

**PENGARUH SYNTHETIC MINORITY OVERSAMPLING  
TECHNIQUE (SMOTE) PADA ANALISIS SENTIMEN  
 MENGGUNAKAN ALGORITMA SUPPORT VECTOR  
 MACHINE (SVM)**

Diajukan Sebagai Syarat Untuk Menyelesaikan  
Pendidikan Program Strata-1 Pada  
Jurusn Teknik Informatika



Oleh:

HELMI FARIZKI  
NIM: 09021281924037

Jurusan Teknik Informatika  
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA  
2023

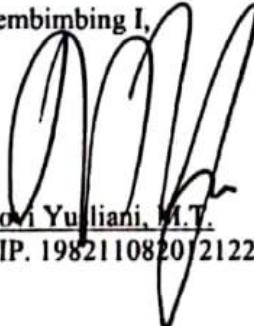
## LEMBAR PENGESAHAN SKRIPSI

### **PENGARUH SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) PADA ANALISIS SENTIMEN MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)**

Oleh:

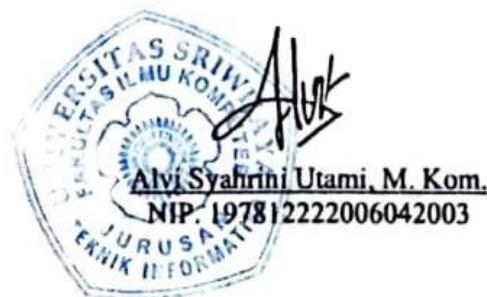
Helmi Farizki  
NIM : 09021281924037

Palembang, 29 Maret 2023

Pembimbing I,  
  
Novi Yuliiani, M.T.  
NIP. 198211082012122001

Pembimbing II,  
  
Annisa Darmawahyuni, M.Kom.  
NIP. 1671147006900002

Mengetahui,  
Ketua Jurusan Teknik Informatika



## TANDA LULUS UJIAN SIDANG SKRIPSI

Pada hari **Jumat** tanggal **24 Maret 2023** telah dilaksanakan ujian sidang skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Helmi Farizki  
NIM : 09021281924037  
Judul : Pengaruh *Synthetic Minority Oversampling Technique* (SMOTE) Pada Analisis Sentimen Menggunakan Algoritma *Support Vector Machine* (SVM)

dan dinyatakan **LULUS**

1. Ketua Pengaji

Yunita, M.Cs.

NIP. 198306062015042002

.....

2. Pengaji I

Alvi Syahrini Utami, M.Kom.

NIP. 197812222006042003

.....

3. Pembimbing I

Novi Yusliani, M.T.

NIP. 198211082012122001

.....

4. Pembimbing II

Annisa Darmawahyuni, M.Kom.

NIP. 1671147006900002

.....

Mengetahui,

Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom

NIP. 197812222006042003



## HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Helmi Farizki

NIM : 09021281924037

Program Studi : Teknik Informatika

Judul Skripsi :*PENGARUH SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) PADA ANALISIS SENTIMENT MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)*

Hasil Pengecekan Software Thenticate/Turmitin : 13 %

Menyatakan bahwa Laporan Proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.



Inderalaya, 29 Maret 2023



Helmi Farizki

NIM. 09021281924037

## **MOTTO**

*Jangan Memperlambat Suatu Pekerjaan Hanya dikarenakan Malas Karena Akhir  
dari Permasalahan Tersbeut Hanyalah Penyesalan.*

Ku persembahkan karya tulis ini kepada:

- Allah Swt.
- Keluarga
- Sahabat
- Universitas Sriwijaya

## **ABSTRACT**

Imbalanced data is a problem that often occurs when conducting research in the field of sentiment analysis. This problem occurs when the dataset being analyzed has more positive classes than negative classes or vice versa so that the terms majority and minority data appear. If the majority data is more dominant then the classification process tends to produce a classification into the majority class. This causes the need for a solution to overcome the problem of imbalanced data. One of the methods to overcome this data is by using the Synthetic Minority Oversampling Technique (SMOTE) by creating synthetic data on minority data so that the data will be balanced with the majority data. This study aims to see the effect of SMOTE on sentiment analysis using the Support Vector Machine (SVM) algorithm. Based on the evaluation results using two different datasets, it was found that the results of the analysis using the SVM method resulted in an average accuracy of around 79.1% in the covid-19 dataset and 75% in the tv dataset. There was an increase in accuracy when applying the SVM + SMOTE method with an average accuracy of around 93.2% in the covid-19 dataset and 84,7% in the tv dataset.

Keywords: Imbalanced data, Support Vector Machine, Synthetic Minority Oversampling Technique (SMOTE), sentiment analysis.

## **ABSTRAK**

*Imbalanced data* merupakan permasalahan yang sering terjadi saat melakukan penelitian di bidang analisis sentimen. Permasalahan ini terjadi ketika *dataset* yang dianalisis memiliki kelas positif lebih banyak dari kelas negatif ataupun sebaliknya. Sehingga munculah istilah data mayoritas dan minoritas. Apabila data mayoritas lebih dominan maka proses klasifikasi cenderung menghasilkan klasifikasi kedalam kelas mayoritas. Sehingga diperlukan sebuah solusi untuk mengatasi permasalahan *imbalanced data*. Salah satu metode untuk mengatasi permasalahan *imbalanced data* dapat menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE akan menyelesaikan permasalahan tersebut dengan cara membuat data *synthetic* pada data minoritas sehingga data tersebut akan menjadi seimbang dengan data mayoritas. Penelitian ini akan melihat pengaruh SMOTE terhadap analisis sentimen menggunakan algoritma *Support Vector Machine* (SVM). Berdasarkan hasil evaluasi menggunakan dua *dataset* berbeda didapatkan bahwa hasil analisis menggunakan metode SVM menghasilkan akurasi rata-rata kisaran 79,1% pada *dataset* covid-19 dan 75% pada *dataset* tv. Lalu terjadi peningkatan akurasi ketika menerapkan metode SVM+SMOTE dengan akurasi rata-rata kisaran 93,2% pada *dataset* covid-19 dan 84,7% pada *dataset* tv.

Kata Kunci: *Imbalanced data*, *Support Vector Machine* (SVM), *Synthetic Minority Oversampling Technique* (SMOTE), analisis sentimen.

## **KATA PENGANTAR**

Puji syukur kepada Allah SWT atas segala rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi ini dengan baik. Skripsi ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan tugas akhir pendidikan program Strata-1 Program Studi Teknik Informatika pada Fakultas Ilmu Komputer, Universitas Sriwijaya.

Penulis ingin mengucapkan terima kasih kepada seluruh pihak yang telah memberikan bantuan dan dukungan baik dalam materil maupun moril selama proses pembuatan tugas akhir ini, yaitu :

1. Kedua orang tua penulis yang selalu memberikan dukungan dan doa kepada penulis setiap waktu agar dapat menyelesaikan semua kegiatan yang dilakukan dengan baik.
2. Kedua saudara penulis yang selalu memberikan semangat dan dukungan untuk menyelesaikan skripsi ini.
3. Bapak Jaidan Jauhari, M.T. selaku Dekan Fakultas Ilmu Komputer.
4. Ibu Alvi Syahrini Utami, M.Kom selaku Ketua Jurusan Teknik Informatika dan penguji saat sidang skripsi.
5. Ibu Novi Yusliani, M.T. dan Ibu Annisa Darmawahyuni, M.Kom. selaku pembimbing I dan pembimbing II yang telah memberikan arahan, bimbingan, kritik dan saran kepada penulis untuk menyelesaikan tugas akhir.
6. Bapak Muhammad Qurhanul Rizqie, M.T., Ph.D selaku pembimbing akademik yang telah memberikan bimbingan serta arahan dalam urusan akademik.

7. Seluruh Bapak dan Ibu Dosen Program Studi Teknik Informatika yang telah memberikan ilmu kepada penulis selama masa perkuliahan.
8. Staff Jurusan Teknik Informatika beserta staff Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi dan akademik selama masa perkuliahan.
9. Teman seperjuangan penulis, Muhammad Rizqi Assabil dan Ahmad Jailani, yang selalu memberikan semangat dan bantuan jika penulis terdapat kendala saat pengerjaan skripsi.
10. Sahabat penulis, Luthfi, Abdul Khoir, Saiful, Amos, Aldi, Josie, Adahwiya, Amalia, Amelia, dan seluruh teman kelas “Reguler C 2019” yang telah menemani penulis selama masa perkuliahan.
11. Semua pihak lainnya yang tidak dapat disebutkan satu persatu yang telah berperan bagi penulis dalam menyelesaikan tugas akhir, Terima Kasih banyak atas semuanya.
12. Semoga tugas akhir ini dapat bermanfaat bagi penulis khususnya maupun pembaca pada umumnya. Serta dapat menjadi referensi dan rujukan bagi hal-hal yang bermanfaat

Indralaya, 2 April 2023

Penulis,

Helmi Farizki

## DAFTAR ISI

<b>COVER .....</b>	i
<b>LEMBAR PENGESAHAN SKRIPSI .....</b>	ii
<b>TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI .....</b>	iii
<b>HALAMAN PERNYATAAN</b>	
Error! Bookmark not defined.	
<b>MOTTO .....</b>	v
<b>ABSTRACT .....</b>	vi
<b>ABSTRAK .....</b>	vii
<b>KATA PENGANTAR.....</b>	viii
<b>DAFTAR ISI.....</b>	x
<b>DAFTAR TABEL .....</b>	xiii
<b>DAFTAR GAMBAR.....</b>	xv
<b>BAB I PENDAHULUAN.....</b>	I-1
1.1 Pendahuluan .....	I-1
1.2 Latar Belakang Masalah .....	I-1
1.3 Rumusan Masalah .....	I-3
1.4 Tujuan Penelitian.....	I-4
1.5 Manfaat Penelitian.....	I-4
1.6 Batasan Masalah .....	I-5
1.7 Sistematika Penulisan.....	I-5
<b>BAB II KAJIAN LITERATUR .....</b>	II-1
2.1 Pendahuluan .....	II-1
2.2 Analisis Sentimen.....	II-1
2.3 Text Preprocessing .....	II-2
2.4 Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) .....	II-5
2.5 Support Vector Machine (SVM) .....	II-7
2.6 Synthetic Minority Oversampling Technique (SMOTE) .....	II-11
2.7 Confusion Matrix.....	II-13

2.8 RUP .....	II-15
2.9 Penelitian Lain yang Relevan.....	II-16
<b>BAB III METODOLOGI PENELITIAN .....</b>	<b>III-1</b>
3.1 Pendahuluan .....	III-1
3.2 Pengumpulan Data.....	III-1
3.3 Tahapan Penelitian .....	III-2
3.3.1 Kerangka Kerja .....	III-2
3.3.2 Menetapkan Kriteria Pengujian .....	III-6
3.3.3 Format Analisis Pengujian .....	III-6
3.3.4 Alat yang Digunakan dalam Pelaksanaan Penelitian.....	III-7
3.3.5 Pengujian Penelitian.....	III-7
3.3.6 Analisis Pengujian dan Membuat Kesimpulan .....	III-8
3.4 Metode Pengembangan Perangkat Lunak .....	III-8
3.4.1 Fase Insepsi (Inception) .....	III-9
3.4.2 Fase Elaborasi (Elaboration).....	III-9
3.4.3 Fase Konstruksi.....	III-10
3.4.4 Fase Transisi .....	III-10
3.5 Manajemen Proyek Penelitian.....	III-10
<b>BAB IV PENGEMBANGAN PERANGKAT LUNAK .....</b>	<b>IV-1</b>
4.1 Pendahuluan.....	IV-1
4.2 Fase Insepsi.....	IV-1
4.2.1 Pemodelan Bisnis .....	IV-1
4.2.2 Kebutuhan Sistem.....	IV-2
4.2.3 Analisis dan Desain .....	IV-3
4.3 Fase Elaborasi.....	IV-34
4.3.1 Pemodelan Bisnis .....	IV-34
4.3.2 Perancangan Data.....	IV-34
4.3.3 Perancangan <i>Interface</i> .....	IV-35
4.3.4 Kebutuhan Sistem .....	IV-38
4.3.5 Diagram Aktivitas .....	IV-38
4.3.6 Diagram Sequence .....	IV-41

4.4 Fase Konstruksi .....	IV-43
4.4.1 Diagram Kelas.....	IV-43
4.4.2 Implementasi.....	IV-44
4.5 Fase Transisi.....	IV-49
4.5.1 Pemodelan Bisnis.....	IV-49
4.5.2 Rencana Pengujian.....	IV-49
4.5.3 Implementasi.....	IV-50
4.6 Kesimpulan.....	IV-52
<b>BAB V HASIL DAN ANALISIS PENELITIAN.....</b>	<b>V-1</b>
5.1 Pendahuluan .....	V-1
5.2 Data Hasil Penelitian .....	V-1
5.2.1 Konfigurasi Percobaan.....	V-1
5.2.2 Hasil Pengujian <i>Dataset Sentimen Covid-19</i> .....	V-2
5.2.3 Hasil Pengujian <i>Dataset Sentimen Tayangan TV</i> .....	V-7
5.3 Analisis Hasil Pengujian Secara Keseluruhan.....	V-13
5.4 Kesimpulan.....	V-17
<b>BAB VI KESIMPULAN DAN SARAN.....</b>	<b>VI-1</b>
6.1 Pendahuluan .....	VI-1
6.2 Kesimpulan.....	VI-1
6.3 Saran .....	VI-2
<b>DAFTAR PUSTAKA .....</b>	<b>xi</b>

## DAFTAR TABEL

Tabel II-1. Contoh Proses <i>Cleansing</i> (Tuhuteru, 2020) .....	II-2
Tabel II-2. Contoh Proses <i>Case Folding</i> (Tuhuteru, 2020) .....	II-3
Tabel II- 3. Contoh proses <i>Tokenizing</i> (Tuhuteru, 2020).....	II-3
Tabel II- 4. Contoh proses <i>normalization</i> (Tuhuteru, 2020).....	II-4
Tabel II- 5. Contoh proses <i>Stopword Removal</i> (Tuhuteru, 2020).....	II-5
Tabel II- 6. Contoh proses <i>Stemming</i> (Tuhuteru, 2020).....	II-5
Tabel II- 7. <i>Confusion Matrix</i> Untuk Dua Kelas.....	II-13
Tabel III- 1. Format Analisis Pengujian.....	III-6
Tabel III- 2. Format Hasil Pengujian .....	III-8
Tabel III- 3. Jadwal Rancangan Penelitian .....	III-1
Tabel IV-1. Kebutuhan Fungsional.....	IV-3
Tabel IV-2. Kebutuhan Non-Fungsional.....	IV-3
Tabel IV-3. Contoh Data Sampel Tweet.....	IV-5
Tabel IV-4. Hasil Proses <i>Case Folding</i> .....	IV-6
Tabel IV-5. Hasil Proses <i>Cleansing</i> .....	IV-7
Tabel IV-6. Hasil Proses <i>Tokenizing</i> .....	IV-8
Tabel IV-7. Hasil Proses Normalisasi.....	IV-9
Tabel IV-8. Hasil Proses <i>Stopword</i> .....	IV-9
Tabel IV-9. Hasil Proses <i>Stemming</i> .....	IV-10
Tabel IV-10. Hasil Perhitungan Proses TF dan IDF .....	IV-12
Tabel IV-11. Tabel Hasil Akhir TF-IDF .....	IV-14
Tabel IV-12. Hasil Proses Pembentukan K-NN Data Minoritas .....	IV-17
Tabel IV-13. Hasil Proses Pembentukan Data <i>synthetic</i> .....	IV-20
Tabel IV-14. Hasil Proses Nilai <i>Dot Product</i> .....	IV-23
Tabel IV-15. Hasil Klasifikasi D11 Menggunakan SVM dan SMOTE.....	IV-26
Tabel IV-16. Tabel Definisi Aktor.....	IV-28
Tabel IV-17. Tabel Definisi <i>Use Case</i> .....	IV-28
Tabel IV-18. Skenario Menginput Dataset Sentimen .....	IV-29

Tabel IV-19. Skenario Menampilkan Dashboard Hasil Analisis .....	IV-
<b>Error! Bookmark not defined.</b>	
Tabel IV-20. Skenario Menampilkan Tabel Hasil Klasifikasi SVM .....	IV-31
Tabel IV- 21. Skenario Menampilkan Tabel SVM SMOTE .....	IV-33
Tabel IV- 22. Perancangan Data .....	IV-35
Tabel IV-23. Deskripsi Isi Kelas.....	IV-44
Tabel IV-24. Daftar File Program <i>Interface</i> .....	IV-46
Tabel IV-25. Rencana Pengujian <i>Use Case</i> Menginput <i>Dataset</i> Sentimen... IV-49	
Tabel IV-26. Rencan Pengujian <i>Use Case</i> Analisis Akurasi Klasifikasi SVM Tanpa SMOTE Dan SVM Dengan SMOTE.....	IV-49
Tabel IV-27. Rencana Pengujian <i>Use Case</i> Tabel Hasil Analisis SVM Tanpa SMOTE .....	IV-48
Tabel IV-28. Rencana Pengujian <i>Use Case</i> Tabel Hasil Analisis SVM Dengan Kombinasi SMOTE.....	IV-50
Tabel IV- 29. Pengujian <i>Use Case</i> Menginput <i>Dataset</i> Sentimen.....	IV-50
Tabel IV-30. Pengujian <i>Use Case</i> Analisis Akurasi Klasifikasi SVM Tanpa SMOTE Dan SVM Dengan SMOTE.....	IV-51
Tabel IV-31. Pengujian <i>Use Case</i> Tabel Klasifikasi SVM Tanpa SMOTE ..IV-51	
Tabel IV- 32. Pengujian <i>Use Case</i> Tabel Klasifikasi SVM Dengan Kombinasi SMOTE .....	IV-52
Tabel V-1. <i>Confusion Matrix</i> SVM Tanpa SMOTE <i>Dataset Covid</i> .....	V-2
Tabel V-2. Hasil Evaluasi Klasifikasi SVM Tanpa SMOTE <i>Dataset Covid</i> ... V-3	
Tabel V-3. <i>Confusion Matrix</i> SVM Kombinasi SMOTE <i>Dataset Covid</i> .....	V-4
Tabel V-4. Hasil Evaluasi Klasifikasi SVM Dengan Kombinasi SMOTE <i>Dataset</i> <i>Covid</i> .....	V-4
Tabel V-5. <i>Confusion Matrix</i> SVM Tanpa SMOTE <i>Dataset TV</i> .....	V-8
Tabel V-6. Hasil Evaluasi Klasifikasi SVM Tanpa SMOTE <i>Dataset TV</i> ..... V-8	
Tabel V- 7. <i>Confusion Matric</i> SVM Kombinasi SMOTE <i>Dataset TV</i> .....	V-9
Tabel V-8. Hasil Evaluasi Klasifikasi SVM Dengan Kombinasi SMOTE <i>Dataset</i> <i>TV</i> .....	V-10
Tabel V- 9. Perbandingan Waktu Komputasi .....	V-16

## DAFTAR GAMBAR

Gambar II- 1. Contoh <i>hyperplane</i> pada SVM.....	II-7
Gambar II- 2. Penerapan Arsitektur RUP (Krutchen, 2000).....	II-15
Gambar III- 1. Kerangka Kerja Penelitian .....	III-2
Gambar III- 2. <i>Flowchart</i> Tahapan <i>Text Preprocessing</i> .....	III-4
Gambar IV-1. Gambar <i>Use Case Diagram</i> .....	IV-27
Gambar IV-2. Rancangan Halaman <i>Home Input Dataset</i> .....	IV-36
Gambar IV-3. Rancangan Halaman Dashboard Hasil Analisis .....	IV-36
Gambar IV-4. Rancangan Halaman Tabel Analisis SVM .....	IV-37
Gambar IV-5. Rancangan Halaman Tabel Analisis SVM+SMOTE .....	IV-37
Gambar IV-6. Diagram Aktivitas <i>input Dataset</i> .....	IV-39
Gambar IV-7. Diagram Aktivitas Dashboard Hasil Analisis Sentimen.....	IV-39
Gambar IV-8. Diagram Aktivitas Tabel Hasil Klasifikasi SVM .....	IV-40
Gambar IV-9. Diagram Aktivitas Hasil Tabel Klasifikasi SVM Kombinasi SMOTE .....	IV-40
Gambar IV-10. Diagram Sequence Input Dataset.....	IV-41
Gambar IV-11. Diagram <i>Squence</i> Dashboard Hasil Analisis .....	IV-42
Gambar IV-12. Diagram <i>Squence</i> Tabel Hasil Analisis SVM.....	IV-42
Gambar IV-13. Diagram <i>Squence</i> Tabel Analisis SVM Kombinasi SMOTE .....	IV-43
Gambar IV-14. Diagram Kelas .....	IV-44
Gambar IV-15. <i>Interface Home Input Dataset</i> .....	IV-47
Gambar IV-16. <i>Interface Dashboard</i> Hasil Analisis Klasifikasi SVM Tanpa SMOTE Dan SVM Dengan SMOTE.....	IV-47
Gambar IV-17. <i>Interface</i> Tabel Label Dari Klasifikasi SVM Tanpa SMOTE .....	IV-48
Gambar IV-18. <i>Interface</i> Tabel Label Dari Klasifikasi SVM Dengan Kombinasi SMOTE .....	IV-48
Gambar V-1. Grafik Perbandingan Hasil <i>Accuracy</i> Pada <i>Dataset Covid</i> .....	V-5

Gambar V- 2. Grafik Perbandingan Hasil <i>Precision</i> Pada <i>Dataset Covid</i> ....	V-6
Gambar V-3. Grafik Perbandingan Hasil <i>Recall</i> Pada <i>Dataset Covid</i> .....	V-6
Gambar V-4. Grafik Perbandingan Hasil <i>F-Measure</i> Pada <i>Dataset Covid</i> ..	V-7
Gambar V- 5. Grafik Perbandingan Hasil <i>Accuracy</i> Pada <i>Dataset TV</i> .....	V-11
Gambar V- 6. Grafik Perbandingan Hasil <i>Precision</i> Pada <i>Dataset TV</i> .....	V-11
Gambar V- 7. Grafik Perbandingan Hasil <i>Recall</i> Pada <i>Dataset TV</i> .....	V-12
Gambar V- 8. Grafik Perbandingan Hasil <i>F-Measure</i> Pada <i>Dataset TV</i> .....	V-12
Gambar V- 9. Grafik Hasil Perbandingan <i>Accuracy</i> Seluruh Dataset .....	V-13
Gambar V- 10. Grafik Hasil Perbandingan <i>Precision</i> Seluruh Dataset .....	V-14
Gambar V- 11. Grafik Hasil Perbandingan <i>Precision</i> Seluruh Dataset .....	V-14
Gambar V- 12. Grafik Hasil Perbandingan <i>F-Measure</i> Seluruh Dataset.....	V-15

# **BAB I**

## **PENDAHULUAN**

### **1.1 Pendahuluan**

Bab ini akan membahas latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan batasan masalah. Bagian ini juga memuat penjelasan mengenai gambaran umum dari keseluruhan kegiatan penelitian yang dilakukan terhadap pengaruh metode *Synthetic Minority Oversampling Technique* (SMOTE) terhadap analisis sentimen dengan menggunakan metode algoritma *Support Vector Machine* (SVM).

### **1.2 Latar Belakang Masalah**

Analisis sentimen merupakan metode yang digunakan untuk mengkategorikan bahasa tertulis apakah bernilai kategori positif atau negatif (Alsaeedi & Khan, 2019). Melalui analisis sentimen menggunakan algoritma *machine learning* maka dapat menganalisis kategori *tweet* apakah termasuk dalam kategori positif atau negatif dalam jumlah besar yang mana sulit dilakukan oleh manusia secara manual.

Banyak algoritma *machine learning* yang diimplementasikan untuk analisis sentimen mendapatkan akurasi baik seperti algoritma *Support Vector Machine* (SVM), *Naive Bayes*, dan *Logistic Regression* (Satriaji & Kusumaningrum, 2018). Algoritma SVM menerapkan metode *hyperplane* yang berfungsi untuk

memisahkan kelas sentimen positif atau negatif (Fitri, 2020). Penerapan metode ini membuat klasifikasi SVM banyak digunakan untuk penelitian analisis sentimen.

Berdasarkan penelitian dari Ratino et al.(2020) pada analisis sentimen informasi *covid-19* menggunakan algoritma SVM dan *Naive Bayes*, didapatkan hasil pengujian bahwa metode SVM mendapatkan akurasi lebih baik yaitu 80,23% dibandingkan dengan akurasi *Naive Bayes* sebesar 78,02%.

Dalam melakukan analisis sentimen munculah permasalahan mengenai data yang dianalisis tidak seimbang (*imbalanced data*). *Imbalanced data* disebabkan karena *dataset* yang digunakan salah satu kelasnya lebih banyak ke positif atau sebaliknya (Satriaji & Kusumaningrum, 2018) sehingga munculah istilah data mayoritas dan minoritas. Permasalahan *imbalanced data* menjadi masalah utama karena ketidakseimbangan ini menyebabkan dalam proses klasifikasi cenderung menghasilkan klasifikasi kedalam kelas mayoritas (Johnson & Khoshgoftaar, 2019). Hal ini dikarenakan peningkatan probabilitas ke kelas mayoritas menyebabkan kelas minoritas cenderung terjadi kesalahan saat diklasifikasikan.

Permasalahan tersebut dapat diatasi dengan menggunakan metode *resample data*. Metode *resample data* yang sering digunakan adalah metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE adalah sebuah metode yang menggunakan pendekatan *oversampling* dengan cara mensintesis kelas data minor sehingga data tersebut akan menjadi seimbang antara kelas minoritas dan mayoritas (Chawla et al., 2002).

Beberapa penelitian sebelumnya yang menerapkan metode SMOTE menggunakan data tidak seimbang menunjukkan hasil akurasi yang cenderung

meningkat dibandingkan tanpa menerapkan metode SMOTE. Berdasarkan penelitian dari Flores et al. (2018) mengenai penerapan kombinasi metode SVM dan SMOTE. Didapatkan hasil SVM tanpa SMOTE sebesar 80,87% sedangkan SVM yang dikombinasikan dengan SMOTE sebesar 83,16%. Lalu juga berdasarkan penelitian dari Putri dan Kharisudin (2022) mengenai analisis sentimen marketplace Tokopedia menggunakan algoritma SVM dan *Naive Bayes* yang dikombinasikan dengan SMOTE. Didapatkan hasil SVM tanpa SMOTE sebesar 90,88% sedangkan SVM + SMOTE sebesar 91,84%. Lalu hasil akurasi dari *Naive Bayes* tanpa SMOTE sebesar 83,52% sedangkan *Naive Bayes* + SMOTE sebesar 85,76%.

Berdasarkan uraian penelitian diatas, maka penelitian ini akan melakukan pengujian untuk melihat pengaruh metode *Synthetic Minority Oversampling Technique* (SMOTE) terhadap analisis sentimen dengan menggunakan metode algoritma *Support Vector Machine* (SVM).

### 1.3 Rumusan Masalah

Berdasarkan uraian dari latar belakang sebelumnya maka rumusan masalah dari penelitian ini adalah pengaruh *Synthetic Minority Oversampling Technique* (SMOTE) terhadap sentimen analisis menggunakan algoritma *Support Vector Machine* (SVM). Sehingga dari penelitian ini terdapat beberapa RQ (*Research Question*) sebagai berikut:

1. Bagaimana mengimplementasikan metode SVM untuk mengklasifikasi analisis sentimen?

2. Bagaimana mengimplementasikan metode SVM dan metode SMOTE untuk mengklasifikasi analisis sentimen?
3. Bagaimana kinerja metode SVM pada analisis sentimen?
4. Bagaimana kinerja metode SVM yang dikombinasikan dengan SMOTE pada analisis sentimen?

#### **1.4 Tujuan Penelitian**

Berikut tujuan dari penelitian ini:

1. Menghasilkan sistem analisis sentimen menggunakan metode SVM untuk analisis sentimen.
2. Menghasilkan sistem analisis sentimen menggunakan metode SVM dan metode SMOTE untuk mengatasi *imbalanced data* pada analisis sentimen.
3. Mengetahui kinerja metode SVM pada analisis sentimen.
4. Mengetahui kinerja metode SVM yang dikombinasikan dengan SMOTE pada analisis sentimen.

#### **1.5 Manfaat Penelitian**

Berikut manfaat dari penelitian ini:

1. Memahami mekanisme metode SVM yang dikombinasikan dengan metode SMOTE untuk mengatasi *imbalance data*.
2. Hasil laporan penelitian dapat menjadi sumber referensi bagi peneliti lain terkait penanganan *imbalanced data* menggunakan metode SMOTE.

## 1.6 Batasan Masalah

Berikut batasan masalah dari penelitian ini:

1. *Dataset* analisis sentimen yang digunakan berasal dari opini *tweet* pengguna *twitter*.
2. *Dataset* analisis sentimen yang digunakan berbahasa Indonesia.
3. *Dataset* yang digunakan berformat csv.
4. *Dataset* analisis sentimen yang digunakan terdiri dari dua kelas yaitu positif dan negatif.
5. *Kernel* yang digunakan pada metode SVM yaitu *linear*.
6. Metode pemberian bobot kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*).
7. Metode evaluasi pengujian menggunakan *confusion matrix*.

## 1.7 Sistematika Penulisan

Berikut sistematika penulisan pada penelitian ini:

### **BAB I. PENDAHULUAN**

Bab ini memaparkan penjelasan tentang latar belakang penelitian, rumusan masalah penelitian, tujuan penelitian, manfaat penelitian, batasan masalah, serta sistematika penulisan penelitian.

### **BAB II. KAJIAN LITERATUR**

Bab ini memaparkan penjelasan tentang teori-teori dasar seperti analisis sentimen, *text preprocessing*, SVM, dan SMOTE

serta memaparkan penelitian-penelitian terdahulu yang relevan dengan topik penelitian ini.

### **BAB III. METODELOGI PENELITIAN**

Bab ini memaparkan penjelasan tentang tahapan yang dikerjakan pada penelitian ini. Tahapan yang dijelaskan disusun secara rinci berdasarkan kerangka kerja.

### **BAB IV. PENGEMBANGAN PERANGKAT LUNAK**

Bab ini memaparkan proses pengembangan perangkat lunak analisis sentimen menggunakan algoritma SVM tanpa SMOTE dan SVM dengan SMOTE. Tahapan proses disusun berdasarkan kerangka metode *Rational Unified Porcess* (RUP).

### **BAB V. HASIL DAN ANALISIS PENELITIAN**

Bab ini memaparkan hasil dari pengembangan perangkat lunak yang telah dibuat serta memaparkan hasil dari pengujian pemodelan yang sudah dianalisis.

### **BAB VI. KESIMPULAN DAN SARAN**

Bab ini memaparkan kesimpulan dari hasil penelitian serta saran untuk pengembangan penelitian selanjutnya.

## **DAFTAR PUSTAKA**

- Alsaeedi, A. and Khan, M.Z., 2019. A study on sentiment analysis techniques of Twitter data. *International Journal of Advanced Computer Science and Applications*, 10(2).
- Bordoloi, M. and Biswas, S.K., 2018. Sentiment analysis of product using machine learning technique: a comparison among NB, SVM and MaxEnt. *International Journal of Pure and Applied Mathematics*, 118(19), pp.71-83.
- Buntoro, G.A., 2017. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, 2(1).
- Chamzah, S.M., Lestandy, M., Kasan, N. and Nugraha, A., 2022. Penerapan Synthetic Minority Oversampling Technique (SMOTE) untuk Imbalance Class pada Data Text Menggunakan kNN. *Syntax: Jurnal Informatika*, 11(02), pp.56-67.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp.321-357.
- Ferdiana, R., Jatmiko, F., Purwanti, D.D., Ayu, A.S.T. and Dicka, W.F., 2019. Dataset Indonesia untuk analisis sentimen. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 8(4), pp.334-339.

Fitri, E., 2020. Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. *Jurnal Transformatika*, 18(1), pp.71-80.

Fitriyani, N.K. and Hartanto, A.D., 2020. Analisis Sentimen Terhadap Tokoh Publik Menggunakan Support Vector Machine. *MEANS (Media Informasi Analisa dan Sistem)*, pp.8-12.

Flores, A.C., Icoy, R.I., Pena, C.F. and Gorro, K.D., 2018. An evaluation of svm and naive bayes with smote on sentiment analysis data set. In *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)* (pp. 1-4). IEEE.

Johnson, J.M. and Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), pp.1-54.

Kelvin, K., Banjarnahor, J., Nababan, M.N. and Sinurat, S.H., 2022. Analisis perbandingan sentimen Corona Virus Disease-2019 (Covid19) pada Twitter Menggunakan Metode Logistic Regression Dan Support Vector Machine (SVM). *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, 5(2), pp.47-52.

Nuraliza, H., Pratiwi, O.N. and Hamami, F., 2022. Analisis Sentimen IMBd Film Review Dataset Menggunakan Support Vector Machine (SVM) dan Seleksi Feature Importance. *Jurnal Mirai Management*, 7(1), pp.1-17.

Prastyo, P.H., Sumi, A.S., Dian, A.W. and Permanasari, A.E., 2020. Tweets responding to the Indonesian Government's handling of COVID-19: Sentiment analysis using SVM with normalized poly kernel. *Journal of Information Systems Engineering and Business Intelligence*, 6(2), pp.112-122.

Putri, M.I. and Kharisudin, I., 2022. Analisis Sentimen Pengguna Aplikasi Marketplace Tokopedia Pada Situs Google Play Menggunakan Metode Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 5, pp. 759-766).

Rajvanshi, N. and Chowdhary, K.R., 2017. Comparison of svm and naive bayes text classification algorithms using weka. *International Journal of Engineering Research and, 6(09)*.

Ratino, R.R., Hafidz, N.H.H., Anggraeni, S.A.A. and Gata, W.G.G., 2020. Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *JUPITER (Jurnal Penelitian Ilmu dan Teknologi Komputer)*, 12(2), pp.1-11.

Rofiqoh, U., Perdana, R.S. and Fauzi, M.A., 2017. Analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler indonesia pada twitter dengan metode Support Vector Machine dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, 2548, p.964X.

- Sari, F.V. and Wibowo, A., 2019. Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 10(2), pp.681-686.
- Satriaji, W. and Kusumaningrum, R., 2018. Effect of synthetic minority oversampling technique (smote), feature representation, and classification algorithm on imbalanced sentiment analysis. In *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 1-5). IEEE.
- Setiawan, H. and Khairuzzaman, M.Q., 2017. Perancangan Sistem Informasi Manajemen Proyek: Sistem Informasi Kontraktor. *Jurnal Khatulistiwa Informatika*, 5(2).
- Sridhya, K. and Sowjanya, A.M., 2019. Aspect based sentiment analysis using POS tagging and TFIDF. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8.
- Taufik, I., 2018. *Analisis sentimen terhadap tokoh publik menggunakan Algoritma Support Vector Machine (SVM)* (Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta).
- Tuhuteru, H. and Iriani, A., 2018. Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier. *Jurnal Informatika*, 3(03).

Tuhuteru, H., 2020. Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine. *Journal Information System Development (ISD)*, 5(2).

Pakdeetrakulwong, Udsanee. "Semantic Web-based Approach to Support Rational Unified Process Software Development." In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, pp. 432-438. IEEE, 2018.

Widodo, B.K., Matondang, N.H. and Prasvita, D.S., 2022. Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Penggunaan Aplikasi Jobstreet. *Techno. Com*, 21(3), pp.523-533.

Zheng, Z., Cai, Y. and Li, Y., 2015. Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5), pp.1017-1037.