

**IMPLEMENTASI ALGORITMA SMOTE DAN UMAP
SEBAGAI UPAYA PENANGGULANGAN IMBALANCE
HIGH DIMENSIONAL DATASETS DALAM
ANALISA SENTIMEN**

SKRIPSI

**Diajukan sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer (S.Kom.)**



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

Carissa Komalasari

00000019971

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG**

2021

LEMBAR PENGESAHAN

IMPLEMENTASI ALGORITMA SMOTE DAN UMAP SEBAGAI UPAYA PENANGGULANGAN IMBALANCE HIGH DIMENSIONAL DATASETS DALAM ANALISA SENTIMEN

oleh

Nama : Carissa Komalasari
NIM : 000000199971
Program Studi : Informatika
Fakultas : Teknik dan Informatika

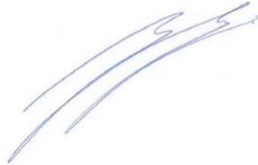
Tangerang, 26 Januari 2021

Ketua Sidang



Moeljono Widjaja, M.Sc., Ph.D.

Dosen Penguji



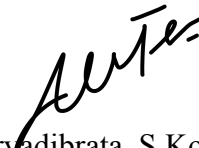
Seng Hansun, S.Si., M.Cs.

Dosen Pembimbing I



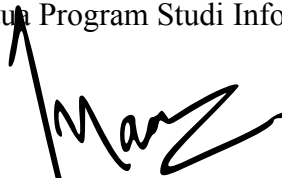
Julio Christian Young, M.Kom.

Dosen Pembimbing II



Alethea Suryadibrata, S.Kom., M.Eng.

Mengetahui,
Ketua Program Studi Informatika,



Marlinda Vasty Overbeek, M.Kom.

PERNYATAAN TIDAK MELAKUKAN PLAGIAT

Dengan ini saya:

Nama : Carissa Komalasari

NIM : 00000019971


Program Studi : Informatika

Fakultas : Teknik dan Informatika

menyatakan bahwa Skripsi yang berjudul “**Implementasi Algoritma SMOTE dan UMAP sebagai Upaya Penanggulangan Imbalance High Dimensional Datasets dalam Analisa Sentimen**” ini adalah karya ilmiah saya sendiri, bukan plagiat dari karya ilmiah yang ditulis oleh orang lain atau lembaga lain, dan semua karya ilmiah orang lain atau lembaga lain yang dirujuk dalam Skripsi ini telah disebutkan sumber kutipannya serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah Skripsi yang telah saya tempuh.

Tangerang, 29 Desember 2020



Carissa Komalasari

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Multimedia Nusantara, saya yang bertanda tangan di bawah ini:

Nama : Carissa Komalasari

NIM : 00000019971

Program Studi : Informatika

Fakultas : Teknik dan Informatika

Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui dan memberikan izin kepada **Universitas Multimedia Nusantara** hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

**Implementasi Algoritma SMOTE dan UMAP
sebagai Upaya Penanggulangan Imbalance High
Dimensional Datasets dalam
Analisa Sentimen**

beserta perangkat yang diperlukan.

Dengan Hak Bebas Royalti Non-eksklusif ini, pihak **Universitas Multimedia Nusantara** berhak menyimpan, mengalihmedia atau *format*-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mendistribusi dan menampilkan atau mempublikasikan karya ilmiah saya di internet atau media lain untuk kepentingan akademis, tanpa perlu meminta izin dari saya maupun memberikan

royalty kepada saya, selama tetap mencantumkan nama saya sebagai penulis karya ilmiah tersebut.

Demikian pernyataan ini saya buat dengan sebenarnya untuk dipergunakan sebagaimana mestinya.

Tangerang, 29 Desember 2020

A handwritten signature in black ink, appearing to be 'Carissa Komalasari', written in a cursive style.

Carissa Komalasari

HALAMAN PERSEMBAHAN / MOTO

*I am a great believer in luck,
and I find the harder I work, the more I have of it.*

— *Thomas Jefferson (1743-1826)*

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas rahmat dan penyertaan-Nya, sehingga laporan Skripsi ini dapat diselesaikan tepat waktu. Skripsi dengan judul: Implementasi Algoritma SMOTE dan UMAP sebagai Upaya Penanggulangan Imbalance High Dimensional Datasets dalam Analisa Sentimen dibuat untuk memenuhi salah satu syarat memperoleh gelar sarjana strata satu Jurusan Informatika Universitas Multimedia Nusantara.

Proses penyusunan laporan tidak terlepas dari bantuan serta dukungan berbagai pihak. Oleh sebab itu, penulis mengucapkan terima kasih kepada:

1. Dr. Ninok Leksono MA selaku Rektor Universitas Multimedia Nusantara yang memberi inspirasi bagi penulis untuk berprestasi,
2. Dr. Eng. Niki Prastomo S.T., M.Sc., Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara,
3. Ibu Marlinda Vasty Overbeek, M.Kom. selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara, yang menerima penulis dengan baik untuk berkonsultasi,
4. Bapak Julio Christian Young, M.Kom., dan Ibu Alethea Suryadibrata, S.Kom., M.Eng. yang telah membimbing pembuatan Skripsi dengan sabar, memberikan ilmu baru, dan mengajarkan penulis tata cara menulis karya ilmiah dengan benar,
5. Orang tua serta keluarga penulis, terutama kakak dan adik yang telah menemani, memberikan dukungan, bantuan, dan semangat mulai dari awal masa perkuliahan hingga tersusunnya laporan Skripsi ini ditengah masa pandemi.

6. Teman-teman terdekat, khususnya grup Jacemart yang terdiri dari Theniarti Ailin, Emily Wiputri, dan Jeanne Lukman, serta Gabriella Valentine atas bantuan, semangat, dan hiburan di kala penulis merasa kesulitan menyelesaikan laporan Skripsi ini.

Semoga Skripsi ini dapat bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi bagi para pembaca.

Tangerang, 29 Desember 2020

A handwritten signature in black ink, appearing to be 'Carissa', with a stylized flourish extending to the right.

Carissa Komalasari

IMPLEMENTASI ALGORITMA SMOTE DAN UMAP SEBAGAI UPAYA PENANGGULANGAN IMBALANCE HIGH DIMENSIONAL DATASETS DALAM ANALISA SENTIMEN

ABSTRAK

Class imbalance problem merupakan permasalahan yang sering dihadapi dalam *data mining*. Permasalahan ini pada umumnya diatasi dengan melakukan *oversampling* menggunakan Synthetic Minority Oversampling Technique (SMOTE), namun implementasinya terhadap data berdimensi tinggi terbukti menghasilkan akurasi yang lebih rendah dibandingkan *random undersampling*. Untuk mengatasi masalah ini, terdapat penelitian yang membuktikan melakukan reduksi dimensi dengan PCA sebelum menggunakan SMOTE dapat memberikan performa yang lebih baik untuk mengklasifikasi gambar. Terdapat beberapa algoritma untuk melakukan reduksi dimensi, salah satunya Uniform Manifold Approximation and Projection (UMAP). UMAP memiliki beberapa kelebihan dibandingkan algoritma lain seperti t-SNE dan PCA, yakni merepresentasikan struktur topological dan memperhitungkan struktur data global lebih baik. Berdasarkan fakta dan permasalahan tersebut, penelitian ini bertujuan untuk membandingkan performa SMOTE dengan kombinasi UMAP dan SMOTE dalam mengatasi permasalahan *class imbalance problem* pada *text features* berdimensi tinggi. Percobaan yang dilakukan berbentuk analisa sentimen dengan *sentence embedding* menggunakan *pretrained* Embedding from Language Model (ELMo) dan klasifikasi dengan Multilayer Perceptron (MLP). UMAP diimplementasikan untuk mereduksi dimensi dataset, sehingga secara teoritis performa SMOTE dapat meningkat. Hasil dari *resampling* selanjutnya dikembalikan lagi ke dimensi awal untuk melakukan analisa sentimen. Hasil dari penelitian yang dilakukan menunjukkan UMAP-SMOTE menurunkan rata-rata f-measure SMOTE dengan persentase minimum 27%.

Kata kunci: UMAP, SMOTE, *class imbalance problem*, ELMo, MLP, analisa sentimen.

IMPLEMENTATION OF SMOTE AND UMAP FOR HANDLING HIGH-DIMENSIONAL CLASS IMBALANCED DATASETS IN SENTIMENT ANALYSIS

ABSTRACT

Class imbalance problem is a common problem in data mining. In general, this problem is solved by oversampling using Synthetic Minority Oversampling Technique (SMOTE). However, its implementation towards high-dimensional data is proven to yield lower accuracy compared to random under-sampling. To solve this problem, a study demonstrates reducing the dataset dimension with PCA before using SMOTE could obtain better image classification performance. There are plenty of algorithms to perform dimensional reduction, one of which is the Uniform Manifold Approximation and Projection (UMAP). UMAP has several advantages compared to other algorithms such as t-SNE and PCA, namely that it represents the topological structure of datasets and preserves the global data structure better. Therefore, this study aims to compare the performance of SMOTE to the combination of UMAP and SMOTE in overcoming class imbalance problems in high dimensional text features. The experiment was carried out through sentiment analysis with sentence embedding using pre-trained Embedding from Language Model (ELMo) and classification with Multilayer Perceptron (MLP). UMAP is implemented to reduce the dimensions of the dataset to increase the SMOTE performance theoretically. The resampling result will be returned to its original dimensions afterward to perform sentiment analysis. The research results show that UMAP-SMOTE decreases the f-measure average of SMOTE by a minimum percentage of 27%.

Keywords: UMAP, SMOTE, *class imbalance problem*, ELMo, MLP, sentiment analysis.

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
HALAMAN PERNYATAAN TIDAK MELAKUKAN PLAGIAT.....	iii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN / MOTO.....	vi
KATA PENGANTAR	vii
ABSTRAK.....	ix
ABSTRACT.....	x
DAFTAR ISI.....	xi
DAFTAR GAMBAR	xiii
DAFTAR TABEL.....	xiv
DAFTAR RUMUS	xv
DAFTAR LAMPIRAN.....	xvi
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah.....	1
1.2 Rumusan Masalah	6
1.3 Batasan Masalah.....	6
1.4 Tujuan Penelitian.....	7
1.5 Manfaat Penelitian.....	7
1.6 Sistematika Penulisan.....	7
BAB 2 LANDASAN TEORI.....	9
2.1 Representasi Teks.....	9
2.2 Embedding from Language Model (ELMo)	9
2.3 Uniform Manifold Approximation and Projection (UMAP)	12
2.4 Synthetic Minority Over-sampling Technique (SMOTE)	14
2.5 Multilayer Perceptron (MLP).....	16
2.6 Principal Component Analysis (PCA)	18
2.7 Grid Search.....	20
2.8 Metrik Evaluasi	21
BAB 3 METODOLOGI PENELITIAN.....	24
3.1 Metodologi Penelitian	24
3.2 Perancangan Aplikasi	25
3.2.1 Flowchart Utama	26
3.2.2 Flowchart Sentence Embedding.....	27
3.2.3 Flowchart Reduksi Dimensi dengan PCA.....	29
3.2.4 Flowchart Dataset Tidak Seimbang	30
3.2.5 Flowchart Analisa Sentimen dengan MLP	31
3.2.6 Flowchart Analisa Sentimen dengan SMOTE dan MLP	32
3.2.7 Flowchart Analisa Sentimen dengan UMAP, SMOTE, dan MLP.....	33
BAB 4 HASIL DAN DISKUSI	35
4.1 Spesifikasi Perangkat	35
4.1.1 Perangkat Pertama.....	35
4.1.2 Perangkat Kedua	36
4.2 Implementasi Algoritma.....	36

4.2.1	Potongan Kode Sentence Embedding dengan ELMo	37
4.2.2	Potongan Kode Reduksi Dimensi dengan PCA	39
4.2.3	Potongan Kode Analisa Sentimen.....	41
4.3	Skenario Pengujian.....	44
4.3.1	Pengujian Dataset Hasil Embedding Ketiga Layer ELMo.....	45
4.3.2	Pengujian Dataset Hasil Embedding Layer <i>Semantic</i> ELMo.....	47
4.4	Hasil Pengujian dan Visualisasi	47
4.4.1	Hasil Pengujian Dataset Hasil Embedding Ketiga Layer ELMo	48
4.4.2	Hasil Pengujian Dataset Embedding Layer <i>Semantic</i> ELMo	53
4.4.3	Visualisasi Dataset	55
4.5	Evaluasi Hasil Pengujian.....	61
BAB 5 SIMPULAN DAN SARAN		65
5.1	Simpulan.....	65
5.2	Saran.....	66
DAFTAR PUSTAKA		68

DAFTAR GAMBAR

Gambar 2.1 Transformasi untuk Setiap Token	10
Gambar 2.2 Bidirectional Language Model	11
Gambar 2.3 Perhitungan ELMo untuk Setiap Layer	12
Gambar 2.4 <i>Simplices</i> dalam Dimensi Rendah	13
Gambar 2.5 Pembuatan Data Sintesis SMOTE	15
Gambar 2.6 Single Layer Perceptron	16
Gambar 2.7 Multilayer Perceptron	17
Gambar 2.8 Visualisasi Pencarian <i>Principal Components</i> dengan PCA	19
Gambar 2.9 Alur proses Grid Search	21
Gambar 3.1 <i>Flowchart</i> Utama	26
Gambar 3.2 <i>Flowchart Sentence Embedding</i>	28
Gambar 3.3 <i>Flowchart</i> Reduksi Dimensi dengan PCA	29
Gambar 3.4 <i>Flowchart</i> Dataset Tidak Seimbang	30
Gambar 3.5 <i>Flowchart</i> Analisa Sentimen dengan MLP	32
Gambar 3.6 <i>Flowchart</i> Analisa Sentimen dengan SMOTE dan MLP	33
Gambar 3.7 <i>Flowchart</i> Analisa Sentimen dengan UMAP, SMOTE, dan MLP ...	34
Gambar 4.1 Potongan Kode Fungsi <i>Sentence Embedding</i>	37
Gambar 4.2 Potongan Kode <i>Semantic Sentence Embedding</i>	38
Gambar 4.3 Potongan Kode <i>Sentence Embedding</i> Dalam <i>Batch</i>	39
Gambar 4.4 Potongan Kode Pencarian Jumlah Komponen PCA	40
Gambar 4.5 Grafik <i>Explained Variance Principal Components</i>	40
Gambar 4.6 Potongan kode Reduksi Dimensi PCA	41
Gambar 4.7 Potongan Kode Fungsi Penentuan Model	42
Gambar 4.8 Potongan Kode Analisa Sentimen	42
Gambar 4.9 Potongan Kode Kelas UMAP SMOTE	44
Gambar 4.10 Pesan <i>Warning</i> UMAP dengan Jumlah Komponen Besar	53
Gambar 4.11 Visualisasi 2D Hasil <i>Sentence Embedding</i> Ketiga Layer ELMo	56
Gambar 4.12 Visualisasi 3D Hasil <i>Sentence Embedding</i> Ketiga Layer ELMo	57
Gambar 4.13 Visualisasi Data Menggunakan UMAP dengan Perbandingan 3:4	57
Gambar 4.14 Visualisasi Data Menggunakan UMAP dan SMOTE dengan Perbandingan 3:4	58
Gambar 4.15 Visualisasi Data Menggunakan UMAP dengan Perbandingan 1:10	59
Gambar 4.16 Visualisasi Data Menggunakan UMAP dan SMOTE dengan Perbandingan 1:10	59
Gambar 4.17 Visualisasi 2D Hasil <i>Sentence Embedding</i> Layer <i>Semantic</i> ELMo	60
Gambar 4.18 Visualisasi 3D Hasil <i>Sentence Embedding</i> Layer <i>Semantic</i> ELMo	61

DAFTAR TABEL

Tabel 2.1 <i>Confusion Matrix</i>	22
Tabel 4.1 Hasil Pengujian 1000 Data PCA dengan Perbandingan 3 : 4	48
Tabel 4.2 Hasil Pengujian 1000 Data PCA dengan Perbandingan 2 : 3	48
Tabel 4.3 Hasil Pengujian 1000 Data PCA dengan Perbandingan 1 : 2	48
Tabel 4.4 Hasil Pengujian 1000 Data PCA dengan Perbandingan 1 : 3	48
Tabel 4.5 Hasil Pengujian 1000 Data PCA dengan Perbandingan 1 : 4	49
Tabel 4.6 Hasil Pengujian 1000 Data PCA dengan Perbandingan 1 : 10	49
Tabel 4.7 Hasil Pengujian 1000 Data dengan Perbandingan 3 : 4	49
Tabel 4.8 Hasil Pengujian 1000 Data dengan Perbandingan 2 : 3	49
Tabel 4.9 Hasil Pengujian 1000 Data dengan Perbandingan 1 : 2	50
Tabel 4.10 Hasil Pengujian 1000 Data dengan Perbandingan 1 : 3	50
Tabel 4.11 Hasil Pengujian 1000 Data dengan Perbandingan 1 : 4	50
Tabel 4.12 Hasil Pengujian 1000 Data dengan Perbandingan 1 : 10	50
Tabel 4.13 Hasil Pengujian 5000 Data dengan Perbandingan 3 : 4	51
Tabel 4.14 Hasil Pengujian 5000 Data dengan Perbandingan 2 : 3	51
Tabel 4.15 Hasil Pengujian 5000 Data dengan Perbandingan 1 : 2	51
Tabel 4.16 Hasil Pengujian 5000 Data dengan Perbandingan 1 : 3	51
Tabel 4.17 Hasil Pengujian 5000 Data dengan Perbandingan 1 : 4	51
Tabel 4.18 Hasil Pengujian 5000 Data dengan Perbandingan 1 : 10	52
Tabel 4.19 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 3 : 4	53
Tabel 4.20 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 2 : 3	54
Tabel 4.21 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 1 : 2	54
Tabel 4.22 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 1 : 3	54
Tabel 4.23 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 1 : 4	54
Tabel 4.24 Hasil Pengujian 1000 Data <i>Semantic</i> dengan Perbandingan 1 : 10	54

DAFTAR RUMUS

Rumus 2.1 Representasi Teks ELMo	11
Rumus 2.2 UMAP <i>Cross Entrophy</i>	14
Rumus 2.3 <i>Resampling</i> SMOTE	15
Rumus 2.4 <i>Output</i> Single Layer Perceptron	17
Rumus 2.5 <i>Backpropagation Error</i>	18
Rumus 2.6 <i>Weight Update MLP</i>	18
Rumus 2.7 <i>Covariance</i>	19
Rumus 2.8 <i>Covariance Matrix</i>	20
Rumus 2.9 <i>Eigenvalues</i> dan <i>eigenvectors</i>	20
Rumus 2.10 <i>Precision</i>	22
Rumus 2.11 <i>Recall</i>	22
Rumus 2.12 <i>F-measure</i>	23

DAFTAR LAMPIRAN

Lampiran 1. Hasil Uji Coba	71
Lampiran 2. Daftar Riwayat Hidup.....	75
Lampiran 3. Formulir Bimbingan	76