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.)

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	Dosen Penguji
Moeljono Widjaja, M.Sc., Ph.D.	Seng Hansun, S.Si., M.Cs.
Dosen Pembimbing I	Dosen Pembimbing II
Julio Christian Young, M.Kom.	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 **Universtias 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

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

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

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR RUMUS

DAFTAR LAMPIRAN