SENTIMEN ANALISIS REVIEW FILM PADA IMDB MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION

Diajukan Sebagai Syarat Untuk Menyelesaikan

Pendidikan Program Strata-1 Pada

Jurusan Teknik Informatika

Oleh:

Muhammad Azka Faridi

NIM: 09021382227132

JURUSAN TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2025

ABSTRACT

Sentiment Analysis is a subfield of Machine Learning that focuses on analyzing opinions expressed in textual data. IMDb, as a widely used platform, provides a space for movie enthusiasts worldwide to share their thoughts and reviews. User feedback can serve as a benchmark for evaluating a film's success. This research aims to classify reviews into positive and negative categories using the Logistic Regression algorithm combined with Grid Search and Active Learning methods. The classification results show that the highest accuracy achieved is 90.90% using Logistic Regression with the combination of Grid Search and Active Learning. Meanwhile, Logistic Regression with Active Learning alone achieved an accuracy of 90.58%, Logistic Regression with Grid Search reached 90.17%, and the basic Logistic Regression model achieved an accuracy of 89.81%.

Keywords: Sentiment Analysis, Logistic Regression, Grid Search, Active Learning

ABSTRAK

Sentiment analysis merupakan salah satu bidang dalam machine learning yang berfokus pada analisis opini dalam bentuk teks. IMDb, sebagai platform yang telah lama digunakan, menyediakan informasi sekaligus menjadi wadah bagi para pecinta film di seluruh dunia untuk berbagi pendapat. Respons yang diberikan oleh pengguna dapat menjadi tolak ukur keberhasilan suatu film. Penelitian ini bertujuan untuk mengklasifikasikan opini menjadi dua kategori, yaitu positif dan negatif, dengan menggunakan algoritma Logistic Regression yang dikombinasikan dengan metode Grid Search dan Active Learning. Hasil klasifikasi menunjukkan bahwa akurasi tertinggi yang diperoleh adalah sebesar 90,90% pada Logistic Regression dengan kombinasi metode Grid Search dan Active Learning. Sementara itu, Logistic Regression dengan metode Active Learning menghasilkan akurasi sebesar 90,58%, dan dengan metode Grid Search sebesar 90,17%. Adapun Logistic Regression tanpa metode optimasi menghasilkan akurasi sebesar 89,81%.

Kata Kunci: Sentimen Analisis, Logistic Regression, Grid Search, Active Learning

KATA PENGANTAR

Segala Puji dan syukur kepada Allah Swt atas berkat, rahmat, dan RidhoNya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan skripsi ini yang berjudul “**Sentimen Analisis Review Film pada IMDb Menggunakan Algoritma Logistic Regression**”. Skripsi ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya

Dalam penulisan ini penulis menyadari bahwa tanpa adanya bimbingan bantuan serta dukungan maupun petunjuk dari semua pihak, tidak mungkin laporan ini dapat diselesaikan. Oleh karena itu pada kesempatan ini penulis menyampaikan rasa terima kasih kepada:

1. Allah Subhanahu Wa Ta’ala yang telah memberikan hamba kesehatan dan kelancaran sehingga dapat menyelesaikan laporan ini.
2. Keluarga tercinta yang selalu memberikan semangat.
3. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Hadipurnawan Satria, S.Kom., M.Sc., Ph.D.selaku Ketua Jurusan Teknik Informatika Fakultas Komputer Universitas Sriwijaya.
5. Ibu Junia Kurniati, M.Kom. selaku pembimbing kerja praktik yang telah memberikan arahan kepada penulis untuk menyelesaikan laporan ini..
6. Seluruh Bapak dan Ibu Dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya atas ilmu yang telah diberikan.
7. Seluruh Tata Usaha Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya atas kemudahan administrasi yang telah diberikan.
8. Teman-teman seperjuangan.

Penulis juga mengucapkan terima kasih kepada semua pihak yang telah membantu penulis dalam menyelesaikan laporan ini dan tidak dapat disebutkan satu per satu. Penulis berharap semoga tulisan ini dapat menjadi bahan pembelajaran dimasa yang akan datang. Penulis juga meminta maaf apabila terdapat kesalahan dan kekhilafan dalam penulisan laporan ini. Penulis sangat mengharapkan kritik dan saran yang sifatnya membangun demi kesempurnaan tulisan ini.

Palembang,

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1. Latar Belakang

Di era digital saat ini, opini dan ulasan yang diberikan oleh pengguna di berbagai platform online memiliki peran yang penting dalam membentuk pandangan publik terhadap suatu produk atau layanan. Salah satu platform yang sering digunakan untuk memberikan ulasan terkait film adalah IMDb, yang bisa menjadi referensi bagi para penikmat film dalam menilai kualitas suatu tayangan film berdasarkan opini, baik yang bersifat positif maupun negatif. Ulasan dari pengguna tidak hanya memengaruhi keputusan individu dalam memilih film, tetapi juga menjadi salah satu indikator keberhasilan sebuah produksi film di industri hiburan. Oleh karena itu, analisis sentimen semakin banyak diterapkan untuk memahami pola opini masyarakat sehingga dapat dimanfaatkan dalam memprediksi potensi penjualan serta membantu investor dalam proses pengambilan keputusan (Singh et al., 2013).

Analisis sentimen merupakan teknik dalam pemrosesan bahasa alami (Natural Language Processing) yang bertujuan untuk mengklasifikasikan suatu kalimat atau dokumen, apakah bersifat positif atau negatif (Haddi et al., 2013). Selain itu, analisis sentimen juga dapat digunakan untuk mengenali serta memahami emosi yang terkandung dalam sebuah teks (Mailoa, 2021). Analisis sentimen melibatkan penambangan teks untuk mengevaluasi, memproses, dan mengekstrak data teks. Proses ini dilanjutkan ke tahap prapemrosesan data, yang mencakup case folding, tokenization, penghapusan data duplikat dan stopword, normalisasi, stemming, serta klasifikasi sentimen (Rifaldi et al., 2023).

Metode analisis sentimen berbasis machine learning telah banyak diterapkan sebagai pendekatan utama untuk mengklasifikasikan serta memahami pola sentimen dalam data teks. Salah satu algoritma machine learning yang sering digunakan adalah Logistic Regression. Algoritma ini memiliki keunggulan dalam menghasilkan prediksi probabilitas yang akurat serta dapat mengklasifikasikan data ke dalam kelas positif, negatif, atau netral. Selain itu, Logistic Regression sering digunakan dalam analisis sentimen karena kemampuannya dalam memberikan interpretasi yang jelas terhadap hubungan antar variabel (Bahtiar et al., 2023).

Penelitian ini juga menerapkan metode Active Learning dan Grid Search. Grid Search digunakan untuk mencari kombinasi hyperparameter terbaik dengan menguji berbagai kemungkinan nilai melalui proses trial and error hingga menemukan hasil yang optimal (Muhamad Malik Matin, 2023). Sedangkan Active Learning digunakan untuk memilih sampel dengan tingkat ketidakpastian tertinggi dalam data uji, lalu menambahkannya ke dalam data latih secara bertahap, sehingga dapat mengurangi kebutuhan pelabelan tanpa mengorbankan akurasi (Schein & Ungar, 2007).

1. Rumusan Masalah

Mengacu pada latar belakang yang telah dijelaskan, muncul sejumlah permasalahan yang perlu diperhatikan, diantaranya:

* 1. Bagaimana performa algoritma Logistic Regression dalam mengklasifikasikan sentimen ulasan film pada dataset IMDb 50K?
  2. Seberapa besar pengaruh penerapan metode Grid Search dan Active Learning terhadap peningkatan performa model?
  3. Apakah model yang dihasilkan dapat memberikan akurasi, presisi, recall, dan F1-score yang memadai dalam tugas klasifikasi sentimen?

1. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis kinerja algoritma Logistic Regression dalam klasifikasi sentimen terhadap ulasan film IMDb.
2. Menerapkan teknik Grid Search untuk mengoptimalkan hyperparameter Logistic Regression guna meningkatkan performa klasifikasi.
3. Mengimplementasikan Active Learning untuk meningkatkan efisiensi proses pelabelan data tanpa mengorbankan akurasi model.
4. Mengevaluasi model menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menilai efektivitas klasifikasi sentimen.
5. Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Memberikan wawasan tentang efektivitas penggunaan Logistic Regression dalam klasifikasi sentimen berbasis teks ulasan.
2. Menunjukkan bagaimana teknik optimasi seperti Grid Search dan Active Learning dapat meningkatkan performa model machine learning.
3. Menjadi referensi bagi peneliti dan praktisi yang ingin membangun sistem analisis sentimen dengan efisiensi pelabelan dan akurasi yang optimal.
4. Mendukung pengambilan keputusan dalam industri hiburan atau platform digital berdasarkan opini publik yang dianalisis secara otomatis.
5. Batasan Masalah

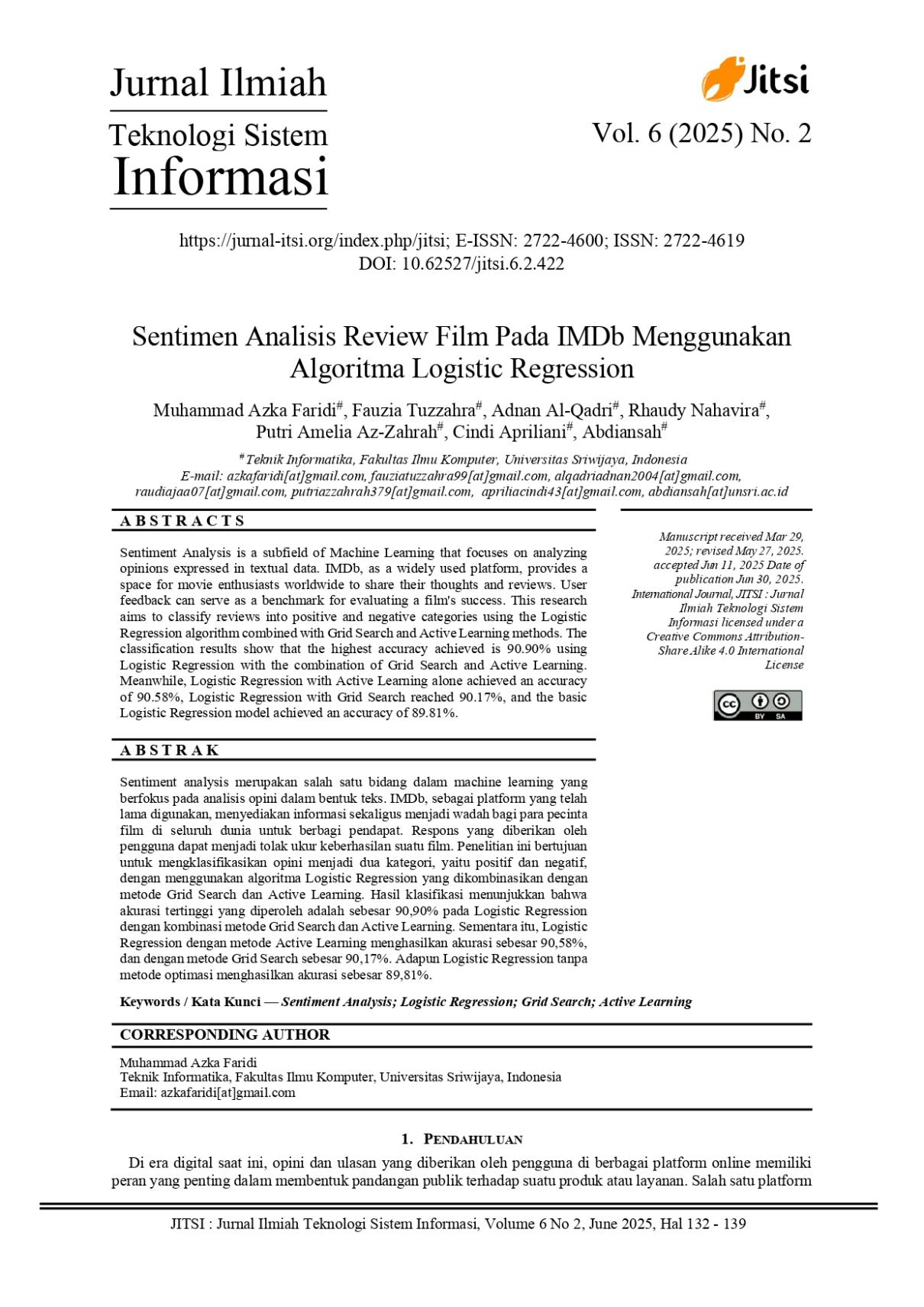
Batasan masalah diperlukan pada penelitian ini, karena luasnya ruang bidang yang dihadapi. Adapun batasan masalah yang ditetapkan dalam penelitian ini yaitu:

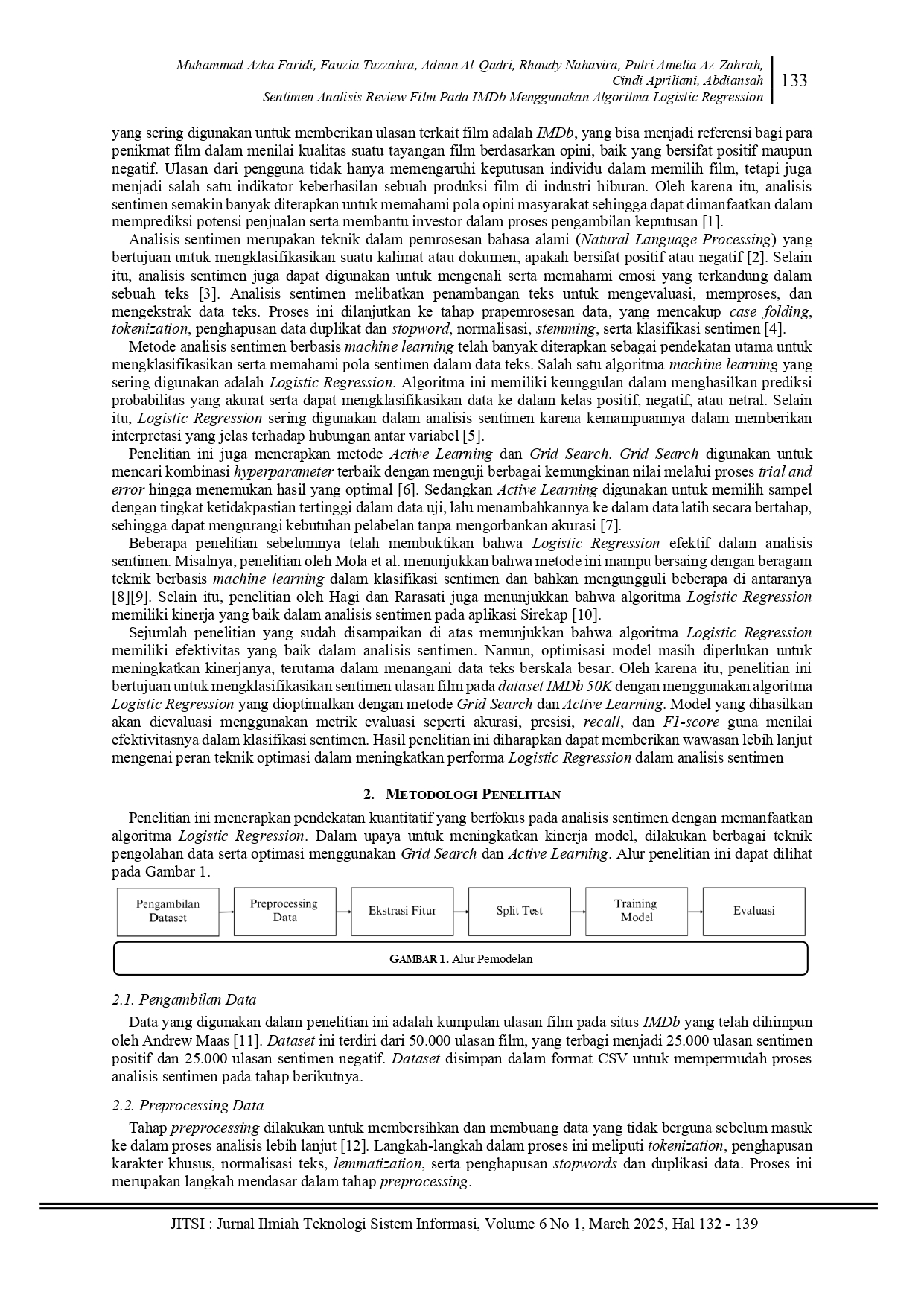
1. Penelitian ini hanya menggunakan dataset IMDb 50K yang berisi ulasan film berbahasa Inggris sebagai objek analisis.
2. Algoritma klasifikasi yang digunakan terbatas pada Logistic Regression dan tidak dibandingkan langsung dengan algoritma lain seperti SVM atau Random Forest.
3. Evaluasi model dibatasi pada metrik akurasi, presisi, recall, dan F1-score, tanpa melibatkan metrik lain seperti AUC-ROC.

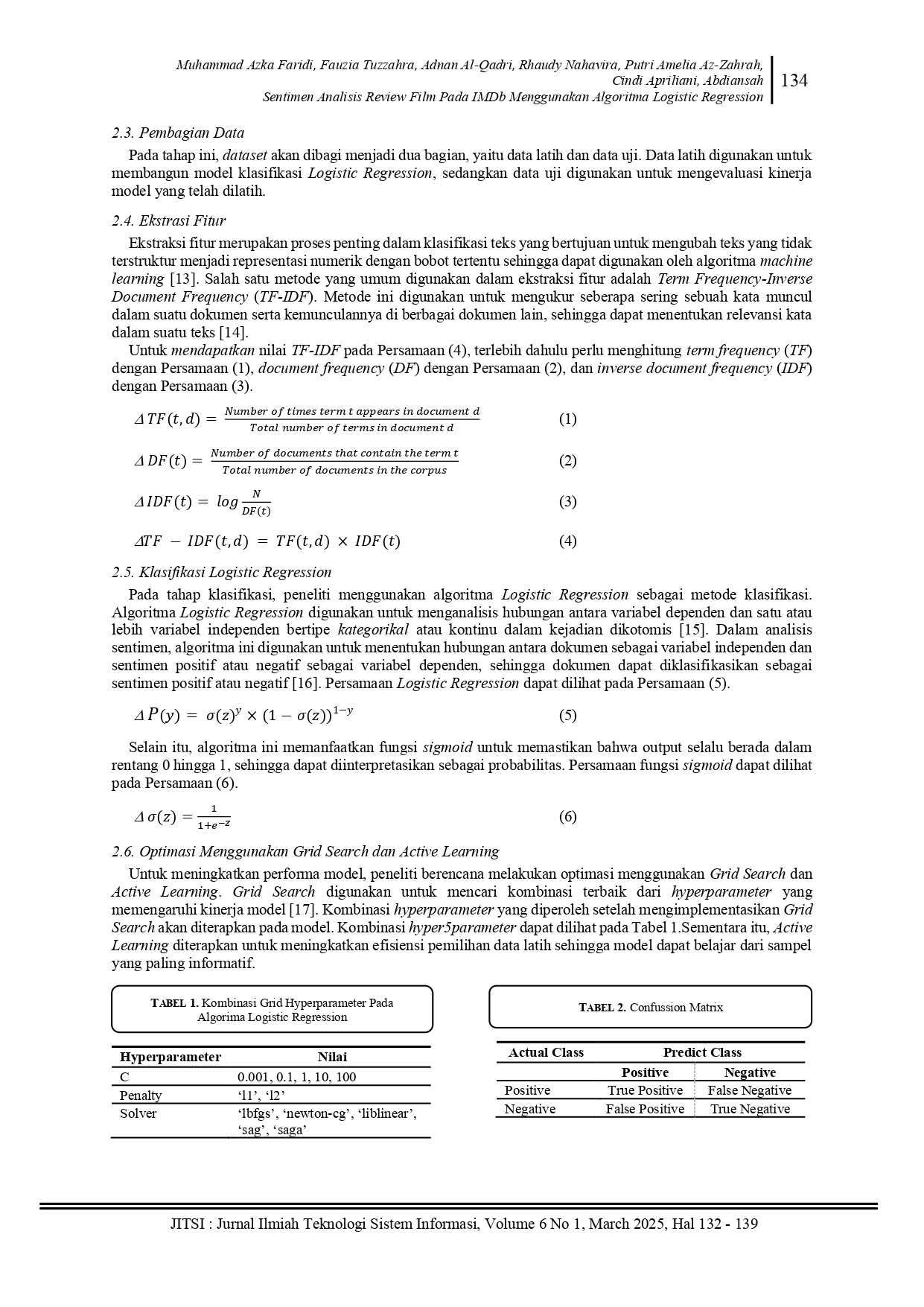
Fokus optimasi hanya mencakup hyperparameter tuning dengan Grid Search dan pemilihan data dengan Active Learning, tanpa mempertimbangkan metode optimasi lanjutan lainnya seperti Bayesian Optimization atau ensemble.

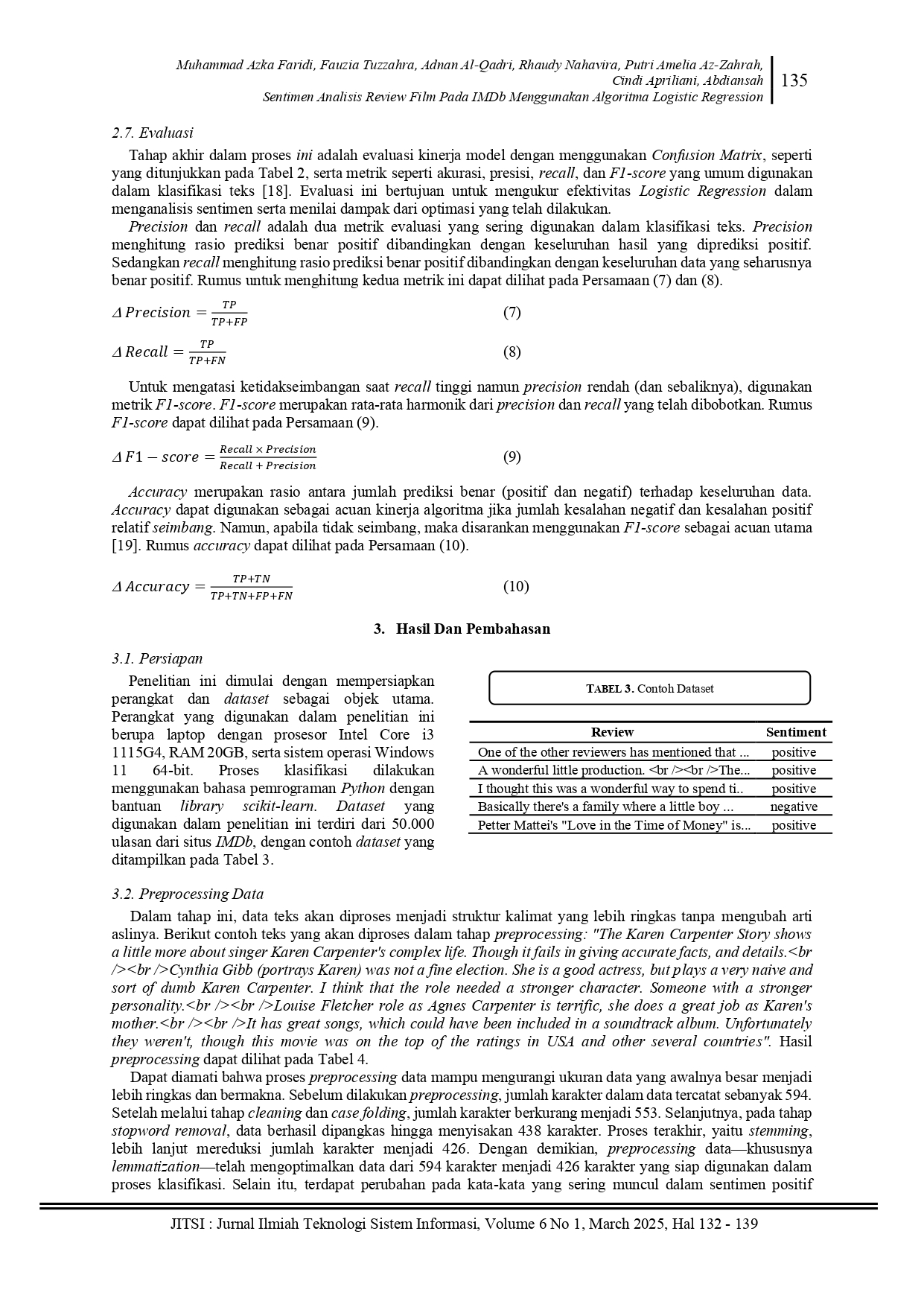
BAB II

ARTIKEL YANG TELAH DIPUBLIKASI

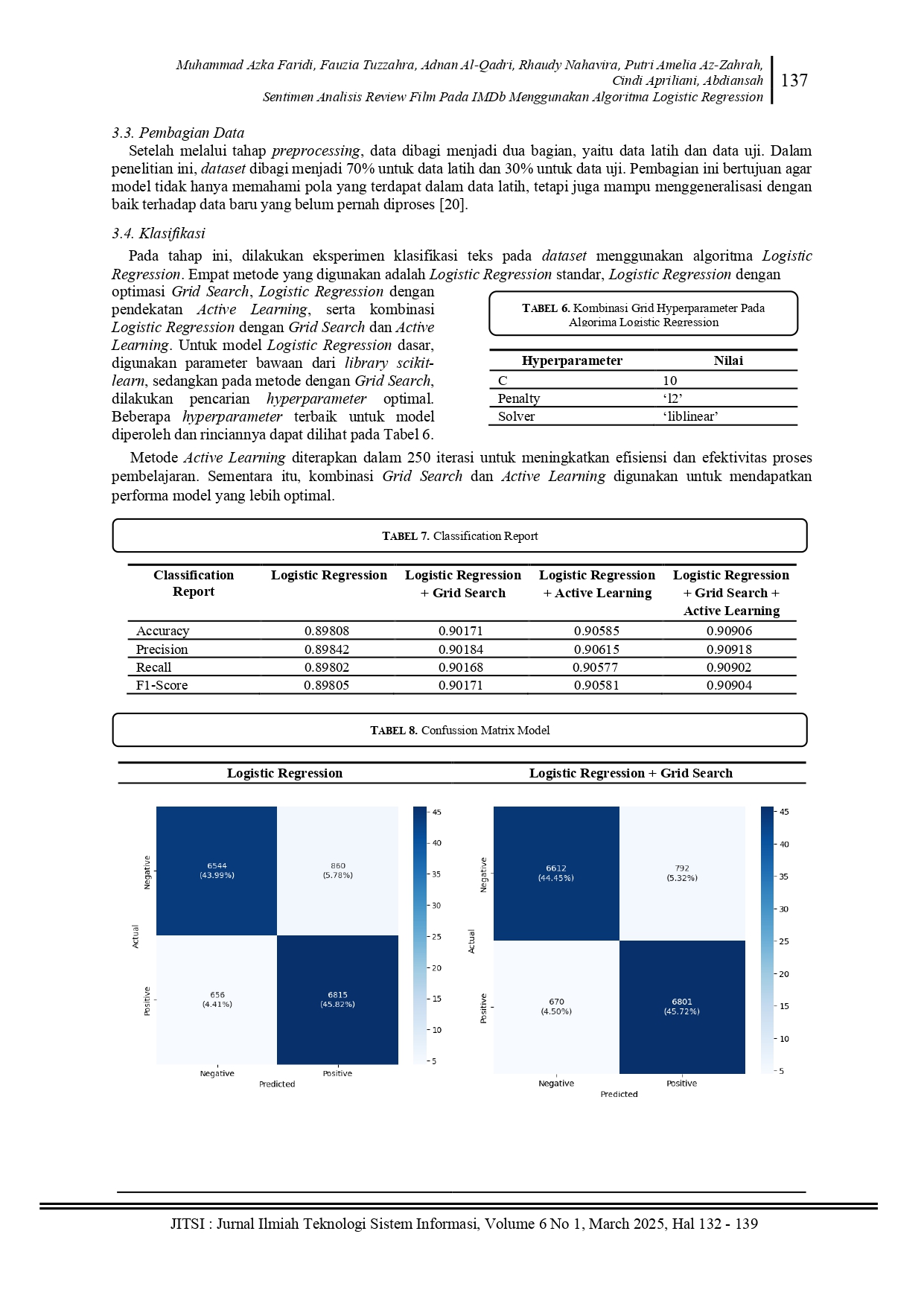


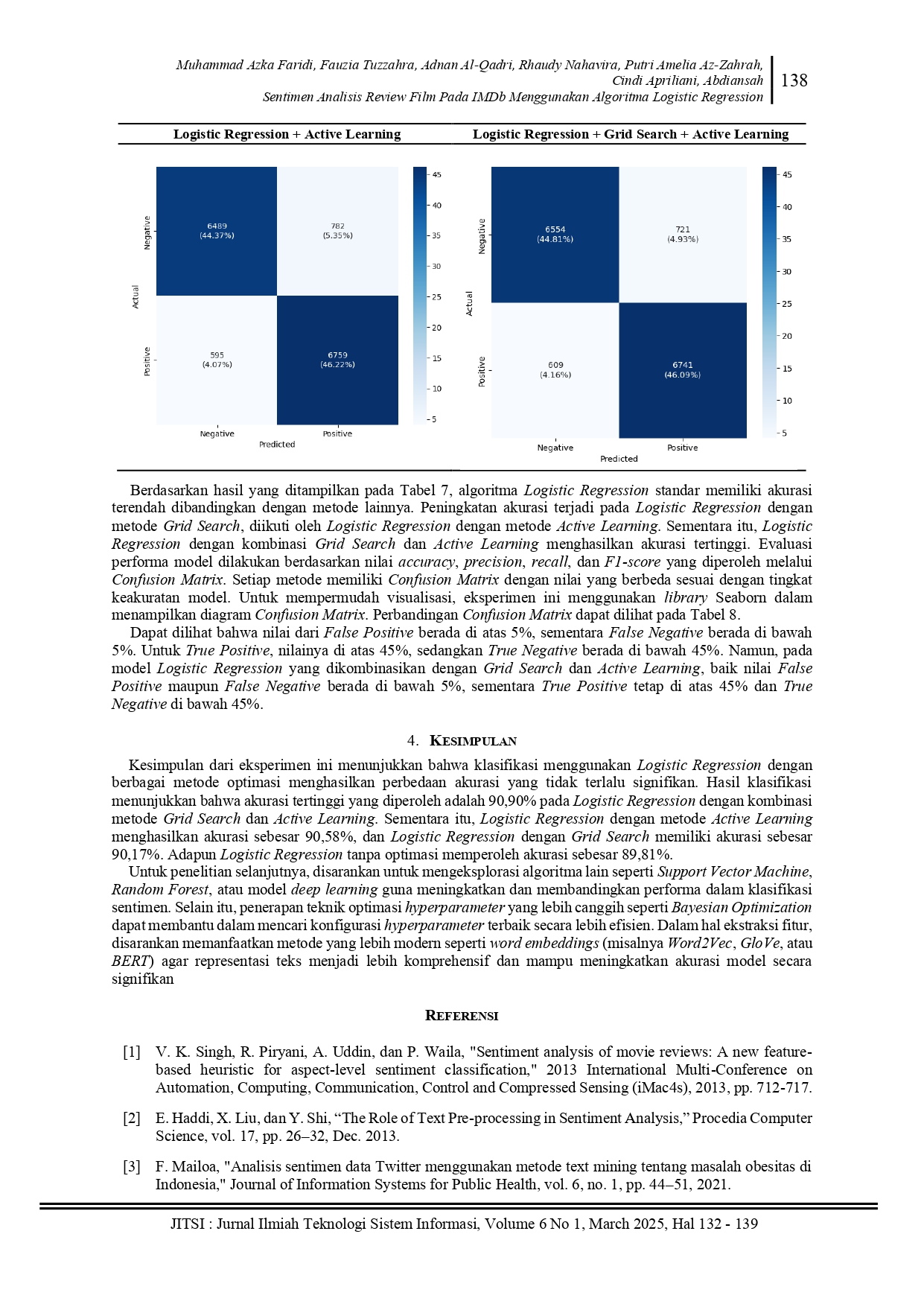


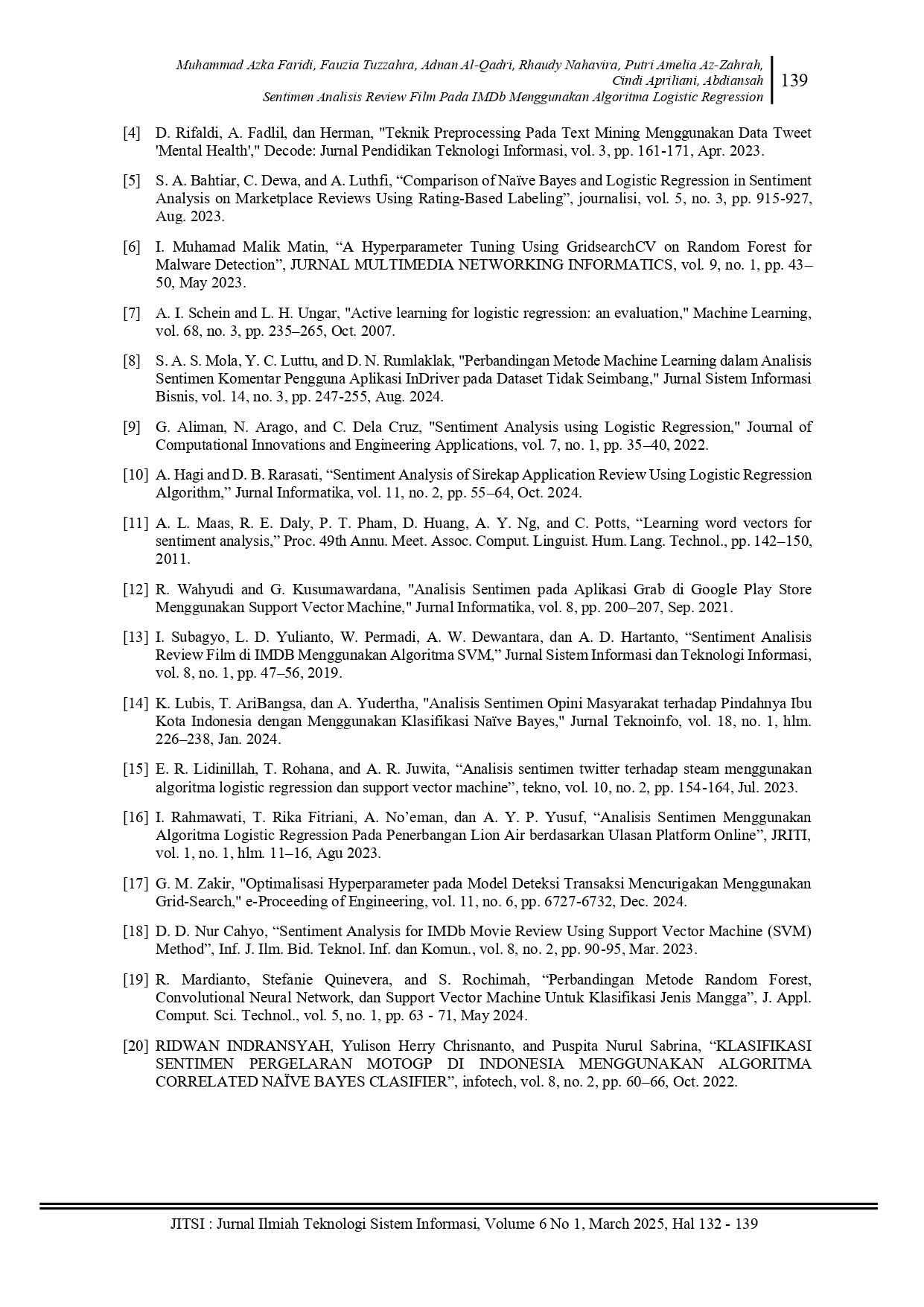












BAB III

METODELOGI PENELITIAN

1. Pendahuluan

BAB VI

KESIMPULAN DAN SARAN

1. Kesimpulan

Berdasarkan keseluruhan proses penelitian yang telah dilaksanakan, mulai dari tahap perumusan masalah, pengumpulan dan analisis data, hingga pengujian model, maka dapat disimpulkan bahwa:

1. Klasifikasi menggunakan Logistic Regression dengan berbagai metode optimasi menunjukkan perbedaan akurasi yang tidak terlalu signifikan.
2. Akurasi tertinggi sebesar 90,90% diperoleh dari Logistic Regression dengan kombinasi Grid Search dan Active Learning.
3. Logistic Regression dengan Active Learning saja menghasilkan akurasi 90,58%.
4. Logistic Regression dengan Grid Search saja memperoleh akurasi 90,17%.
5. Logistic Regression tanpa optimasi memperoleh akurasi terendah, yakni 89,81%.
6. Saran

Berdasarkan penelitian yang telah dilakukan, menghasilkan beberapa saran yang dapat dijadikan acuan bagi peneliti lain yang tertarik ingin mengembangkan atau melakukan studi serupa, di antaranya:

1. Disarankan untuk mengeksplorasi algoritma lain seperti Support Vector Machine, Random Forest, atau model deep learning guna membandingkan performa klasifikasi sentimen.
2. Disarankan untuk menerapkan teknik optimasi hyperparameter yang lebih canggih seperti Bayesian Optimization agar pencarian konfigurasi terbaik lebih efisien.
3. Disarankan untuk menggunakan metode ekstraksi fitur yang lebih modern seperti word embeddings (Word2Vec, GloVe, atau BERT) agar representasi teks lebih komprehensif dan akurasi model meningkat.

DAFTAR PUSTAKA

Bahtiar, S. A. H., Dewa, C. K., & Luthfi, A. (2023). Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling. *Journal of Information Systems and Informatics*, *5*(3), 915–927. https://doi.org/10.51519/journalisi.v5i3.539

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

Hagi, A., & Rarasati, D. B. (2024). Sentiment Analysis of Sirekap Application Review Using Logistic Regression Algorithm. *Jurnal Informatika*, *11*(2), 55–64. https://doi.org/10.31294/inf.v11i2.22066

Jaswanth, Y., Muni, R., Kumar, S., Sudhan, R. M., Vijaya Kumar, M., & Rajagopalam, M. (2020). Sentiment analysis using logistic regression algorithm. *European Journal of Molecular & Clinical Medicine*, *7*(4), 2081–2086. https://ejmcm.com/article\_1947.html

Lidinillah, E. R., Rohana, T., & Juwita, A. R. (2023). Analisis sentimen twitter terhadap steam menggunakan algoritma logistic regression dan support vector machine. *TEKNOSAINS : Jurnal Sains, Teknologi Dan Informatika*, *10*(2), 154–164. https://doi.org/10.37373/tekno.v10i2.440

Lubis, K. A., Bangsa, M. T. A., & Yudertha, A. (2024). Analisis Sentimen Opini Masyarakat terhadap Pindahnya Ibu Kota Indonesia dengan menggunakan Klasifikasi Naïve Bayes. *Jurnal Teknoinfo*, *18*(1), 226–238.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*(February), 142–150.

Mailoa, F. F. (2021). Analisis sentimen data twitter menggunakan metode text mining tentang masalah obesitas di indonesia. *Journal of Information Systems for Public Health*, *6*(1), 44. https://doi.org/10.22146/jisph.44455

Mardianto, R., Stefanie Quinevera, & Rochimah, S. (2024). Perbandingan Metode Random Forest, Convolutional Neural Network, dan Support Vector Machine Untuk Klasifikasi Jenis Mangga. *Journal of Applied Computer Science and Technology*, *5*(1), 63–71. https://doi.org/10.52158/jacost.v5i1.742

Mola, S. A. S., Luttu, Y. C., & Rumlaklak, D. N. (2024). Perbandingan Metode Machine Learning dalam Analisis Sentimen Komentar Pengguna Aplikasi InDriver pada Dataset Tidak Seimbang. *Jurnal Sistem Informasi Bisnis*, *14*(3), 247–255. https://doi.org/10.21456/vol14iss3pp247-255

Muhamad Malik Matin, I. (2023). Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware. *Multinetics*, *9*(1), 43–50. https://doi.org/10.32722/multinetics.v9i1.5578

Nur Cahyo, D. D., Farasalsabila, F., Lestari, V. B., Hanafi, Lestari, T., Al Islami, F. R., & Maulana, M. A. (2023). Sentiment Analysis for IMDb Movie Review Using Support Vector Machine (SVM) Method. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *8*(2), 90–95. https://doi.org/10.25139/inform.v8i2.5700

Rahmawati, I., Rika Fitriani, T., No’eman, A., & Yusuf, A. Y. P. (2023). Analisis Sentimen Menggunakan Algoritma Logistic Regression Pada Penerbangan Lion Air berdasarkan Ulasan Platform Online. *Jurnal Riset Informatika Dan Teknologi Informasi*, *1*(1), 11–16. https://doi.org/10.58776/jriti.v1i1.60

Rifaldi, D., Abdul Fadlil, & Herman. (2023). Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet “Mental Health.” *Decode: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Schein, A. I., & Ungar, L. H. (2007). Active learning for logistic regression: An evaluation. In *Machine Learning* (Vol. 68, Issue 3). https://doi.org/10.1007/s10994-007-5019-5

Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. *Proceedings - 2013 IEEE International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, IMac4s 2013*, *March*, 712–717. https://doi.org/10.1109/iMac4s.2013.6526500

Subagyo, I., Yulianto, L. D., Permadi, W., & Dewantara, A. W. (2019). Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM. *E-Jurnal JUSITI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, *8*–*1*(1), 47–56. https://doi.org/10.36774/jusiti.v8i1.600

Zakir, G. M., Prasasti, A. L., & Paryasto, M. W. (2024). Optimalisasi Hyperparameter pada Model Deteksi Transaksi Mencurigakan Menggunakan Grid-Search. *E-Proceeding of Engineering*, *11*(6), 6728–6731. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/25076/24016