor* sebagai berikut (Dang *et al*., 2018):

1. Melakukan perhitungan jarak antara data uji dan data latih dengan menggunakan metode jarak *euclidean*.
2. Mengurutkan jarak antara data uji dan data latih dari nilai terkecil.
3. Menentukan nilai parameter *k* dengan *k* merupakan jumlah tetanggaan terdekat dari data uji.
4. Menentukan label berdasarkan label mayoritas tetangga terdekat dari data uji.

Kelebihan dari algoritma *K-Nearest Neighbor* sebagai berikut:

1) Non-linear

Algoritma KNN merupakan algoritma pembelajaran mesin yang bersifat nonparametrik. Karena bersifat nonparametrik maka garis keputusan kelas yang dihasilkan menjadi fleksibel dan nonlinear.

2) Bersifat *self-learning*

Algoritma KNN dapat mempelajari struktur data yang ada dan mengkategorikan data tersebut.

3) Memory Based Approach

Algoritma KNN mudah beradaptasi dengan data latih yang baru. Hal ini memungkinkan algoritma dapat merespon dengan cepat perubahan *input* pada *real time*. Pembelajaran algoritma ini berbasis jarak.

Selain kelebihan, algoritma KNN juga memiliki beberapa kekurangan yaitu algoritma ini sensitif terhadap data pencilan khususnya pencilan yang terletak di tengah-tengah kelas yang berbeda. Algoritma ini juga rentan terhadap data yang berdimensi tinggi dan komputasi yang kompleks (Jaya *et al*., 2020). Algoritma KNN juga bergantung pada penentuan banyaknya parameter *k*. Apabila *k* bernilai 1 maka hasil klasifikasi akan sangat kaku karena hanya memperhitungkan satu tetangga terdekat sedangkan apabila nilai *k* terlalu banyak maka hasil klasifikasi akan samar sehingga diperlukan nilai parameter *k* yang tepat untuk memperoleh klasifikasi dengan tingkat akurasi yang baik (Indrayanti *et al*., 2017). Beberapa penelitian yang menggunakan nilai $k = 1$ hingga $k = 10$ diantaranya penelitian oleh Nasution dan Hayaty (2019) pada analisis sentimen *twitter* dengan membandingkan algoritma KNN dan SVM. Algoritma KNN pada penelitian ini memberikan performa yang lebih baik dibandingkan algoritma SVM dengan nilai akurasi 89,70%. Penelitian selanjutnya oleh Nikmatun dan Waspada (2019) pada implementasi data *mining* untuk klasifikasi masa studi mahasiswa dengan membandingkan kinerja algoritma KNN tanpa dan dengan SMOTE dan diperoleh nilai akurasi terbaik diperoleh tanpa menggunakan SMOTE pada $k = 9$ yaitu sebesar 88%. Terakhir penelitian oleh Briliani dkk (2019) mengenai deteksi ujaran kebencian dalam bahasa Indonesia pada komentar di *instagram* menggunakan algoritma KNN. Akurasi terbaik diperoleh pada nilai $k = 3$ yaitu sebesar 98%.

2.9 *Confusion Matrix*

Confusion matrix merupakan metode yang digunakan untuk menghitung akurasi, presisi dan sensitivitas (*recall*) data dalam proses *data mining*. Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting (Ruuska *et al*., 2018). Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data (Ihsan, 2018). Terdapat empat komponen dalam *confusion matrix* yang dibutuhkan untuk menghitung kinerja dari algoritma klasifikasi yaitu:

1) TP adalah *true positive* yaitu jumlah data positif yang terklasifikasi benar sebagai data positif. Berikut contoh data *true positive*:

Tabel 2.1 Contoh data *true positive*

Tweet	Label Aktual	Label Prediksi
ngonser udah pd ga pake masker ya.. gue jg pingin lepas masker, ikut nyanyi keras2 tp inget kesehatan yg utama ðŸ˜”	Positif	Positif

Tabel 2.1 merupakan contoh data *true positive* karena data memiliki label aktual positif dan benar diklasifikasikan sebagai data positif.

2) TN adalah *true negative* yaitu jumlah data negatif yang terklasifikasi benar sebagai data negatif. Berikut contoh data *true negative*:

Tabel 2.2 Contoh data *true negative*

Tweet	Label Aktual	Label Prediksi
Tetep pakai. Aku soalnya kalau lepas masker di ruang bebas langsung kena flu. Terus juga agak parno, soalnya banyak penyakit baru lagi kan ya. https://t.co/xAwOjwXZIQ	Negatif	Negatif

Tabel 2.2 merupakan contoh data *true negative* karena data memiliki label aktual negatif dan benar diklasifikasikan sebagai data negatif.

3) FP adalah *false positive* yaitu jumlah data positif namun terklasifikasi salah sebagai data negatif. Berikut contoh data *false positive*:

Tabel 2.3 Contoh data *false positive*

Tweet	Label Aktual	Label Prediksi
@adityarahman129 Lepas Masker boleh, tapi ya harus tetap waspada Jokowi Majukan Negeri	Positif	Negatif

Tabel 2.3 merupakan contoh data *false positive* karena data memiliki label aktual positif namun diklasifikasikan sebagai data negatif.

4) FN adalah *false negative* yaitu jumlah data negatif namun terklasifikasi salah sebagai data positif. Berikut contoh data *false negative*:

Tabel 2.4 *Contoh data false negative*

Tweet	Label Aktual	Label Prediksi
Tetap pakai masker meski sudah boleh lepas dan bebas. Karena apa? Karena sudah terbiasa dan biar lebih pede saja. ðŸ¤£	Negatif	Positif

Tabel 2.4 merupakan contoh data *false negative* karena data memiliki label aktual negatif namun diklasifikasikan sebagai data positif. Selanjutnya pengukuran kinerja algoritma klasifikasi sebagai berikut:

1) Akurasi

Akurasi digunakan untuk menghitung ketepatan klasifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap kategorinya. Persamaan untuk menghitung akurasi sebagai berikut (Juba & Le, 2019):

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2.7)$$

2) Presisi

Presisi merupakan rasio jumlah data yang terprediksi benar positif dari seluruh data yang diklasifikasi sebagai data positif. Persamaan untuk menghitung presisi sebagai berikut (Juba & Le, 2019):

$$presisi = \frac{TP}{TP + FP} \times 100\% \quad (2.8)$$

3) *Recall*

Recall digunakan untuk melihat ketepatan antara kelas data positif yang dihasilkan sistem dengan kelas sebenarnya (Juba & Le, 2019). Persamaan untuk menghitung presisi sebagai berikut:

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (2.9)$$