

# **EVALUASI PEMBELAJARAN *MANIFOLD* DAN TEKNIK *UNDERSAMPLING* PADA KLASIFIKASI SENTIMEN ULASAN FILM DENGAN KELAS TIDAK SEIMBANG**

**L.M. RISMAN DWI JUMANSYAH**



**PROGRAM STUDI STATISTIKA DAN SAINS DATA  
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2025**



## @Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

## PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Evaluasi Pembelajaran *Manifold* dan Teknik *Undersampling* pada Klasifikasi Sentimen Ulasan Film dengan Kelas Tidak Seimbang” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Mei 2025

L.M. Risman Dwi Jumansyah  
G1501222058



## RINGKASAN

L.M. Risman Dwi Jumansyah. Evaluasi Pembelajaran *Manifold* dan Teknik *Undersampling* pada Klasifikasi Sentimen Ulasan Film dengan Kelas Tidak Seimbang. Dibimbing oleh Agus Mohamad Soleh dan Utami Dyah Syafitri.

Ulasan film merupakan sumber informasi penting di platform media sosial karena secara langsung mencerminkan persepsi penonton terhadap berbagai aspek teknis dan non-teknis dari suatu film. Analisis sentimen terhadap ulasan film memainkan peran penting dalam menentukan kesuksesan sebuah film, karena ulasan negatif dapat memengaruhi keputusan calon penonton untuk tidak menonton film tersebut. *Support Vector Machine* (SVM) merupakan algoritma pembelajaran mesin yang banyak digunakan dalam klasifikasi sentimen ulasan film karena kemampuannya dalam memisahkan sentimen positif dan negatif secara efektif. Namun, kinerja algoritma ini dapat menurun ketika dihadapkan pada data yang tidak seimbang, yaitu kondisi di mana distribusi kelas tidak merata, suatu hal yang umum terjadi dalam data ulasan film yang biasanya didominasi oleh sentimen positif atau negatif. Selain itu, tingginya dimensi data yang disebabkan oleh banyaknya peubah dalam bentuk kata-kata unik di seluruh korpus ulasan dapat mengurangi kemampuan generalisasi model serta meningkatkan beban komputasi, yang pada akhirnya menurunkan efisiensi klasifikasi.

Penelitian ini mengatasi kedua permasalahan tersebut dengan menerapkan pembelajaran *manifold* untuk reduksi dimensi pada data hasil ekstraksi fitur, serta menggunakan teknik *undersampling* untuk menyeimbangkan distribusi antara kelas mayoritas dan minoritas. Metode pembelajaran *manifold* yang dieksplorasi meliputi *Principal Component Analysis* (PCA) dan *Laplacian Eigenmaps* (LE), yang dibandingkan dengan pendekatan *baseline* tanpa reduksi dimensi. Penanganan ketidakseimbangan kelas menggunakan *Random Undersampling* (RUS) dan *EasyEnsemble* dievaluasi terhadap pendekatan tanpa penanganan ketidakseimbangan dan metode *Random Oversampling* (ROS).

Penelitian ini menggunakan dua gugus data yaitu data ulasan film *The Raid 2: Berandal* dan gugus data ulasan film IMDB. Data ulasan film *The Raid 2: Berandal* sangat tidak seimbang, dengan 1763 ulasan positif dan hanya 70 ulasan negatif, mencerminkan isu ketidakseimbangan kelas yang menjadi fokus penelitian ini. Sementara itu, gugus data IMDB yang diperoleh dari Kaggle digunakan untuk simulasi berbagai skenario dengan menetapkan jumlah sampel kelas minoritas sebanyak 70 dan 140, serta menerapkan berbagai rasio ketidakseimbangan (*Imbalance Ratio*/IR) sebesar 4, 9, dan 19. Tujuan penggunaan gugus data ini adalah untuk mengevaluasi keandalan dan generalisasi metode yang diterapkan pada berbagai skenario ukuran kelas minoritas dan tingkat intensitas ketidakseimbangan.

Temuan penelitian menunjukkan bahwa pada tugas klasifikasi sentimen untuk *The Raid 2: Berandal*, PCA sebagai teknik pembelajaran *manifold linear* menunjukkan performa yang lebih baik dibandingkan LE. PCA mampu mereduksi dimensi fitur TF-IDF dari 4836 menjadi 435 komponen utama, yang berdampak pada peningkatan keseimbangan klasifikasi, sebagaimana ditunjukkan oleh nilai rata-rata G-Mean sebesar 55.7%. Meskipun sensitivitas model menggunakan PCA lebih rendah dibandingkan metode berbasis Word2Vec, pendekatan ini mampu

mempertahankan stabilitas performa. Di antara teknik *undersampling*, *EasyEnsemble* terbukti lebih efektif dibandingkan RUS dan ROS, dengan *G-Mean* tertinggi sebesar 79.6% saat diterapkan pada fitur TF-IDF dengan *kernel linear*. Metode ini meningkatkan spesifisitas dan kemampuan model dalam mengklasifikasikan ulasan negatif dengan jumlah data yang kecil. Kombinasi terbaik ditemukan pada penggunaan *EasyEnsemble* dan PCA dengan SVM *kernel linear* pada fitur TF-IDF, yang menghasilkan rata-rata *G-Mean* sebesar 63.6% serta mampu mengurangi waktu komputasi dari 72.125 detik menjadi 59.115 detik. PCA juga menunjukkan stabilitas yang lebih tinggi dibandingkan LE dalam berbagai skenario, meskipun LE diberesapa kondisi menghasilkan *G-Mean* tertinggi.

Analisis terhadap skenario simulasi pada gugus data IMDB mengonfirmasi bahwa penerapan teknik pembelajaran *manifold* dan *undersampling* berkontribusi terhadap peningkatan performa klasifikasi. PCA secara konsisten menunjukkan stabilitas yang lebih baik dibandingkan LE dalam mereduksi dimensi fitur TF-IDF dan Word2Vec tanpa kehilangan informasi yang signifikan. PCA dengan SVM *kernel linear* unggul dibandingkan LE, terutama pada skenario dengan ketidakseimbangan tinggi (IR 9 dan 19), di mana model tanpa reduksi dimensi mengalami kesulitan dalam mengidentifikasi kelas minoritas. *EasyEnsemble* melampaui kinerja RUS ketika jumlah sampel kelas minoritas mencapai 140, dengan menunjukkan kemampuan mempertahankan spesifisitas model meskipun IR meningkat.

Kombinasi antara pembelajaran manifold dan teknik *undersampling* terbukti meningkatkan performa klasifikasi. PCA dengan ROS yang diterapkan pada fitur Word2Vec atau gabungan Word2Vec+TF-IDF mencapai *G-Mean* tertinggi sebesar 65%, sementara LE yang dikombinasikan dengan *EasyEnsemble* mencapai hingga 79,7% dalam skenario tertentu. PCA secara konsisten menunjukkan stabilitas yang lebih tinggi dibandingkan LE, khususnya dalam kondisi ketidakseimbangan kelas yang ekstrem. Pendekatan paling efektif untuk gugus data IMDB melibatkan kombinasi PCA dan ROS pada fitur Word2Vec atau Word2Vec+TF-IDF, yang memberikan keseimbangan terbaik dalam mengklasifikasikan ulasan positif dan negatif.

Studi ini menunjukkan bahwa kombinasi teknik pembelajaran *manifold* dan *undersampling* dapat meningkatkan performa klasifikasi sentimen dalam kondisi ketidakseimbangan kelas yang ekstrem sekaligus mengurangi kompleksitas komputasi. Meskipun performanya belum melampaui metode yang berfokus secara khusus pada penanganan ketidakseimbangan kelas, pendekatan yang diusulkan ini menawarkan alternatif yang layak dan dapat diandalkan.

Kata kunci: ketidakseimbangan kelas, pembelajaran *manifold*, reduksi dimensi, ulasan film, *undersampling*



## SUMMARY

L.M. Risman Dwi Jumansyah. Evaluation of Manifold Learning and Undersampling Techniques in Sentiment Classification of Imbalanced Movie Reviews Data. Supervised by Agus Mohamad Soleh and Utami Dyah Syafitri.

Movie reviews represent a crucial source of information on social media platforms, as they directly reflect viewers' perceptions of various technical and non-technical aspects. Sentiment analysis of movie reviews plays a significant role in determining a film's success since negative reviews can influence prospective viewers' decisions not to watch a particular film. Support Vector Machine (SVM) is a widely used machine learning algorithm for sentiment classification of movie reviews because it can effectively separate positive and negative sentiments. However, the performance of this algorithm may decline when faced with imbalanced data, a condition in which class distribution is unequal, a common occurrence in movie review data dominated by either positive or negative sentiments. The high dimensionality of the data, resulting from many variables in the form of unique words appearing across the entire review corpus, can also reduce the model's generalization ability and increase the computational burden, thereby decreasing classification efficiency.

This study addresses both challenges by applying manifold learning for dimensionality reduction of feature-extracted data and employing undersampling techniques to balance the distribution between majority and minority classes. The manifold learning methods explored include Principal Component Analysis (PCA) and Laplacian Eigenmaps (LE), which are compared with a baseline approach that does not apply dimensionality reduction. For class imbalance handling, Random Undersampling (RUS) and EasyEnsemble are evaluated against approaches without imbalance handling and the Random Oversampling (ROS) method.

Two datasets are utilized in this study: reviews of The Raid 2: Berandal and the IMDB movie review dataset. The Raid 2: Berandal dataset is heavily skewed toward positive reviews, with 1763 positive and only 70 negative reviews, reflecting the class imbalance issue addressed in this research. The IMDB dataset, obtained from Kaggle, simulates various scenarios by setting the number of minority class samples to 70 and 140 and applying different imbalance ratios (IR) of 4, 9, and 19. The purpose of using this dataset is to evaluate the reliability and generalizability of the applied methods across varying scenarios of minority class size and imbalance severity.

The findings indicate that in the sentiment classification task for The Raid 2: Berandal, PCA, as a linear manifold learning technique, outperforms LE. PCA reduced the dimensionality of TF-IDF features from 4836 to 435 principal components, improving classification balance, as indicated by an average G-Mean of 55.7%. Although model sensitivity using PCA is lower than Word2Vec-based methods, this approach maintains performance stability. Among the undersampling techniques, EasyEnsemble proves more effective than RUS and ROS, achieving the highest G-Mean of 79.6% when applied to TF-IDF features with a linear kernel. This method enhances specificity and improves the model's classification of the underrepresented negative reviews. The best-performing combination integrates EasyEnsemble and PCA using linear SVM on TF-IDF features, which yields an

@Hak Cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

average G-Mean of 63.6% and reduces computation time from 72.125 seconds to 59.115 seconds. PCA also demonstrates greater stability than LE across different scenarios, although LE sometimes achieves the highest G-Mean.

Analysis of the IMDB simulation scenarios confirms that manifold learning and undersampling techniques contribute to improved classification performance. PCA consistently exhibits greater stability than LE in reducing the dimensionality of TF-IDF and Word2Vec features without significant information loss. PCA with linear SVM performs better than LE, particularly in high-imbalance scenarios (IR 9 and 19), where models without dimensionality reduction struggle to identify minority classes. EasyEnsemble outperforms RUS when the number of minority class samples reaches 140, showing the ability to maintain model specificity even as IR increases. The combination of manifold learning and undersampling yields improved classification performance. PCA with ROS applied to Word2Vec or Word2Vec+TF-IDF features achieves the highest G-Mean of 65%, while LE combined with EasyEnsemble reaches up to 79.7% in specific scenarios. PCA consistently proves to be more stable than LE, particularly under conditions of high-class imbalance. The most effective approach for the IMDB dataset involves combining PCA and ROS on Word2Vec or Word2Vec+TF-IDF features, providing the best balance in classifying positive and negative reviews.

This study demonstrates that combining manifold learning and undersampling techniques can enhance sentiment classification performance under extreme class imbalance conditions while reducing computational complexity. Although the performance does not yet surpass that of methods focused solely on class imbalance handling, the proposed approach offers a viable an

**Keywords:** class imbalance, dimensionality reduction, manifold learning, movie reviews, undersampling



## @Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

# **EVALUASI PEMBELAJARAN *MANIFOLD* DAN TEKNIK *UNDERSAMPLING* PADA KLASIFIKASI SENTIMEN ULASAN FILM DENGAN KELAS TIDAK SEIMBANG**

**L.M. RISMAN DWI JUMANSYAH**

Tesis  
sebagai salah satu syarat untuk memperoleh gelar  
Magister pada  
Program Studi Statistika dan Sains Data

**PROGRAM STUDI STATISTIKA DAN SAINS DATA  
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Tesis:  
Cici Suhaeni, S.Si., M.Si., Ph.D





## @Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Judul Tesis : Evaluasi Pembelajaran *Manifold* dan Teknik *Undersampling* pada  
Klasifikasi Sentimen Ulasan Film dengan Kelas Tidak Seimbang  
Nama : L.M. Risman Dwi Jumansyah  
NIM : G1501222058

Disetujui oleh

Pembimbing 1:  
Dr. Agus Mohamad Soleh, S.Si., M.T.



Pembimbing 2:  
Dr. Utami Dyah Syafitri, S.Si., M.Si.



Diketahui oleh

Ketua Program Studi Magister Statistika dan Sains Data :  
Dr. Agus Mohamad Soleh, S.Si., M.T.  
NIP 19750315 199903 1 004



Dekan Sekolah Sekolah Sains Data, Matematika dan Informatika :  
Prof. Dr. Ir. Agus Buono, M.Si., M.Kom  
NIP 19660702 199302 1 001



Tanggal Ujian:  
25 Maret 2025

Tanggal Lulus:

## PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian ini ialah klasifikasi teks, dengan judul “Evaluasi Pembelajaran *Manifold* dan Teknik *Undersampling* pada Klasifikasi Sentimen Ulasan Film dengan Kelas Tidak Seimbang”.

Penulis menyampaikan rasa syukur yang mendalam atas terselesaikannya tesis ini. Proses penyusunan tesis tidak akan berhasil tanpa dukungan dari berbagai pihak yang telah memberikan kontribusinya secara langsung maupun tidak langsung. Oleh karena itu, penulis menyampaikan ucapan terima kasih kepada:

1. Bapak Dr. Agus Mohamad Soleh, S.Si., M.T. dan Ibu Dr. Utami Dyah Syafitri, S.Si., M.Si., selaku komisi pembimbing, yang senantiasa meluangkan waktu untuk memberikan bimbingan, masukan, dukungan, dan arahan selama penyusunan tesis ini.
2. Ibu Cici Suhaeni, S.Si., M.Si., Ph.D, selaku dosen penguji luar komisi pembimbing, atas kritik dan saran yang membangun.
3. Bapak Prof. Dr. Ir. I Made Sumertajaya, M.Si., selaku pimpinan sidang tesis, atas arahan dan masukannya selama proses sidang tesis.
4. Seluruh dosen pengajar dan staf Departemen Statistika, atas ilmu, motivasi, serta bantuan yang telah diberikan selama masa perkuliahan.
5. Situs IMDb dan Letterboxd, atas ketersediaan data ulasan film yang sangat membantu dalam pengumpulan data untuk keperluan penelitian ini.
6. Teman-teman Pascasarjana Statistika dan Sains Data 2022 serta Himpunan Profesi Statistika dan Sains Data periode 2023, atas kebersamaan dan dukungannya selama studi berlangsung.
7. Ayah, Ibu, Kakak, Adik, serta seluruh keluarga besar yang senantiasa memberikan doa, dukungan, dan semangat dalam setiap langkah yang penulis tempuh.

Perlu dicatat bahwa dalam penelitian ini terdapat beberapa kata yang memiliki konotasi negatif, mengingat data yang digunakan berasal dari ulasan film dengan sentimen beragam. Kata-kata tersebut muncul sebagai bagian dari analisis klasifikasi sentimen dan tidak mencerminkan pandangan pribadi penulis. Penggunaan kata-kata tersebut semata-mata bertujuan untuk mendukung proses penelitian dan pengolahan data secara objektif.

Semoga kedepannya karya ilmiah ini dapat bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Mei 2025

*L.M. Risman Dwi Jumansyah*

## DAFTAR ISI

|                                                                                                                             |     |
|-----------------------------------------------------------------------------------------------------------------------------|-----|
| DAFTAR TABEL                                                                                                                | xiv |
| DAFTAR GAMBAR                                                                                                               | xiv |
| DAFTAR LAMPIRAN                                                                                                             | xvi |
| I PENDAHULUAN                                                                                                               | 1   |
| 1.1 Latar Belakang                                                                                                          | 1   |
| 1.2 Rumusan Masalah                                                                                                         | 4   |
| 1.3 Tujuan                                                                                                                  | 4   |
| 1.4 Manfaat                                                                                                                 | 5   |
| II TINJAUAN PUSTAKA                                                                                                         | 7   |
| 2.1 Pembelajaran <i>Manifold</i>                                                                                            | 7   |
| 2.2 <i>Principal Component Analysis</i>                                                                                     | 7   |
| 2.3 <i>Laplacian Eigenmaps</i>                                                                                              | 8   |
| 2.4 <i>Undersampling</i>                                                                                                    | 9   |
| 2.5 <i>EasyEnsemble</i>                                                                                                     | 9   |
| 2.6 <i>Support Vector Machine</i>                                                                                           | 10  |
| 2.7 Evaluasi Kinerja Klasifikasi                                                                                            | 12  |
| III METODE                                                                                                                  | 15  |
| 3.1 Data                                                                                                                    | 15  |
| 3.2 Tahapan Penelitian                                                                                                      | 15  |
| 3.3 Pra Pemrosesan Teks                                                                                                     | 21  |
| 3.4 Ekstraksi Fitur                                                                                                         | 23  |
| 3.5 <i>Hyperparameter Tuning</i>                                                                                            | 27  |
| IV HASIL DAN PEMBAHASAN                                                                                                     | 29  |
| 4.1 Eksplorasi Data                                                                                                         | 29  |
| 4.2 Performa Klasifikasi Sentimen Ulasan Film dengan Pembelajaran <i>Manifold</i>                                           | 33  |
| 4.3 Performa Klasifikasi Sentimen Ulasan Film dengan Teknik <i>Undersampling</i>                                            | 39  |
| 4.4 Performa Klasifikasi Sentimen Ulasan Film dengan Kombinasi Pembelajaran <i>Manifold</i> dan Teknik <i>Undersampling</i> | 46  |
| V SIMPULAN DAN SARAN                                                                                                        | 57  |
| 5.1 Simpulan                                                                                                                | 57  |
| 5.2 Saran                                                                                                                   | 58  |
| DAFTAR PUSTAKA                                                                                                              | 59  |
| LAMPIRAN                                                                                                                    | 63  |
| RIWAYAT HIDUP                                                                                                               | 99  |



## DAFTAR TABEL

|   |                                                                                       |    |
|---|---------------------------------------------------------------------------------------|----|
| 1 | <i>Confusion Matrix</i>                                                               | 12 |
| 2 | Hasil Pra Pemrosesan Teks                                                             | 23 |
| 3 | Hasil TF-IDF                                                                          | 25 |
| 4 | Hasil Word2Vec                                                                        | 26 |
| 5 | <i>Hyperparameter</i> yang digunakan                                                  | 28 |
| 6 | Jumlah dan Persentase Ulasan                                                          | 29 |
| 7 | Perbandingan Rata-Rata Waktu Komputasi Metode Kombinasi dengan Tanpa Metode Kombinasi | 51 |

## DAFTAR GAMBAR

|    |                                                                                                                                                                                                                                                                                                                                                                                                    |    |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Ilustrasi <i>Random Undersampling</i>                                                                                                                                                                                                                                                                                                                                                              | 9  |
| 2  | Skema Algoritma <i>EasyEnsemble</i>                                                                                                                                                                                                                                                                                                                                                                | 10 |
| 3  | Ilustrasi <i>hyperplane</i> pada SVM                                                                                                                                                                                                                                                                                                                                                               | 11 |
| 4  | Komponen <i>Boxplot</i> (A) <i>Boxplot Konvensional</i> (B) <i>Violin Plot</i>                                                                                                                                                                                                                                                                                                                     | 13 |
| 5  | Diagram Alir Penelitian                                                                                                                                                                                                                                                                                                                                                                            | 16 |
| 6  | Proses Pembentukan Gugus Data dengan Skenario Simulasi                                                                                                                                                                                                                                                                                                                                             | 18 |
| 7  | Ilustrasi perhitungan Word2Vec terboboti TF-IDF                                                                                                                                                                                                                                                                                                                                                    | 27 |
| 8  | Visualisasi Awan Kata (a) Ulasan Positif (b) Ulasan Negatif                                                                                                                                                                                                                                                                                                                                        | 29 |
| 9  | Awan Kata Tanpa Kata Umum (a) Ulasan Positif (b) Ulasan Negatif                                                                                                                                                                                                                                                                                                                                    | 30 |
| 10 | Kata-Kata dengan Rata-Rata nilai TF-IDF Tertinggi pada (a) Ulasan Positif (b) Ulasan Negatif                                                                                                                                                                                                                                                                                                       | 31 |
| 11 | <i>Cosine Similarity</i> Word2Vec untuk Kata (a) <i>Action</i> (b) <i>Actor</i> (c) <i>Character</i> (d) <i>Direct</i> (e) <i>Fight</i> (f) <i>Scene</i> (g) <i>Sound</i> (h) <i>Story</i>                                                                                                                                                                                                         | 32 |
| 12 | Performa Spesifisitas Data Empiris dengan Pembelajaran <i>Manifold</i> untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                           | 34 |
| 13 | Performa <i>G-Mean</i> Data Empiris dengan Pembelajaran <i>Manifold</i> untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                          | 34 |
| 14 | Performa Sensitivitas Data Empiris dengan Pembelajaran <i>Manifold</i> untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                           | 35 |
| 15 | Waktu Komputasi SVM <i>Linear</i> dan SVM <i>Linear</i> dengan PCA                                                                                                                                                                                                                                                                                                                                 | 36 |
| 16 | Performa Spesifisitas dengan Pembelajaran <i>Manifold</i> Gugus Data dengan Skenario Simulasi (1) $n_{minoritas} = 70$ dan IR = 4 (2) $n_{minoritas} = 70$ dan IR = 9 (3) $n_{minoritas} = 70$ dan IR = 19 (4) $n_{minoritas} = 140$ dan IR = 4 (5) $n_{minoritas} = 140$ dan IR = 9 (6) $n_{minoritas} = 140$ dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF | 36 |
| 17 | Performa Sensitivitas dengan Pembelajaran <i>Manifold</i> Gugus Data dengan Skenario Simulasi (1) $n_{minoritas} = 70$ dan IR = 4 (2) $n_{minoritas} = 70$ dan IR = 9 (3) $n_{minoritas} = 70$ dan IR = 19 (4) $n_{minoritas} = 140$ dan IR = 4 (5) $n_{minoritas} = 140$ dan IR = 9 (6) $n_{minoritas} = 140$ dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF | 37 |

- 18 Performa *G-Mean* dengan Pembelajaran *Manifold* Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan IR = 4 (2)  $n_{minoritas} = 70$  dan IR = 9 (3)  $n_{minoritas} = 70$  dan IR = 19 (4)  $n_{minoritas} = 140$  dan IR = 4 (5)  $n_{minoritas} = 140$  dan IR = 9 (6)  $n_{minoritas} = 140$  dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 38
- 19 Performa Sensitivitas Data Empiris dengan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 39
- 20 Performa Spesifisitas Data Empiris dengan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 40
- 21 ..Performa *G-Mean* Data Empiris dengan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 40
- 22 Prediksi Enam Ulasan pada Data Uji menggunakan teknik *EasyEnsemble* 41
- 23 Performa Spesifisitas dengan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan IR = 4 (2)  $n_{minoritas} = 70$  dan IR = 9 (3)  $n_{minoritas} = 70$  dan IR = 19 (4)  $n_{minoritas} = 140$  dan IR = 4 (5)  $n_{minoritas} = 140$  dan IR = 9 (6)  $n_{minoritas} = 140$  dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 43
- 24 Performa *G-Mean* dengan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan IR = 4 (2)  $n_{minoritas} = 70$  dan IR = 9 (3)  $n_{minoritas} = 70$  dan IR = 19 (4)  $n_{minoritas} = 140$  dan IR = 4 (5)  $n_{minoritas} = 140$  dan IR = 9 (6)  $n_{minoritas} = 140$  dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 44
- 25 Performa Sensitivitas dengan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan IR = 4 (2)  $n_{minoritas} = 70$  dan IR = 9 (3)  $n_{minoritas} = 70$  dan IR = 19 (4)  $n_{minoritas} = 140$  dan IR = 4 (5)  $n_{minoritas} = 140$  dan IR = 9 (6)  $n_{minoritas} = 140$  dan IR = 19 menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 45
- 26 Performa Sensitivitas Data Empiris dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 48
- 27 Performa Spesifisitas Data Empiris dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 49
- 28 Performa *G-Mean* Data Empiris dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 50
- 29 Performa Spesifisitas dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan IR = 4 (2)  $n_{minoritas} = 70$  dan IR = 9 (3)  $n_{minoritas} = 70$  dan IR = 19 (4)  $n_{minoritas} = 140$  dan IR = 4

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



30

@Hak cipta milik IPB University

(5)  $n_{minoritas} = 140$  dan  $IR = 9$  (6)  $n_{minoritas} = 140$  dan  $IR = 19$  menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 54

Performa Sensitivitas dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan  $IR = 4$  (2)  $n_{minoritas} = 70$  dan  $IR = 9$  (3)  $n_{minoritas} = 70$  dan  $IR = 19$  (4)  $n_{minoritas} = 140$  dan  $IR = 4$  (5)  $n_{minoritas} = 140$  dan  $IR = 9$  (6)  $n_{minoritas} = 140$  dan  $IR = 19$  menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 55

31

.Performa *G-Mean* dengan Kombinasi Pembelajaran *Manifold* dan Penanganan Ketidakseimbangan Kelas pada Gugus Data dengan Skenario Simulasi (1)  $n_{minoritas} = 70$  dan  $IR = 4$  (2)  $n_{minoritas} = 70$  dan  $IR = 9$  (3)  $n_{minoritas} = 70$  dan  $IR = 19$  (4)  $n_{minoritas} = 140$  dan  $IR = 4$  (5)  $n_{minoritas} = 140$  dan  $IR = 9$  (6)  $n_{minoritas} = 140$  dan  $IR = 19$  menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF 56

## DAFTAR LAMPIRAN

|    |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |    |
|----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Lampiran 1 Gugus Data Ulasan Film IMDB (6 Data Contoh)                                                                                                                                                                                                                                                                                                                                                                                                  | 65 |
| 2  | Lampiran 2 Ulasan Film <i>The Raid 2</i> : Berandal (6 Data Contoh)                                                                                                                                                                                                                                                                                                                                                                                     | 66 |
| 3  | Lampiran 3 Syntax Python <i>Laplacian Eigenmaps</i>                                                                                                                                                                                                                                                                                                                                                                                                     | 69 |
| 4  | Lampiran 4 Syntax R <i>EasyEnsemble SVM</i>                                                                                                                                                                                                                                                                                                                                                                                                             | 74 |
| 5  | Lampiran 5 Syntax Python <i>EasyEnsemble SVM</i>                                                                                                                                                                                                                                                                                                                                                                                                        | 75 |
| 6  | Lampiran 6 Syntax R Pra Pemrosesan Teks                                                                                                                                                                                                                                                                                                                                                                                                                 | 81 |
| 7  | Lampiran 7 Syntax R Ekstraksi Fitur                                                                                                                                                                                                                                                                                                                                                                                                                     | 83 |
| 8  | Lampiran 8 Performa Klasifikasi Sentimen pada Data Ulasan Film <i>The Raid 2</i> : Berandal dengan Penerapan Pembelajaran <i>Manifold</i>                                                                                                                                                                                                                                                                                                               | 85 |
| 9  | Lampiran 9 Tiga Metode Terbaik Tiap Skenario Simulasi pada Klasifikasi Sentimen Ulasan Film pada Gugus Data dengan Penerapan Pembelajaran <i>Manifold</i>                                                                                                                                                                                                                                                                                               | 86 |
| 10 | Lampiran 10 Grafik Akurasi Klasifikasi Sentimen Ulasan Film <i>The Raid 2</i> : Berandal dengan Pembelajaran <i>Manifold</i> untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                          | 87 |
| 11 | Lampiran 11 Grafik Akurasi Klasifikasi Sentimen Gugus Data Ulasan Film IMDB menggunakan Pembelajaran <i>Manifold</i> dengan Skenario Simulasi (1) $n_{minoritas} = 70$ dan $IR = 4$ (2) $n_{minoritas} = 70$ dan $IR = 9$ (3) $n_{minoritas} = 70$ dan $IR = 19$ (4) $n_{minoritas} = 140$ dan $IR = 4$ (5) $n_{minoritas} = 140$ dan $IR = 9$ (6) $n_{minoritas} = 140$ dan $IR = 19$ pada ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF | 88 |
| 12 | Lampiran 12 Performa Klasifikasi Sentimen pada Data Ulasan Film <i>The Raid 2</i> : Berandal dengan Penerapan Penanganan Ketidakseimbangan Kelas                                                                                                                                                                                                                                                                                                        | 89 |
| 13 | Lampiran 13 Tiga Metode Terbaik Tiap Skenario Simulasi pada Klasifikasi Sentimen Ulasan Film pada Gugus Data dengan Penerapan Penanganan Ketidakseimbangan Kelas                                                                                                                                                                                                                                                                                        | 90 |

|    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |    |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 14 | Lampiran 14 Grafik Akurasi Klasifikasi Sentimen Ulasan Film <i>The Raid</i> 2: Berandal dengan Penerapan Penanganan Ketidakseimbangan Kelas untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                                                            | 91 |
| 15 | Lampiran 15 Grafik Akurasi Klasifikasi Sentimen Gugus Data Ulasan Film IMDB menggunakan Penanganan Ketidakseimbangan Kelas dengan Skenario Simulasi (1) $n_{minoritas} = 70$ dan $IR = 4$ (2) $n_{minoritas} = 70$ dan $IR = 9$ (3) $n_{minoritas} = 70$ dan $IR = 19$ (4) $n_{minoritas} = 140$ dan $IR = 4$ (5) $n_{minoritas} = 140$ dan $IR = 9$ (6) $n_{minoritas} = 140$ dan $IR = 19$ pada ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                            | 92 |
| 16 | Lampiran 16 Performa Klasifikasi Sentimen pada Data Ulasan Film <i>The Raid</i> 2: Berandal dengan Penerapan Pembelajaran <i>Manifold</i> dan Penanganan Ketidakseimbangan Kelas                                                                                                                                                                                                                                                                                                                         | 93 |
| 17 | Lampiran 17 Tiga Metode Terbaik Tiap Skenario Simulasi pada Klasifikasi Sentimen Ulasan Film pada Gugus Data dengan Penerapan Pembelajaran <i>Manifold</i> dan Penanganan Ketidakseimbangan Kelas                                                                                                                                                                                                                                                                                                        | 95 |
| 18 | Lampiran 18 Grafik Akurasi Klasifikasi Sentimen Ulasan Film <i>The Raid</i> 2: Berandal dengan Kombinasi Penanganan untuk (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF                                                                                                                                                                                                                                                                                                                                    | 96 |
| 19 | Lampiran 15 Grafik Akurasi Klasifikasi Sentimen Gugus Data Ulasan Film IMDB dengan Kombinasi Pembelajaran <i>Manifold</i> dan Penanganan Ketidakseimbangan Kelas pada Skenario Simulasi (1) $n_{minoritas} = 70$ dan $IR = 4$ (2) $n_{minoritas} = 70$ dan $IR = 9$ (3) $n_{minoritas} = 70$ dan $IR = 19$ (4) $n_{minoritas} = 140$ dan $IR = 4$ (5) $n_{minoritas} = 140$ dan $IR = 9$ (6) $n_{minoritas} = 140$ dan $IR = 19$ menggunakan ekstraksi fitur (a) TF-IDF (b) Word2Vec (c) Word2Vec+TF-IDF | 97 |