PENERAPAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* PADA SENTIMEN ANALISIS KEBIJAKAN PEMINDAHAN IBUKOTA NEGARA INDONESIA MENGGUNAKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

SKRIPSI

IHKSAN HERIANSYAH

H051181022

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

JULI 2023

PENERAPAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* PADA SENTIMEN ANALISIS KEBIJAKAN PEMINDAHAN IBUKOTA NEGARA INDONESIA MENGGUNAKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

SKRIPSI

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

IHKSAN HERIANSYAH

H051181022

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

JULI 2023

LEMBAR PERNYATAAN KEOTENTIKAN

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

PENERAPAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* PADA SENTIMEN ANALISIS KEBIJAKAN PEMINDAHAN IBUKOTA NEGARA INDONESIA MENGGUNAKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

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

Makassar, 4 Juli 2023

Ihksan Heriansyah

NIM H051181022

PENERAPAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* PADA SENTIMEN ANALISIS KEBIJAKAN PEMINDAHAN IBUKOTA NEGARA INDONESIA MENGGUNAKAN ALGORITMA *NAÏVE BAYES CLASSIFIER*

Disetujui Oleh:

Pembimbing Utama	**Pembimbing Pertama**
Andi Kresna Jaya, S.Si., M.Si.	**Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.**
NIP. 19731228 200003 1 001	**NIP. 19740713 199903 2 001**

Ketua Program Studi

Dr. Anna Islamiyati, S.Si., M.Si.
NIP. 19770808 200501 2 002

Pada 4 Juli 2023

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama	: Ihksan Heriansyah
NIM	: H051181022
Program Studi	: Statistika
Judul Skripsi	: Penerapan *Synthetic Minority Oversampling Technique* Pada Sentimen Analisis Kebijakan Pemindahan Ibukota Negara Indonesia Menggunakan Algoritma *Naïve Bayes Classifier*

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 : Andi Kresna Jaya, S.Si., M.Si. (.....)
2. Sekretaris : Sri Astuti Thamrin,S.Si.,M.Stat.,Ph.D (.....)
3. Anggota : Sitti Sahriman,S.Si.,M.Si. (.....)
4. Anggota : Dra. Nasrah Sirajang,M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 4 Juli 2023

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahi rabbil'alamin, Puji syukur kepada Allah *Subhanallahu Wa Ta'ala* atas segala limpahan berkah, rahmat dan hidayah-Nya yang telah diberikan kepada penulis sampai saat ini sehingga dapat menyelesaikan skripsi dengan judul "Analisis Sentimen Pemindahan Ibukota Negara Indonesia Menggunakan Algoritma *Naïve Bayes Classifier* Dengan Penerapan Metode *Balancing* Data *Synthetic Minority Oversampling Technique*" sebagai salah satu syarat memperoleh gelar sarjana pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

Sholawat serta salam senantiasa tercurahkan kepada baginda Rasulullah Muhammad *Shallallahu'Alaihi Wa sallam*, kepada para keluarga, sahabat, serta orang-orang sholeh yang haq hingga kadar Allah berlaku atas diri-diri mereka.

Perjalanan penjang telah penulis lalui demi sampai pada titik ini. Banyak hambatan yang dihadapi dalam penyusunan skripsi ini, namun berkat dukungan dan bantuan dari pihak yang selalu ada, peduli dan menyayangi penulis sehingga skripsi ini dapat terselesaikan. Oleh karena itu, penulis haturkan rasa terima kasih yang setulus-tulusnya serta penghargaan yang setinggi-tingginya untuk orang tua penulis, Ibunda **Kartini** yang telah memberikan dukungan penuh, pengorbanan, kesabaran hati, cinta dan kasih sayang, serta dengan ikhlas telah mengiringi setiap langkah penulis dengan doa dan restunya.

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

1. **Bapak Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.**, selaku Rektor Universitas Hasanuddin berserta 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 membekali ilmu dan kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika.

4. **Bapak Andi Kresna Jaya, S.Si., M.Si.**, selaku Pembimbing Utama dan **Ibu Sri Astuti Thamrin,S.Si.,M.Stat.,Ph.D.**, selaku Pembimbing Pertama yang dengan penuh kesabaran telah meluangkan waktu dan pemikirannya di tengah berbagai kesibukan dan prioritasnya untuk senantiasa memberikan arahan, dorongan, dan motivasi kepada penulis mulai dari awal hingga selesainya penulisan tugas akhir ini.
5. **Ibu Sitti Sahriman,S.Si.,M.Si.** dan **Dra. Nasrah Sirajang,M.Si.**, selaku Tim Penguji yang telah memberikan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini serta waktu yang telah diberikan kepada penulis.
6. **Ibu Sitti Sahriman,S.Si.,M.Si.**, selaku Penasehat Akademik penulis. Terima kasih atas segala bantuan, nasehat serta motivasi yang selalu diberikan kepada Penulis selama menjalani pendidikan di Departemen Statistika.
7. Seluruh guru TK, SD, SMP, hingga SMA serta para dosen atas ilmu dan pengalaman yang di ajarkan kepada penulis. Sehingga penulis mampu menjadi pribadi yang baik dan mampu mengaplikasikan ilmunya dalam kehidupan sehari hari
8. Teman-teman **Statistika 2018**, terima kasih untuk kebersamaan yang telah dilewati selama menempuh pendidikan di universitas hasanuddin. Untuk suka dan duka dalam kerja sama serta dukungan yang telah diberikan kepada penulis saat menghadapi kesulitan atau hambatan dalam penyelesaian tugas akhir ini.
9. Teman-teman **LT.5, Ari Budak Genshin, Edhy, Haksar, Syahrul, Afik, Dayat, Hapis, Kepin, Ikky, Safir, Nehe**, terima kasih telah selalu membersama penulis selama masa perkuliahan ini. Banyak suka dan duka dari awal perkuliahan ini.
10. Penghuni ruang asisten yakni **Acca Prananda, Haksar, Kepin, Nehe, Hapis, Agra, Noormanto, Afik, Ika dan Rael,** semoga loker menyertai saudara-saudara sekalian**.** Terima kasih telah ada bersama di saat penyelesaian tugas akhir ini. Canda dan tawa selalu ada dalam pertemuan kita sambil mengerjakan tugas akhir masing-masing.
11. Teman-teman **PB. Metaverse** terima kasih untuk dukungan dan kebersamaannya.

12. **Teman teman IPA 3 SMAN 4 Palu angkatan 2018**, terima kasih telah bersama dari tahun 2015 sampai sekarang. Terima kasih atas kebersamaan, canda dan tawanya.
13. **Tina Mahmud**, kamu telah melakukan banyak hal yang luar biasa, saya ingin mengucapkan terima kasih yang sebesar besarnya karena telah hadir dalam hidupku, terima kasih atas kebaikan, dukungan dan perhatiannya. Semoga suatu saat kita dapat bersatu dalam sebuah ikatan. *Love you sweetheart*.
14. Kepada semua pihak yang tidak dapat penulis sebutkan satu-persatu, semoga segala dukungan dan partisipasi yang diberikan kepada penulis bernilai ibadah disisi Allah *Subhanallahu Wa Ta'ala.*

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

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Makassar, 4 Juli 2023

Ihksan Heriasnyah

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

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

Nama	: Ihksan Heriansyah
NIM	: H051181022
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-ekslusif** (***Non-exclusive Royalty- Free Right***) atas tugas akhir saya yang berjudul:

"Penerapan *Synthetic Minority Oversampling Technique* Pada Sentimen Analisis Kebijakan Pemindahan Ibukota Negara Indonesia Menggunakan Algoritma *Naïve Bayes Classifier*"

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 pada tanggal, 4 Juli 2023

Yang menyatakan

(Ihksan Heriansyah)

ABSTRAK

Salah satu topik ataupun isu yang cukup sering dibahas oleh masyarakat pada media sosial twitter belakangan ini adalah wacana pemindahan ibukota negara. Wacana pemindahan ibukota Indonesia memiliki urgensi yang sangat penting bagi kemajuan bangsa utamanya dalam mendongkrak perekonomian. Namun hal ini menimbulkan perdebatan di berbagai kalangan sehingga diperlukannya pendekatan untuk mengekstraksi informasi opini yang terkandung dalam data *text* dari *platform twitter*. Untuk mengekstraksi informasi opini masyarakat mengenai suatu topik pada *twitter* dapat menggunakan analisis sentimen. Proses analisis sentimen melibatkan metode klasifikasi dari teks ke dalam kategori seperti positif ataupun negatif. Dalam penelitian ini, Naïve bayes classifier digunakan untuk proses klasifikasi sentimen yang di kombinasikan dengan balancing terhadap data menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) mengenai pemindahan ibukota negara Indonesia. Penelitian ini menggunakan 10000 data tweet yang kemudian menjadi 5879 tweet setelah melalui tahapan preprocess dengan distribusi kelas setelah melakukan pelabelan manual 2197 tweet bersentimen positif dan 3682 tweet bersentimen negatif, kemudian dilakukan balancing data dengan menambahkan data sintetik pada kelas negatif menggunakan metode SMOTE sehingga diperoleh total data 7364 dengan distribusi kelas yang seimbang. Hasil ketepatan klasifikasi menggunakan naïve bayes classifier pada data yang telah seimbang memiliki akurasi dan f-measure masing-masing 80,72% dan 82,228% nilai ini lebih baik dibandingkan tanpa melakukan penyeimbangan data.

Kata Kunci: Analisis Sentimen, *Balancing* data, *Naïve Bayes Classifier*, Pemindahan Ibukota Negara Indonesia, *Synthetic Minority Oversampling Technique*

ABSTRACT

One of the topics or issues that is quite often discussed by the public on social media twitter lately is the discourse on relocating the country's capital. The discourse on moving the capital of Indonesia has a very important urgency for the progress of the nation, especially for improving the economy. However, this has caused controversy in various circles so that an approach is needed to extract opinion information contained in text data from the twitter platform. To extract public opinion information about a topic on twitter can use sentiment analysis. The sentiment analysis process involves a classification method of text into categories such as positive or negative. In this research, Naïve Bayes classifier is used for sentiment classification process combined with data balancing using Synthetic Minority Oversampling Technique (SMOTE) regarding the relocation of Indonesia's capital city. This study uses 10000 tweet data which then becomes 5879 tweets after going through the preprocess stage with a class distribution after manually labeling 2197 tweets with positive sentiment and 3682 tweets with negative sentiment, then balancing the data by adding synthetic data to the negative class using the SMOTE method so that a total of 7364 data is obtained with a balanced class distribution. The results of classification accuracy using naïve bayes classifier on balanced data have accuracy and f-measure of 80.72% and 82.228% respectively, this value is better than without balancing the data.

Keywords: *Sentiment Analysis, Data Balancing, Naïve Bayes Classifier, Relocation of Indonesia's National Capital, Synthetic Minority Oversampling Technique*

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN JUDUL ii
LEMBAR PERNYATAAN KEONTENTIKAN iii
HALAMAN PERSETUJUAN PEMBIMBING iv
HALAMAN PENGESAHAN v
KATA PENGANTAR vi
HALAMAN PERSETUJUAN PUBLIKASI ix
ABSTRAK x
ABSTRACT xii
DAFTAR ISI xii
DAFTAR TABEL xiv
DAFTAR GAMBAR xv
DAFTAR LAMPIRAN xvi
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Penelitian 4
1.5 Manfaat Penelitian 4
BAB 2 TINJAUAN PUSTAKA 5
2.1 Twitter 5
2.2 Analisis Sentimen 6
2.3 *Preprocessing* 6
2.4 *Term Frequency - Inverse Document Frequency* (TF-IDF) 8
2.5 *Naïve Bayes Classifier* 9
2.6 *Imbalanced Class* 11
2.7 *Synthetic Minority Oversampling Technique* (SMOTE) 11
2.8 Evaluasi Performa Klasifikasi 12

BAB 3 METODELOGI 144
3.1 Sumber Data 14
3.2 Struktur Data 14
3.3 Prosedur Kerja 15
3.3.1 Prosedur Kerja Tanpa Melakukan Balancing Pada Data 16
3.3.2 Prosedur Kerja Dengan Melakukan Balancing Pada Data 16
BAB 4 HASIL DAN PEMBAHASAN 18
4.1 Deskripsi Data 18
4.2 *Preprocess* 21
4.3 *Term Frequency - Inverse Document Frequency* 24
4.4 Klasifikasi Sentimen 24
4.4.1 Data Latih dan Data Uji 24
4.4.2 *Naïve Bayes Classifier* 25
4.5 Kinerja Klasifikasi *Naïve Bayes Classifier* Pada Data *Imbalanced* 36
4.6 Penanganan Data Tidak Seimbang 37
4.7 Kinerja Klasifikasi *Naïve Bayes Classifier* Pada Data *Balanced* 38
4.8 Perbandingan Klasifikasi pada data *Imbalanced* dan data *balanced* 40
BAB 5 KESIMPULAN DAN SARAN 41
5.1 Kesimpulan 41
5.2 Saran 41
DAFTAR PUSTAKA 42
LAMPIRAN 45

DAFTAR TABEL

Tabel 2.1 *Cofusion matrix* 11
Tabel 3.1 Struktur Data 14
Tabel 4.1 Hasil Pengumpulan Data Pada Media Sosial Twitter 18
Tabel 4.2 *Tweet* sebelum dan sesudah dilakukan normalisasi kalimat 21
Tabel 4.3 *Tweet* sebelum dan sesudah dilakukan cleansing data 22
Tabel 4.4 *Tweet* sebelum dan sesudah dilakukan case folding 22
Tabel 4.5 *Tweet* sebelum dan sesudah dilakukan tokenizing 23
Tabel 4.6 *Tweet* sebelum dan sesudah dilakukan stopwords removal 23
Tabel 4.7 *Tweet* sebelum dan sesudah dilakukan stemming 23
Tabel 4.8 Jumlah pembagian data latih dan data uji 25
Tabel 4.9 Contoh 5 kalimat data latih 25
Tabel 4.10 Contoh Data Uji 26
Tabel 4.11 Perbandingan klasifikasi *naïve bayes classifier* pada data *imbalanced* dan klasifikasi *naïve bayes classifier* pada data *balanced* 40

DAFTAR GAMBAR

Gambar 4.1 Pie Chart pendapat pemindahan ibukota Indonesia..... 20
Gambar 4.2 Confusion matrix data tweet pendapat pemindahan ibukota Indonesia menggunakan naïve bayes classifier..... 36
Gambar 4.3 Proporsi kelas data sebelum dan setelah penerapan metode SMOTE 38
Gambar 4.5 Confusion matrix data tweet pendapat pemindahan ibukota Indonesia yang telah seimbang menggunakan naïve bayes classifier 39

DAFTAR LAMPIRAN

Lampiran 1. Hasil perhitungan *Term Frequency* (TF) pada data *tweet* pendapat pemindahan ibukota Indonesia. 46

Lampiran 2. Hasil perhitungan nilai *Inverse Document Frequency* (IDF) pada data *tweet* pendapat pemindahan ibukota Indonesia 47

Lampiran 3. Bobot dari setiap kata dasar pada data *tweet* pendapat pemindahan ibukota Indonesia dari hasil perkalian nilai TF dan IDF 48

Lampiran 4. Data *Training* hasil *splitting dataset*..... 50

Lampiran 5. Data *Test* hasil *splitting dataset*..... 51

BAB I
PENDAHULUAN

1.1 Latar Belakang

Microblog adalah *platform* bagi pengguna dunia maya bertukar informasi dalam bentuk pesan pendek, alamat *website*, gambar, dan video. *Platform microblog* telah menjadi sumber informasi mengenai banyak permasalahan karena sifatnya yang *real-time*. *Twitter* merupakan salah satu bentuk *microblog*, menjadi wadah untuk bertukar pikiran tentang berbagai topik di berbagai bidang dan masalah kehidupan sehari-hari. Misalnya, mengungkapkan opini positif tentang suatu produk atau opini negatif tentang suatu kebijakan pemerintah.

Salah satu topik ataupun isu yang cukup sering dibahas oleh masyarakat pada media sosial twitter adalah wacana pemindahan ibukota negara yang disampaikan oleh bapak presiden joko widodo pada tanggal 29 april 2019, Presiden Joko widodo memutuskan untuk memindahkan ibukota negara Indonesia keluar dari pulau jawa dan dicantumkan dalam RPJMN 2020-2024. Wacana pemindahan ibukota Indonesia memiliki urgensi yang sangat penting bagi kemajuan bangsa utamanya dalam mendongkrak perekonomian. Namun hal ini menimbulkan perdebatan opini di berbagai kalangan masyarakat indonesia khususnya pada media sosial twitter.

Perdebatan mengenai wacana pemindahan ibukota menunjukkan perhatian yang kolektif sehingga diperlukannya pendekatan untuk mengekstraksi informasi opini yang terkandung dalam data *text* dari *platform twitter* (Rintyarna, 2017). Dikarenakan semakin berkembangnya teknologi dan pertumbuhan jumlah data *digital*. Pentingnya pengolahan data *digital* karena manfaatnya yang cukup berpengaruh. Salah satunya dalam bidang pemerintahan, tanggapan masyarakat mengenai suatu kebijakan dapat dijadikan pertimbangan dalam implementasinya.

Analisis sentimen adalah teknik mengekstraksi tanggapan masyarakat mengenai suatu topik pada *twitter*. Analisis sentimen dapat didefinisikan sebagai proses yang dirancang agar dapat bekerja secara otomatis agar memperoleh sikap, pandangan, opini, dan emosi dari ucapan, teks, *tweet*, dan sumber basis data. Proses ini melibatkan metode klasifikasi dari teks ke dalam kategori seperti positif ataupun negatif (Uma dan Aloysius, 2018). Ada beberapa metode klasifikasi yang dapat

digunakan analisis sentimen, diantaranya: *support vector machine, decision tree*, *naïve bayes classifier,* dan lain-lain.

Salah satu metode klasifikasi yang sering digunakan adalah metode naïve bayes yang merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Pada penelitian sebelumnya, yang termuat pada penelitian Fikri, dkk. (2020) tentang Analisis Sentimen di media sosial *twitter* untuk mengetahui opini masyarakat mengenai Universitas Muhammadiyah Malang, membuktikan bahwa metode *naïve bayes classifier* memiliki akurasi yang lebih tinggi dibandingkan metode *Support Vector Machine* yaitu 73,65%

Semua metode klasifikasi yang ada mengasumsikan keseimbangan antar sampel pada data berlabel. Kasus yang lebih umum terjadi pada data twitter adalah distribusi kelas mengalami ketidakseimbangan. Kelas yang mempunyai data sampel sangat besar disebut kelas mayoritas, sedangkan kelas yang memiliki data sampel yang sedikit disebut kelas minoritas. Pada masalah distribusi kelas mengalami ketidakseimbangan, lebih sulit untuk mengklasifikasikan anggota dari kelas minoritas daripada anggota kelas mayoritas dikarenakan kurangnya sampel pada data minoritas sehingga hasil prediksi cenderung menghasilkan kelas mayoritas (Gopalakrishnan dan Ramaswamy, 2014).

Salah satu pendekatan dalam menangani permasalahan distribusi data yang tidak seimbang adalah *Synthetic Minority Oversampling Technique* (SMOTE) yang menggunakan teknik *over-sampling* dengan mensintesis data pada kelas minor sehingga data menjadi seimbang antara data pada kelas mayoritas dan minoritas adalah definisi dari metode SMOTE (Chawla dkk, 2002). Pada penelitian Erlin, dkk. (2022) yang mengkombinasikan metode klasifikasi *random forest* dengan SMOTE untuk mengklasifikasikan dataset penyakit jantung, membuktikan kombinasi antara *random forest* dan SMOTE mencapai akurasi klasifikasi yang cukup memuaskan dibandingkan dengan hanya menggunakan *random forest*.

Berdasarkan hal tersebut, sehingga penelitian yang dilakukan adalah analisis sentimen pada opini masyarakat berupa sentimen yang didapatkan dari media sosial *twitter* menggunakan metode *naïve bayes classifier* dengan mengkombinasikan metode SMOTE untuk mengetahui pro dan kontra masyarakat mengenai kebijakan

pemindahan ibukota negara Indonesia, serta memperoleh hasil klasifikasi terbaik dari pengujian sistem yang dilakukan.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas maka yang akan dibahas dalam penulisan ini yaitu:

1. Bagaimana hasil klasifikasi sentimen analisis pada opini masyarakat terhadap kebijakan pemindahan ibukota negara Indonesia Indonesia dengan metode *naïve bayes classifier* yang dikombinasikan dengan *synthetic minority oversampling technique*?
2. Bagaimana hasil kinerja metode *naïve bayes classifier* yang dikombinasikan dengan *synthetic minority oversampling technique* dalam klasifikasi opini masyarakat tentang kebijakan pemindahan ibukota negara Indonesia?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab permasalahan yang telah dirumuskan sebelumnya, yaitu:

1. Memperoleh hasil klasifikasi sentimen analisis pada opini masyarakat terhadap kebijakan pemindahan ibukota negara Indonesia Indonesia dengan metode *naïve bayes classifier* yang dikombinasikan dengan *synthetic minority oversampling technique.*
2. Memperoleh hasil kinerja metode *naïve bayes classifier* yang dikombinasikan dengan *synthetic minority oversampling technique* dalam klasifikasi opini masyarakat tentang kebijakan pemindahan ibukota negara Indonesia.

1.4 Batasan Penlitian

Penelitian yang dilakukan difokuskan kepada setiap *tweet* masyarakat yang berhubungan dengan kebijakan pemindahan ibukota pada platform media sosial *twitter* dengan jumlah *tweet* yang akan digunakan yakni sebanyak 10000 *tweet*. Data dari *Twitter* tersebut akan diproses dalam 2 skema yakni dengan menggunakan algoritma *naïve bayes classifier* dan kombinasi antara algoritma *naïve bayes classifier* dengan *synthetic minority oversampling technique*.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah peneliti mengharapkan dapat membantu pihak-pihak yang ingin menganalisa pandangan atau sentimen masyarakat terhadap suatu kejadian atau isu yang sedang beredar dengan persentase antara kelas mayoritas dan kelas minoritas yang tidak seimbang serta memberikan gambaran hasil performansi *naïve bayes classifier* yang dipadukan dengan *synthetic minority oversampling technique* untuk mengatasi ketidakseimbangan kelas dalam sentimen analisis.

BAB II
TINJAUAN PUSTAKA

2.1 *Twitter*

Twitter adalah layanan jejaring sosial *microblog* gratis yang memungkinkan anggota yang terdaftar menyiarkan pesan singkat yang disebut *tweet*. Pengguna *twitter* dapat menyiarkan *tweet* dan mengikuti *tweet* pengguna lain. *Tweet* dan balasan ke *tweet* dapat dikirim melalui pesan teks ponsel, *klien desktop* atau dengan memasang pada situs twitter.com. Tidak seperti media sosial lainnya, *twitter* dianggap *microblog* karena aktivitas utamanya berkisar pada *posting Update* singkat atau *tweet* menggunakan *website* atau ponsel. Kepopuleran *twitter* mengakibatkan banyak peneliti yang tertarik menganalisis data *twitter* untuk berbagai tugas berbeda seperti membuat prediksi, mendeteksi sentimen pengguna terhadap topik yang berbeda, mendeteksi emosi pengguna dan mendeteksi ironi. (Simanjuntak dan Pramana, 2021)

2.2 Analisis Sentimen

Analisis sentimen merupakan sebuah metode yang digunakan untuk mengekstrak data opini, memahami serta mengolah tekstual data secara otomatis untuk melihat sentimen yang terkandung dalam sebuah opini. (Sari dan Wibowo, 2019). Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif dan negatif.

Analisis sentimen dapat dibedakan berdasarkan sumber datanya, beberapa level yang sering digunakan dalam penelitian sentimen analisis adalah pada level dokumen dan pada level kalimat. Berdasarkan level sumber datanya sentimen analisis terbagi menjadi dua kelompok besar yaitu (Fatima dkk, 2020):

1. *Coarse-grained Sentiment Analysis*

   Pada *coarse-grained sentiment analysis*, analisis sentimen yang dilakukan adalah pada level dokumen. Secara garis besar fokus utama dari analisis sentimen jenis ini adalah menganggap seluruh isi dokumen sebagai sebuah sentimen positif atau sentimen negatif.

2. *Fined-grained Sentiment Analysis*

   Fined-grained sentiment analysis adalah analisis sentimen pada level kalimat. Fokus utama *fined-grained sentiment analysis* adalah menentukan sentimen pada setiap kalimat.

2.3 *Preprocessing*

Dokumen pada umumnya mempunyai struktur yang sembarangan atau tidak terstruktur. Oleh karena itu, diperlukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur ke dalam bentuk data yang terstruktur. Struktur dari dokumen tersebut memiliki beragam variasi mulai dari bentuk huruf sampai tanda baca. Variasi huruf harus diseragamkan yaitu dengan menjadikan huruf besar saja atau huruf kecil saja. Selain itu, proses penghilangan tanda baca dilakukan untuk menghilangkan noise pada saat pengambilan informasi. (Jumeilah, 2017).

Preprocessing data merupakan langkah penting dalam proses penemuan pengetahuan, karena keputusan-keputusan yang berkualitas harus didasarkan pada data yang berkualitas. *Preprocessing* data seringkali digunakan untuk mengurangi kesalahan data dan sistematis bias dalam data mentah sebelum analisis apapun terjadi. Ada banyak faktor yang menimbulkan *problem* peforma klasifikasi. Terutama adalah bentuk dan kualitas data, bila data mengandung *noise*, redundansi dan mempunyai data yang tidak relevan, kondisi tersebut membuat proses pengekstraksian fitur selama fase pelatihan lebih sulit (Setyohadi dkk, 2017). Berikut tahapan dalam preprocessing data:

1. *Case Folding* adalah proses mengubah format data teks yang memiliki bentuk *uppercase* (huruf kapital) menjadi format data teks yang memiliki bentuk *lowercase* (huruf kecil).
2. Data *Cleansing* merupakan proses membersihkan data teks yang akan digunakan dari karakter-karakter hingga kata-kata yang tidak diperlukan. Karakter yang dianggap tidak valid seperti angka, tanda baca, *Uniform Resources Locator* (URL), *hashtag*, dan segala jenis karakter lainnya yang dapat menimbulkan proses perhitungan dalam pengklasifikasian tidak berjalan optimal.
3. *Stemming* adalah proses mengubah kata yang mendapatkan imbuhan menjadi kata dasarnya. Hal ini bertujuan untuk meningkatkan performa dan mengurangi

penggunaan *resource* dari sistem dengan mengurangi jumlah *unique word* yang harus diakomodasikan oleh sistem.

4. *Stopwords* adalah kata-kata yang tidak terlalu penting untuk digunakan di *text mining*. Biasanya kata-kata ini disaring atau dibuang dari teks yang akan dievaluasi karena *stopwords* ini akan menyebabkan banyak informasi yang tidak perlu.
5. *Tokenizing* merupakan tugas untuk memecah kalimat menjadi bagian-bagian yang disebut token dan menghilangkan bagian tertentu seperti tanda baca. *Tokenizing* memberikan urutan karakter dan sebuah unit dokumen terdefinisi.
6. Normalisasi Kalimat merupakan proses mengubah kata-kata yang tidak baku menjadi kata yang baku

2.3 Term Frequency - Inverse Document Frequency (TF-IDF)

Setelah tahap *preprocessing*, selanjutnya data yang diperoleh dilakukan proses pembobotan data berupa kata menjadi numerik dengan menggunakan metode *Term Frequency Inverse Document Frequency* (TF-IDF). Pembobotan TF-IDF merupakan ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen atau di dalam sebuah kalimat. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting dan seberapa umum kata tersebut di dalam dokumen tersebut. Bobot kata semakin besar jika sering muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen (Melita dkk, 2018). Pembobotan TF-IDF diperoleh dengan menggunakan persamaan:

$$idf_j = \log \frac{N}{df_j} \quad (2.1)$$

$$W_{i,j} = tf_{ij} \times idf_j \quad (2.2)$$

Keterangan:

$W_{i,j}$ = Bobot *term* ke – j terhadap dokumen ke - i

tf_{ij} = Jumlah kemunculan *term* j ke dalam dokumen i

N = Jumlah dokumen secara keseluruhan

df_j = Jumlah dokumen yang mengandung *term* j

2.5 *Naïve Bayes Classifier*

Salah satu tugas Data *Mining* adalah klasifikasi data, yaitu memetakan data ke dalam satu kelas atau beberapa kelas yaang sudah didefinisikan sebelumnya. Salah satu metode dalam klasifikasi data adalah *Naïve Bayes* yaitu salah satu metode *machine learning* yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Cara kerja *Naive Bayes* yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Dasar dari *Naïve Bayes* yang dipakai dalam pemrograman adalah rumus *Bayes* seperti pada persamaan berikut (Darwis dkk, 2021):

$$P(C|X) = \frac{(P(X|C) \times P(C))}{P(X)} \tag{2.3}$$

dengan:

X : Atribut

C : Kelas

$P(C|X)$: Peluang bersyarat dari peristiwa C yang terjadi mengingat X terjadi

$P(C)$: Peluang peristiwa C

$P(X)$: Peluang peristiwa X

Proses klasifikasi *Naïve Bayes Classifier* terhadap dokumen yaitu dengan merepresentasikan setiap dokumen atau kalimat berlabel yang dapat ditulis sebagai atribut $X = (x_1, x_2, \dots, x_n)$ yang mempunyai makna bahwa x_1 untuk kata pertama, x_2 adalah kata kedua, dan seterusnya. Setiap dokumen atau kalimat berlabel tersebut digolongkan dalam k kelas $C_1, C_2, \dots, C_k$. *Naïve bayes classifie*r akan mengklasifikasikan sampel X yang diberikan sedemikian sehingga sampel tersebut berada pada kelas yang memiliki peluang posterior tertinggi. Yaitu, X, diprediksi berada pada kelas C_i jika dan hanya jika (Buche dkk, 2013)

$$P(C_i|X) > P(C_j|X), \qquad untuk\ 1 \le i, j \le k, \qquad i \ne j$$

Berdasarkan Teorema Bayes

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$P(C_i|X) = \frac{P(x_1, x_2, \dots, x_n|C_i)P(C_i)}{P(x_1, x_2, \dots, x_n)} \tag{2.4}$$

Nilai dari $P(x_1, x_2, x_3, \dots, x_n)$ adalah konstan untuk semua kategori C_i maka dari persamaan (2.4) dapat ditulis:

$$P(C_i|X) = P(x_1, x_2, \dots, x_n|C_i)P(C_i)$$
$$P(C_i|X) = P(x_1|C_i)P(x_2, \dots, x_n|C_i, x_1)P(C_i)$$
$$P(C_i|X) = P(x_1|C_i)\, P(x_2|C_i, x_1)P(x_n|C_i, x_1, x_2, \dots, x_{n-1})P(C_i) \quad (2.5)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleks faktor – faktor yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk di analisa satu persatu. Akibatnya, perhitungan menjadi sulit untuk dilakukan, maka digunakan asumsi independensi sangat tinggi (*naïve*), bahwa masing – masing $(x_1, x_2, \dots, x_n)$ saling bebas satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(x_i|x_j) = \frac{P(x_i \cap x_j)}{P(x_j)}$$

$$P(x_i|x_j) = \frac{P(x_i)P(x_j)}{P(x_j)}$$

$$P(x_i|x_j) = P(x_i)$$

Untuk $i \neq j$, sehingga

$$P(x_i|C_i, x_j) = P(x_i|C_i)$$

Atau dapat dituliskan dalam notasi

$$P(C_i|x_1, x_2, \dots, x_n) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \quad (2.6)$$

Dalam Naïve Bayes Classifiers kita perlu memaksimalkan nilai probabilitas setiap kelas, maka

$$P(C_i|x_1, x_2, \dots, x_n) = argmax\, P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \quad (2.7)$$

Sehingga ditemukan kelas yang memaksimumkan $P(C_i|X)$. Nilai maksimum $P(C_i|X)$ untuk kelas C_i dinamakan *maximum posterior hypothesis*.

Nilai $P(C_i)$ dapat dihitung menggunakan persamaan berikut:

$$P(C_i) = \frac{|docs_i|}{|contoh|} \quad (2.8)$$

Keterangan:

docs_i = Jumlah dokumen pada setiap kategori j

contoh = Jumlah dokumen yang ada

Peluang fitur bersyarat $P(x_i|C_i)$ dapat dihitung dengan rumus sebagai berikut

$$P(x_i|v_j) = \frac{count(x_i, C_i)}{count(C_i)} \quad (2.9)$$

Namun jika kata x_k tidak tersedia dalam data *training* dari kelas C_i tapi muncul dalam teks dokumen peluang (x_k, C_i) menjadi 0 yang berarti peluang data *testing* berada pada kelas C_i menjadi 0. Untuk menghindari hal itu kita menggunakan laplace smoothing dan peluang fitur bersyaratnya di hitung dengan rumus berikut:

$$P(x_i|v_j) = \frac{count(x_i, C_i) + 1}{count(C_i) + Jumlah\ bobot\ kata\ pada\ dokumen} \quad (2.10)$$

2.6 *Imbalanced Class*

Imbalance class adalah kondisi distribusi antar kelas yang tidak seimbang pada suatu dataset, dimana salah satu kelasnya memiliki jumlah data yang sangat besar (kelas mayoritas) dibanding kelas lainnya (kelas minoritas). Perbedaan jumlah data yang besar antar kelas dapat mengakibatkan model klasifikasi sering tidak dapat memprediksikan kelas minoritas dengan tepat sehingga banyak data tes yang seharusnya berada pada kelas minoritas diprediksikan salah oleh model klasifikasi.

Untuk mengatasi permasalahan *imbalance class*, salah satu metode yang digunakan adalah *sampling*. Metode *sampling* melakukan modifikasi terhadap distribusi data antar kelas mayoritas dan kelas minoritas pada dataset *training* untuk menyeimbangkan jumlah data tiap kelas. Salah satu metode sampling yang sering digunakan adalah *Syntetic Minority Over-sampling Technique* (Sutoyo dan Fadlurrahman, 2020).

2.7 *Synthetic Minority Oversampling Technique*

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla sebagai salah satu solusi dalam menangani data tidak seimbang dengan prinsip yang berbeda. Bila Metode *oversampling* berprinsip memperbanyak pengamatan secara acak maka, metode SMOTE

menambah jumlah data kelas minoritas agar setara dengan kelas mayoritas dengan cara membangkitkan data buatan. Data buatan atau sintetis tersebut dibuat berdasarkan *k-nearest neighbor*. (Saputra dkk, 2020)

Metode ini bekerja dengan menambahkan jumlah data pada kelas atau kategori minoritas dengan cara membangkitkan data baru berdasarakan *k-nearest neighbor*. Penentuan jumlah replikasi yang dilakukan disesuaikan dengan jumlah anggota pada kelas mayoritas. Penambahan jumlah pada data kelas minoritas dilakukan dengan membuat data sintetis di sepanjang garis yang menghubungkan salah satu atau semua *k-nearest neighbor* data kelas minoritas tersebut. Dalam menentukan *k-nearest neighbor* digunakan perhitungan jarak *euclidean*. (Wijayanti dkk, 2021).

Misalkan terdapat dua struktur data dengan n dimensi yaitu $X = (x_1, x_2, \dots, x_n)$ dan $Y = (y_1, y_2, \dots, y_n)$, maka jarak *euclidean* $d(x, y)$ yang dihasilkan antara kedua data ditunjukkan pada persamaan berikut:
(Saputra dkk, 2020)

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2} \qquad (2.11)$$

Dalam Jo*urnal of Artificial Intelligence Research* dengan judul SMOTE: *Synthetic Minority Over-sampling Technique* karya Chawla dkk. (2002) jarak *euclidean* sebagai metode perhitungan jarak dalam teknik SMOTE dapat dipengaruhi oleh beberapa alasan

1. Kesederhanaan dan Efektivitas: Metode jarak Euclidean relatif sederhana, tetapi cukup efektif dalam banyak skenario. Kelebihan kesederhanaannya termasuk ketersediaan implementasi yang luas, penggunaan sumber daya yang lebih sedikit, dan kemampuan untuk mengelola data berdimensi tinggi dengan baik.
2. Kesesuaian dengan Data Numerik: SMOTE adalah teknik yang umumnya digunakan pada data numerik atau data berdimensi tinggi. Jarak Euclidean telah terbukti berhasil dalam banyak kasus pengolahan data numerik dan dapat memberikan hasil yang baik dalam mengukur kedekatan antara data dalam ruang berdimensi tinggi.

3. Interpretasi yang Intuitif: Jarak Euclidean memiliki interpretasi yang intuitif dalam ruang geometris. Dalam banyak kasus, kita dapat membayangkan jarak Euclidean sebagai panjang garis lurus yang menghubungkan dua titik dalam ruang. Ini membuatnya lebih mudah dipahami dan diinterpretasikan secara visual.

Algoritma yang bekerja pada SMOTE pertama akan mengambil nilai selisih antara vektor dari fitur pada kelas minoritas dan nilai *nearest neighbor* dari kelas minoritas lalu mengalikan nilai tersebut dengan angka acak antara 0 sampai 1. Selanjutnya, hasil kalkulasi tersebut ditambahkan dengan vektor fiturnya sehingga didapatkan hasil nilai vektor yang baru (Edi dan Asri, 2020).

$$x_{syn} = x_i + (x_{knn} - x_i) \times \tau \quad (2.12)$$

Keterangan:

x_{syn} = Data hasil replikasi

x_i = Data yang akan direplikasi

x_{knn} = *K-nearest neighbor* untuk x_i

τ = Bilangan random 0 sampai 1

2.8 Evaluasi Performa Klasifikasi

Confusion matrix merupakan tabel yang menggambarkan performa dari sebuah algoritma atau metode klasifikasi dalam memprediksi kelas suatu data. Pada penelitian ini data diklasifikasikan menjadi positif dan negatif. Setiap baris mempresentasikan kelas aktual dari data dan setiap kolom mempresentasikan kelas prediksi dari data. Tabel 2.1 menunjukkan tabel *confusion matrix*.

Tabel 2.1 *Cofusion matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP (*True Positive*)	FP (*False Positive*)
Negatif	FN (*False Negative*)	TN (*True Negative*)

Dalam pengukuran performa klasifikasi terdapat beberapa cara, namun cara yang paling sering digunakan adalah dengan menghitung akurasi, *precision*, *recall* dan *f-measure*. Akurasi merupakan persentase dari total sentimen yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data

sentimen yang benar dengan total data dan data uji. Untuk menghitung nilai akurasinya dilakukan dengan menggunakan Persamaan

$$Akurasi = \frac{Jumlah\ Sentimen\ Benar}{Jumlah\ Data\ Tes} \times 100\% \quad (2.13)$$

Untuk pengukuran performa klasifikasi cara berikutnya yang dapat digunakan adalah menghitung *precission, recall* dan *f-measure*. *Precision* merupakan rasio data terprediksi benar positif dibandingkan dengan keseluruhan data yang diprediksi positif. Perhitungan *precision* dilakukan dengan cara membagi jumlah data benar yang bernilai positif dibagi dengan jumlah data benar yang bernilai positif dan data salah yang bernilai positif.. Untuk menghitung nilai precision dapat dilakukan dengan menggunakan Persamaan

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2.14)$$

Recall merupakan rasio data terprediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Perhitungan *recall* dilakukan dengan cara membagi data benar bernilai benar positif dengan hasil penjumlahan dari data yang bernilai benar positif dan data yang bernilai negatif namun seharusnya bernilai positif. Perhitungan *recall* dapat menggunakan Persamaan:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2.15)$$

F-measure merupakan rata – rata harmonic dari *precision* dan *recall*. Nilai *f-measure* didapat dari perhitungan hasil perkalian *precision* dan *recall* dibagi dengan hasil penjumlahan *precision* dan *recall* kemudian dikalikan dua dan perhitungan *f-measure* menggunakan Persamaan (Sulistiyono dkk, 2021):

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.16)$$