PENERAPAN ALGORITMA RANDOM UNDER DAN OVER SAMPLING UNTUK MENGATASI CLASS IMBALANCE DALAM KLASIFIKASI TOPIK FORUM

Disusun sebagai salah satu syarat untuk memperoleh gelar

Sarjana Teknik Informatika

Disusun oleh:

NAWWAB ZIA AJNADEN

18106050042

PROGRAM STUDI INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA YOGYAKARTA

2023

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-209/Un.02/DST/PP.00.9/01/2023

Tugas Akhir dengan judul : Penerapan Algoritma Random Under dan Over Sampling Untuk Mengatasi Class Imbalance Dalam Klasifikasi Topik Forum

yang dipersiapkan dan disusun oleh:

Nama : NAWWAB ZIA AJNADEN
Nomor Induk Mahasiswa : 18106050042
Telah diujikan pada : Selasa, 10 Januari 2023
Nilai ujian Tugas Akhir : A-

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang

Nurochman, S.Kom., M.Kom
SIGNED

Valid ID: 63ce11315b352

Penguji I

Dr. Ir. Shofwatul 'Uyun, S.T., M.Kom.
SIGNED

Valid ID: 63c5082f7d569

Penguji II

Ir. Maria Ulfah Siregar, S.Kom., MIT., Ph.D.
SIGNED

Valid ID: 63c8e3de6f28d

Yogyakarta, 10 Januari 2023
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 63cf36ea8bd58

SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi
Lamp :

Kepada
Yth. Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta
di Yogyakarta

Assalamu'alaikum wr. wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka kami selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama	:	Nawwab Zia Ajnaden
NIM	:	18106050042
Judul Skripsi	:	Penerapan Algoritma Random Under dan Over Sampling Untuk Mengatasi Class Imbalance Dalam Klasifikasi Topik Forum

sudah dapat diajukan kembali kepada Program Studi Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Informatika

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara tersebut di atas dapat segera dimunaqsyahkan. Atas perhatiannya kami ucapkan terima kasih.

Wassalamu'alaikum wr. wb.

Yogyakarta, 3 Januari 2023
Pembimbing

Nurochman, S.Kom., M.Kom
NIP. 19801223 200901 1 007

PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini:

Nama	:	Nawwab Zia Ajnaden
NIM	:	18106050042
Program Studi	:	Informatika
Fakultas	:	Sains dan Teknologi

Menyatakan bahwa skripsi saya yang berjudul **"Penerapan Algoritma Random Under dan Over Sampling Untuk Mengatasi Class Imbalance Dalam Klasifikasi Topik Forum"** merupakan hasil penelitian saya sendiri, tidak terdapat pada karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi, dan bukan plagiasi karya orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 3 Januari 2023

Penulis

Nawwab Zia Ajnaden
NIM. 18106050042

KATA PENGANTAR

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat, nikmat dan petunjuk-Nya, sehingga saya dapat menyelesaikan tugas akhir ini. Shalawat serta salam penulis tujukan untuk Nabi Muhammad SAW, yang telah membawa kabar baik yakni agama Islam ke dalam peradaban umat manusia, dan semoga kita semua diberikan syafaat di akhirat kelak.

Selama proses penyusunan tugas akhir dengan judul "Penerapan Algoritma Random Under Dan Over Sampling Untuk Mengatasi Class Imbalance Dalam Klasifikasi Topik Forum" penulis mendapat banyak bantuan, saran, dan kritik yang diberikan dari berbagai pihak. Oleh karena itu pada kesempatan ini, izinkan penulis berterima kasih kepada:

1. Bapak, Ibu, saudara kembar, dan keluarga saya atas segala kasih sayang, kesabaran, dan pengorbanan dalam merawat dan menemani setiap langkah.
2. Bapak Nurochman, S.Kom., M.Kom. selaku Dosen Pembimbing Skripsi atas kesempatan, pikiran, saran, dan *support* selama pengerjaan skripsi.
3. Bapak Agus Mulyanto, S.Si., M.Kom. selaku Dosen Pembimbing Akademik yang membantu kegiatan akademik dan perkuliahan.
4. Mas Helmy dan senior di Symbolic.id, Galih, Irfan, Ridwan, Uqi, Davin, Fakhry (homwok), Fahry, Aman, Novan, Topik, Fikran dkk. Terima Kasih atas kesempatan dan waktu luang kalian.

5. Teman-teman angkatan 18, adik serta kakak kelas, dan Bapak serta Ibu Dosen dan Pegawai sebagai komponen memori saya terhadap UIN Sunan Kalijaga.
6. Semua pihak yang terlibat dalam hal apapun terkait penyusunan tugas akhir. Ucapan terima kasih saya tidak akan tereduksi meskipun tidak saya tulis namanya satu persatu.

Akhir kata, penulis sadar bahwa tugas akhir ini masih jauh dari kesempurnaan. Oleh karena itu, kritik dan saran dari segala aspek sangat diharapkan. Semoga terwujudnya tugas akhir ini dapat memberi bermanfaat untuk berbagai pihak tanpa terkecuali.

Yogyakarta, 3 Januari 2023
Penulis,

Nawwab Zia Ajnaden

HALAMAN PERSEMBAHAN

Untuk Bapak & Ibu untuk kesabaran dan kasih sayangnya, percayalah akan rasa sayang dan peduliku walau mungkin tak kau ketahui.

Untuk orang-orang yang saya kenal atas kesempatan dan kasih sayang.

HALAMAN MOTO

ready for unpredicted

iterate small

swallow failure

embrace imperfection

let life yield itself.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

PENERAPAN ALGORITMA RANDOM UNDER DAN OVER SAMPLING UNTUK MENGATASI CLASS IMBALANCE DALAM KLASIFIKASI TOPIK FORUM

Nawwab Zia Ajnaden
18106050042

INTISARI

Klasifikasi *user generated content* pada aplikasi sosial media memberikan beberapa kesempatan untuk memberikan manfaat yang berguna bagi pengembang dan pengguna. Faktor seperti keterbatasan waktu pembuatan fitur dan keterbatasan sumber data pada waktu tertentu dapat menyebabkan ketimpangan jumlah kelas pada label tertentu atau disebut *class imbalance* pada *dataset*. Teknik *resampling* seperti *random over* dan *under* sampling menjadi salah satu solusi dalam menyelesaikan permasalahan ini.

Penelitian ini membandingkan tiga model (dengan *classifier naive bayes*) dalam mengklasifikasi dua *dataset* yakni data *post* maupun data gabungan antara *post* dan komentar, ketiga model tersebut yakni: model yang tidak dikenai *resampling*, model dengan Random Over Sampling (ROS), serta model dengan Random Under Sampling (RUS). Seluruh data memiliki total 12 kelas topik.

Hasil penelitian menunjukkan peningkatan nilai *balanced accuracy* pada seluruh model yang dilengkapi dengan *resampling* (Tanpa *resampling*: 0.5792 dan 0.5078, ROS: 0.6148 dan 0.5570, RUS: 0.6040 dan 0.5225 untuk data *post*, serta data gabungan *post* dan komentar). Peningkatan terjadi hanya pada model yang dilatih dengan data *post* pada F1-*score* (Tanpa *resampling*: 0.5780 dan 0.5315, ROS: 0.6027 dan 0.5260, RUS: 0.5754 dan 0.4711 untuk data *post*, serta data gabungan *post* dan komentar) dan *precision* (tanpa *resampling*: 0.5816 dan 0.5774, ROS: 0.6027 dan 0.5155, RUS: 0.5754 dan 0.4653 untuk data *post*, serta data gabungan *post* dan komentar). Namun, seluruh model dengan *resampling* meningkatkan nilai *recall* (tanpa *resampling*: 0.5792 dan 0.5078, ROS: 0.6148 dan 0.5570, RUS: 0.6040 dan 0.5225 untuk data *post*, serta data gabungan *post* dan komentar) dan sejalan dengan kenaikan kuantitas hasil prediksi pada beberapa kelas minoritas (baik itu *true positives* maupun *false negatives*).

Kata kunci: Klasifikasi, Ketidakseimbangan Kelas, *Resampling*, Topik Forum

THE IMPLEMENTATION OF RANDOM UNDER AND OVER SAMPLING ALGORITHMS TO ADDRESS CLASS IMBALANCES IN FORUM TOPIC CLASSIFICATION.

Nawwab Zia Ajnaden
18106050042

ABSTRACT

Classification of user-generated content in social media applications provides several opportunities to provide useful benefits to developers and users. Factors such as limited time to create features and limited data sources at a certain time can cause an imbalance in the number of classes in a particular label in the dataset. Resampling techniques such as random over and under sampling are one of the solutions to solving this problem.

This research compares three models (with the naive bayes classifier) in classifying two datasets, namely, post data and combined data between posts and comments, the three models are: models that are not subject to resampling, models with Random Over Sampling (ROS), and models with Random Under Sampling (RUS). All data has a total of 12 topic classes.

The results showed an increase in the balanced accuracy value in all models equipped with resampling (Without resampling: 0.5792 and 0.5078, ROS: 0.6148 and 0.5570, RUS: 0.6040 and 0.5225 went to post data, and combined post and comment data, respectively). Improvement occurred only in models trained with post data on F1-score (Without resampling: 0.5780 and 0.5315, ROS: 0.6027 and 0.5260, RUS: 0.5754 and 0.4711 went to post data, and combined post and comment data, respectively) and precision (without resampling: 0.5816 and 0.5774, ROS: 0.6027 and 0.5155, RUS: 0.5754 and 0.4653 went to post data, and combined post and comment data, respectively). However, all models with resampling improved recall values (without resampling: 0.5792 and 0.5078, ROS: 0.6148 and 0.5570, RUS: 0.6040 and 0.5225 went to post data, and combined post and comment data, respectively) and correspond to an increase in the number of predicted results in some minority classes (both true positives and false negatives prediction).

Keywords: Classification, Class Imbalance, Resampling, Forum Topic

BAB I
PENDAHULUAN

1.1. Latar Belakang

Kesempatan masyarakat untuk berkumpul, berdiskusi, dan berinteraksi semakin luas dengan adanya sosial media. Semakin lama, dengan bertambahnya jumlah pengguna, beberapa fitur yang mempermudah baik itu pengguna dan developer perlu dibuat. Salah satunya dengan melakukan klasifikasi, mempelajari pola dan *behaviour* yang dapat membagi pengguna dalam beberapa segmen, dan memberikan suatu perilaku khusus pada tiap segmen tersebut, salah satu contohnya adalah dalam periklanan (Kosinski et al., 2013).

Usaha klasifikasi tentu memiliki banyak tantangan. Beberapa faktor pada data yang digunakan pada klasifikasi seperti jumlah pengguna pada aplikasi dan minat pada tiap forum, waktu yang tersedia untuk memulai atau mengeksekusi fitur, dan lama perusahaan pengembang berdiri pada sosial media dapat menimbulkan ketidakseimbangan kelas atau label sebagai bahan latih serta uji yang membuat bias pada kelas tertentu saja. Ketidakseimbangan kelas juga dapat menyebabkan klasifikasi yang tidak tepat pada kelas minoritas (Fahrudy, 2023).

Usaha dalam menyelesaikan ketidakseimbangan ini telah diteliti dan dilakukan, beberapa diantaranya ialah dengan menggunakan teknik *resampling*. Pada penelitian ini teknik *resampling* yang akan digunakan adalah mereduksi data yang terlalu banyak (atau disebut *random under sampling*), maupun dengan memperbanyak data yang terlalu sedikit (atau disebut *random over sampling*).

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah yang akan diselesaikan dalam penelitian ini adalah:

1. Bagaimana cara menerapkan *random over sampling* dan *random under sampling* untuk menyelesaikan permasalahan ketidakseimbangan kelas pada data.

2. Bagaimana perbandingan akurasi setelah dikenai *random over sampling* dan *random under sampling*.

1.3. Batasan Masalah

Pada penelitian ini didefinisikan beberapa batasan sebagai berikut:

a. Seluruh data berbentuk *text* disertai dengan label, data diambil dari aplikasi Symbolic.id dari tanggal 20 Januari 2020 sampai tanggal 20 Juni 2022. Data topik baik itu *post* maupun komentar akan diambil dari aplikasi Symbolic.id pada fitur *public space*. Seluruh karakter yang digunakan adalah huruf abjad.
b. Algoritma klasifikasi yang digunakan adalah *multinomial naive bayes*.
c. Perlakuan yang diberikan pada data adalah teknik *resampling*: *random under sampling* dan *random over sampling*.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini yaitu sebagai berikut:

a. Mengetahui implementasi *random over sampling* dan *random under sampling* untuk menyelesaikan permasalahan ketidakseimbangan kelas pada data.
b. Mengetahui perbedaan hasil akurasi pasca implementasi *random over sampling* dan *random under sampling*.

1.5. Manfaat Penelitian

Penelitian ini akan menjadi masukan bagi peneliti, pengembang khususnya yang berkutat pada sosial media, dan pembaca sebagai gambaran dan komparasi beberapa perlakuan untuk menyelesaikan masalah ketidakseimbangan data yang sering ditemukan pada *dataset* yang diambil langsung dari aplikasi terutama pada aplikasi sosial media.

1.6. Kebaruan Penelitian

Keaslian penelitian ini diperlukan sebagai bukti agar tidak plagiarisme berdasarkan beberapa penelitian terdahulu yang mempunyai karakteristik yang sama dalam hal metode, namun berbeda secara objek dan hasil penelitian.

BAB V
PENUTUP

5.1. Kesimpulan

Setelah melakukan rangkaian proses penelitian, dapat disimpulkan bahwa:

1. Setelah data dibagi untuk menghindari data leaking, data dimasukkan ke dalam *pipeline* yang kemudian dilatih dan dicari model serta parameter paling optimal. Proses *resampling* hanya terjadi saat proses pelatihan model.
2. Dalam kasus klasifikasi 12 kelas (Sains, Pertanian, Ekonomi Bisnis, Filsafat, Sosial Budaya, Politik, Psikologi, Agama, Pendidikan, Sejarah, Islam, dan Kesehatan) *posts* dan komentar pada aplikasi Symbolic.id dengan data latih sebesar 80% dan data uji sebesar 20% pada *dataset* baik itu hanya data *post* maupun gabungan antara *post* serta komentar terjadi peningkatan pada nilai *balanced accuracy* dan *recall* pada model dengan teknik resampling khususnya model dengan *random over sampling* (pada data *post, balanced accuracy* meningkat dari 0.5792980998 ke 0.6148711363 dan *recall* dari 0.57929809978804 ke 0.614871136330679).
3. Kenaikan juga terjadi pada nilai F1 dan *precision* pada model yang dilatih dan dikenai *random over sampling* pada data yang hanya berisi *posts*. Namun, kenaikan ini tidak terjadi pada model dengan data gabungan *posts* dan komentar.
4. Teknik resampling pada dua *dataset* (*post* dan gabungan *post* serta komentar) dapat meningkatkan nilai recall pada kelas minoritas. hal ini dibuktikan dengan meningkatnya jumlah label prediksi pada beberapa kelas minoritas.
5. Pada kedua metode *resampling*, metode *random over sampling* mendapatkan skor lebih baik pada seluruh matrik jika dibandingkan dengan *random under sampling*.

5.2. Saran

Dalam melakukan penelitian selanjutnya, peneliti menyarankan bahwa:

1. Seluruh model memiliki nilai skor pada seluruh matrik berada dibawah 0.7, pembersihan data pada *pipeline* atau penerapan *stemming* dan *lemmatization* mungkin dapat menaikkan skor dan dapat memperjelas perbedaan pada tiap model.
2. Pada penelitian ini, penggunaan dua variasi *dataset* bertujuan untuk menghilangkan faktor keberuntungan yang dikeluarkan oleh fungsi yang mengambil parameter *random state*. Perlakuan lain seperti variasi rasio pembagian data latih dan uji atau penggunaan *dataset* dengan rentang waktu yang berbeda dapat dilakukan.
3. Metode *resampling* yang digunakan pada penelitian ini terbatas pada dua metode dan tidak diatur rasio duplikasi atau penghapusannya, juga terdapat metode selain *random sampling* seperti SMOTE, TOMEK Links, dan lainnya.

DAFTAR PUSTAKA

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR), 49(2), 1-50.

Fahrudy, D. (2023). Classification of Student Graduation using Naïve Bayes by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection. *JOIV: International Journal on Informatics Visualization*, *6*(4).

Fahrudy, D., et al. (2022). Intelligent System For Classification Of Student Personality With Naive Bayes Algorithm. *Sintech (Science and Information Technology) Journal*, *5*(1), 1-9.

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence, 1(1), 89-101.

Kosinski, M., et al. (2014). Manifestations of user personality in website choice and behaviour on online social networks. Machine learning, 95(3), 357-380.

Lever, J. (2016). Classification evaluation: it is important to understand both what a classification metric expresses and what it hides. Nature Methods, 13(8), 603+. https://link.gale.com/apps/doc/A459507798/HRCA?u=anon~2474b51b&sid=googleScholar&xid=d1137cad

Muhsina, E. A., & Nurochman, N. (2017). Sistem Pakar Rekomendasi Profesi Berdasarkan Multiple Intelligences Menggunakan Teorema Bayesian. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *2*(1), 16-25.

Nugraha, W., & Sabaruddin, R. (2021). Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4. 5, Random Forest, dan SVM. Techno. Com, 20(3), 352-361.

Pedregosa, et al. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Purwa, T. (2019). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). Jurnal Matematika, Statistika dan Komputasi, 16(1), 58-73.

Saputro, E., & Rosiyadi, D. (2022). Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes. Bianglala Informatika, 10(1), 42-47.

Sasaki, Y., 2007. The Truth Of the F--Measure. Manchester: School of Computer Science, University of Manchester.

Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. Jurnal Informatika, 5(2), 175-185.

Yagis, E., Atnafu, S. W., de Herrera, A. G. S., et al. (2021). Deep learning in brain MRI: Effect of data leakage due to slice-level split using 2D convolutional neural networks.