Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 4 No. 3 (2026) pp: 1966-1976
P-ISSN: 2986-6790, e-ISSN: 2986-6782

Implementasi Klasifikasi Genre Film Pada Platform Imdb Berdasarkan Deskripsi Sinopsis Film Menggunakan Algoritma Naïve Bayes, SVM, Dan Logistic Regression

Muhammad Alfian Ilyassa[1], Satrio Akbar S[2], Ikhwanul Akmal[3], Fakhrullah Abrisam[4], Hanif Zaidan Sinaga[5]
[1,2,3,4,5] Bisnis Digital, Universitas Ibn Khaldun Bogor, Bogor
[1]alfianilyassa22@gmail.com,[2] satrioakbr15@gmail.com, [3]ikhwanulakmalakmal@gmail.com,
[4]fakhrullahabrisam558@gmail.com, [5]hanif.zaidan@gmail.com

Abstract
The rapid growth of digital film data and the extensive use of IMDb create a need for automatic methods that can classify genres consistently from textual information. This study aims to implement Natural Language Processing and compare Multinomial Naïve Bayes, Support Vector Machine LinearSVC, class-weighted balanced SVM, and Logistic Regression for classifying a film's primary genre from its synopsis. The IMDb Top 1000 dataset obtained from Kaggle was filtered to 811 films representing five major genres: Drama, Action, Comedy, Crime, and Biography. The research stages included case folding, cleaning, tokenization, stopword removal, stemming, Term Frequency-Inverse Document Frequency feature extraction, stratified 80:20 data splitting, model training, evaluation using accuracy, precision, recall, F1-score and confusion matrices, and five-fold stratified cross-validation. The majority-class baseline achieved 35.6% accuracy. Logistic Regression produced the best performance, with 43.56% test accuracy and 44.63% mean cross-validation accuracy. Balanced SVM achieved 42.33% accuracy, LinearSVC reached 41.10%, and Naïve Bayes obtained 40.49%. Naïve Bayes showed a strong bias toward the dominant Drama class, whereas SVM and Logistic Regression generated more balanced predictions across genres. The modest overall performance was mainly associated with class imbalance, narrative overlap among genres, short synopsis descriptions, and the reduction of multi-genre information to a single primary label. The findings indicate that Logistic Regression is the most effective classical model under the evaluated feature representation and dataset configuration and provides a suitable baseline for subsequent research.

Keywords: *Natural Language Processing, Film Genre Classification, Machine Learning, Imdb, Film Synopsis, Naïve Bayes, Support Vector Machine, Logistic Regression*

Abstrak
Pertumbuhan data film digital dan meningkatnya penggunaan platform IMDb menuntut metode otomatis yang mampu mengelompokkan genre berdasarkan informasi tekstual secara konsisten. Penelitian ini bertujuan mengimplementasikan Natural Language Processing dan membandingkan kinerja Multinomial Naïve Bayes, Support Vector Machine LinearSVC, SVM dengan pembobotan kelas seimbang, serta Logistic Regression untuk mengklasifikasikan genre primer film berdasarkan sinopsis. Dataset IMDb Top 1000 dari Kaggle diseleksi menjadi 811 film pada lima genre utama, yaitu Drama, Action, Comedy, Crime, dan Biography. Tahapan penelitian meliputi case folding, cleaning, tokenization, stopword removal, stemming, ekstraksi fitur Term Frequency-Inverse Document Frequency, pembagian data secara stratified 80:20, pelatihan model, evaluasi accuracy, precision, recall, F1-score, confusion matrix, dan validasi silang stratified lima lipatan. Baseline kelas mayoritas menghasilkan akurasi 35,6%. Logistic Regression memberikan hasil terbaik dengan akurasi pengujian 43,56% dan rata-rata validasi silang 44,63%. SVM Balanced mencapai akurasi 42,33%, LinearSVC 41,10%, dan Naïve Bayes 40,49%. Naïve Bayes menunjukkan bias kuat terhadap kelas Drama, sedangkan SVM dan Logistic Regression menghasilkan prediksi lebih seimbang. Kinerja yang masih terbatas terutama dipengaruhi ketidakseimbangan kelas, tumpang tindih naratif antargenre, sinopsis yang pendek, dan penyederhanaan label multigenre menjadi genre primer. Hasil penelitian menegaskan bahwa Logistic Regression merupakan model klasik paling efektif pada konfigurasi data dan fitur yang digunakan serta layak dijadikan baseline untuk penelitian lanjutan.
Kata Kunci: Natural Language Processing, Klasifikasi Genre Film, Machine Learning, IMDb, Sinopsis Film, Naïve Bayes, Support Vector Machine, Logistic Regression.

1. Pendahuluan

Perkembangan teknologi informasi dan komunikasi pada era digital telah menghasilkan pertumbuhan data yang sangat besar, khususnya data dalam bentuk teks. Data tekstual saat ini menjadi salah satu sumber informasi yang paling dominan di berbagai sektor, mulai dari bisnis, pendidikan, kesehatan, hingga industri hiburan digital. Dalam konteks industri hiburan, khususnya perfilman, data teks memiliki peran penting karena banyak informasi film

disajikan dalam bentuk deskripsi, ulasan, maupun sinopsis. Sinopsis film merupakan ringkasan cerita yang menggambarkan inti alur, karakter, konflik, dan tema utama dari sebuah film. Informasi ini menjadi elemen penting dalam membantu pengguna memahami isi film sebelum memutuskan untuk menontonnya.

Seiring dengan meningkatnya jumlah produksi film setiap tahun, platform digital seperti IMDb (Internet Movie Database) menjadi salah satu pusat data perfilman terbesar yang menyediakan jutaan informasi terkait film, seperti judul, genre, rating, ulasan pengguna, serta sinopsis film. IMDb menjadi salah satu referensi utama bagi pengguna dalam mencari informasi terkait film karena kelengkapan datanya yang sangat luas. Dataset IMDb juga banyak dimanfaatkan dalam penelitian Machine Learning dan Natural Language Processing (NLP) karena memiliki struktur data yang kaya dan relevan untuk klasifikasi teks, (Mohammad & Al-momani, 2016) .

Genre film merupakan salah satu komponen penting dalam industri perfilman karena berfungsi sebagai sistem kategorisasi yang membantu pengguna dalam menemukan jenis tontonan sesuai preferensi mereka. Genre seperti aksi, drama, horor, komedi, thriller, dan romantis memiliki karakteristik naratif yang berbeda-beda. Pada umumnya, proses pengklasifikasian genre dilakukan secara manual oleh editor atau penyedia platform film. Namun, metode manual memiliki beberapa keterbatasan, seperti memerlukan waktu yang lama, rawan terhadap subjektivitas manusia, serta kurang efisien dalam menangani jumlah data yang terus meningkat. Oleh karena itu, diperlukan suatu sistem otomatis yang mampu mengidentifikasi genre film berdasarkan konten teks yang tersedia, khususnya synopsis, (Mccallum, 1997).

Natural Language Processing (NLP) merupakan salah satu cabang dari kecerdasan buatan (Artificial Intelligence) yang berfokus pada interaksi antara komputer dan bahasa manusia. NLP memungkinkan sistem untuk memahami, mengolah, dan mengekstraksi informasi penting dari data teks yang tidak terstruktur. Dalam konteks klasifikasi genre film, NLP berperan dalam mengidentifikasi pola linguistik, struktur kalimat, serta kata-kata kunci yang menjadi ciri khas dari suatu genre tertentu. Tahapan NLP seperti case folding, tokenization, stopword removal, stemming, dan feature extraction menjadi proses penting untuk mentransformasikan teks mentah menjadi data yang dapat dipahami oleh model Machine Learning, (Manning et al., 2009).

Penerapan Machine Learning dalam klasifikasi genre film telah menjadi perhatian banyak peneliti. Penelitian oleh (Irawan & Abdullah, 2025) membandingkan algoritma K-Nearest Neighbor (KNN), Naïve Bayes, dan Support Vector Machine (SVM) dalam klasifikasi genre film berbasis sinopsis. Hasil penelitian menunjukkan bahwa SVM memiliki performa terbaik dengan tingkat akurasi mencapai 93%, dibandingkan algoritma lainnya (Irawan & Abdullah, 2025). Hal ini menunjukkan bahwa algoritma SVM memiliki kemampuan yang baik dalam menangani data teks berdimensi tinggi dan sparse. Penelitian terkini juga menunjukkan bahwa klasifikasi dan penandaan genre film dapat dikembangkan melalui word embedding, arsitektur hibrida, dan ensemble deep learning (González-Santos et al., 2024; Lin et al., 2024; Shao & Guo, 2024).

Studi oleh (Amelia et al., 2023) dalam survei komprehensif tentang identifikasi genre otomatis menunjukkan bahwa pendekatan modern berbasis transformer memberikan performa yang lebih unggul dibandingkan model klasik. Namun demikian, model klasik seperti Naïve Bayes dan SVM masih tetap relevan karena memiliki keunggulan dari segi efisiensi komputasi, kebutuhan sumber daya yang lebih rendah, serta lebih mudah diimplementasikan pada penelitian skala akademik. Selain itu, Logistic Regression dengan representasi TF-IDF juga dikenal sebagai model baseline yang kompetitif untuk klasifikasi teks, seringkali mengungguli SVM pada dataset kecil hingga menengah dengan distribusi kelas yang tidak seimbang (Lukman & Novianto, 2025) Di luar domain film, kajian terbaru menegaskan pentingnya representasi fitur, efisiensi komputasi, deskripsi kelas, penanganan ketidakseimbangan, dan interpretabilitas pada klasifikasi teks (Xia et al., 2022; De Silva et al., 2023; Henning et al., 2023; Mustafa & Hama Saeed, 2025; Hossain & Kabir, 2025).

(Sebastiani, 2002) menyatakan bahwa representasi dokumen menggunakan Doc2Vec dapat meningkatkan performa klasifikasi teks dan menghasilkan nilai evaluasi yang kompetitif pada berbagai dataset klasifikasi dokumen.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk Language Processing dalam klasifikasi genre film pada platform IMDb berdasarkan deskripsi sinopsis menggunakan algoritma Machine Learning. Model yang digunakan dalam penelitian ini adalah Naïve Bayes dan Support Vector Machine (SVM) sebagai algoritma pembanding dengan Logistic Regression sebagai baseline tambahan. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, dan F1-Score serta validasi silang 5-fold untuk mengukur tingkat performa klasifikasiNovelty penelitian ini terletak pada evaluasi komprehensif empat model klasifikasi teks klasik, yaitu Naïve Bayes, SVM LinearSVC, SVM Balanced, dan Logistic Regression, pada dataset IMDb Top 1000 yang memiliki karakteristik multi-genre namun disederhanakan menjadi single-label berdasarkan genre primer. Penelitian ini juga mengombinasikan evaluasi menggunakan baseline majority class, confusion matrix, serta validasi silang stratified 5-fold untuk mengukur kemampuan model dalam menghadapi ketidakseimbangan kelas dan overlap antar genre. Pendekatan tersebut memberikan gambaran yang lebih komprehensif mengenai performa algoritma klasik pada klasifikasi sinopsis film dibandingkan penelitian sebelumnya yang umumnya hanya membandingkan satu atau dua algoritma tanpa mempertimbangkan baseline maupun pengaruh distribusi kelas.

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2. Metode Penelitian

Metodologi penelitian merupakan tahapan sistematis yang digunakan untuk mencapai tujuan penelitian. Pada penelitian ini, metode yang digunakan berfokus pada implementasi Natural Language Processing (NLP) untuk melakukan klasifikasi genre film berdasarkan sinopsis menggunakan algoritma Machine Learning, yaitu Naïve Bayes dan Support Vector Machine (SVM). Tahapan penelitian dilakukan secara berurutan mulai dari pengumpulan data hingga evaluasi model.

2.1 Tahapan Penelitian

Penelitian ini terdiri dari beberapa tahapan utama yang dilaksanakan secara sistematis sebagai berikut.

1. Pengumpulan Dataset (Data Collection)

Tahap pertama adalah pengumpulan dataset film yang diperoleh dari platform IMDb melalui dataset publik Kaggle, yaitu IMDb Top 1000 Movies. Dataset berisi informasi film berupa judul, genre, dan sinopsis film. Dataset kemudian dilakukan seleksi atribut agar hanya menggunakan data yang relevan, yaitu: Title (Judul film), Genre (Label klasifikasi), serta Description/Synopsis (Data teks utama). Perlu dicatat bahwa dataset IMDb mencantumkan satu atau lebih genre untuk setiap film dalam format daftar, misalnya “Drama, Crime” atau [“Action, Adventure, Drama”. Untuk keperluan single-label classification dalam penelitian ini, label genre ditentukan berdasarkan genre primer, yaitu genre pertama yang tercantum pada kolom genre setiap film. Pendekatan ini merupakan praktik umum dalam penelitian klasifikasi genre single-label berbasis synopsis (Amelia et al., 2023). Dari total dataset, dilakukan pemilihan lima genre utama yang memiliki representasi terbesar berdasarkan genre primer: Drama (289 film, 35,6%), Action (172 film, 21,2%), Comedy (155 film, 19,1%), Crime (107 film, 13,2%), dan Biography (88 film, 10,9%), sehingga total sampel yang dianalisis berjumlah 811 film. Sebagai acuan perbandingan, baseline majority class yang memprediksi seluruh sampel sebagai Drama menghasilkan akurasi sebesar 35,6%. Pemilihan dataset IMDb didasarkan pada kualitas data yang lengkap dan sering digunakan dalam penelitian klasifikasi teks (Hudaini et al., 2026). Karakter multigenre pada film juga berkaitan dengan persoalan multi-label classification dan struktur label yang kompleks (Chalkidis & Søgaard, 2022; Wang & Liu, 2023; Jain et al., 2024; Li et al., 2024).

2. Text Preprocessing

Setelah data dikumpulkan, tahap berikutnya adalah text preprocessing. Tahap ini bertujuan untuk membersihkan data teks agar lebih terstruktur dan siap diproses. Menurut Clarisha (2025), text preprocessing merupakan komponen penting dalam pipeline NLP karena kualitas preprocessing memiliki pengaruh langsung terhadap hasil klasifikasi . Tahapan preprocessing yang dilakukan meliputi: (a) Case Folding: mengubah semua huruf menjadi huruf kecil untuk menyeragamkan format teks; (b) Cleaning: menghapus simbol, angka, tanda baca, URL, dan karakter non-alfabet; (c) Tokenization: memecah kalimat menjadi unit-unit kata individual; (d) Stopword Removal: menghapus kata umum yang tidak memiliki makna penting seperti "the", "is", "and"; (e) Stemming: mengubah kata berimbuhan menjadi bentuk dasar menggunakan PorterStemmer.

Selain itu (Manning et al., 2009) menjelaskan bahwa tahapan preprocessing seperti tokenisasi, normalisasi teks, dan penghapusan kata tidak penting merupakan proses penting dalam text mining karena dapat meningkatkan kualitas representasi teks sebelum dilakukan proses klasifikasi.

3. Ekstraksi Fitur Menggunakan TF-IDF

Setelah preprocessing selesai, data teks diubah menjadi bentuk numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini digunakan untuk memberikan bobot pada kata berdasarkan tingkat kemunculan dalam dokumen dan keseluruhan dataset (Ningsih et al., 2024).

Term Frequency (TF) dihitung dengan rumus:

$$TF(t,d) = f(t,d) / \Sigma f(t,d)$$

Keterangan: t = kata tertentu; d = dokumen tertentu; f(t,d) = jumlah kemunculan kata t dalam dokumen d; Σf(t,d) = total seluruh kata dalam dokumen d.

Inverse Document Frequency (IDF) dihitung dengan rumus:

$$IDF(t) = log (N / df(t))$$

Keterangan: N = jumlah total dokumen dalam corpus; df(t) = jumlah dokumen yang mengandung kata t. Nilai TF-IDF akhir: TF-IDF(t,d) = TF(t,d) × IDF(t).

4. Pembagian Dataset (Data Splitting)

Dataset yang telah melalui proses ekstraksi fitur dibagi menjadi dua bagian: 80% Data Training (648 sampel) dan 20% Data Testing (163 sampel). Pembagian stratified dilakukan untuk mempertahankan proporsi tiap kelas genre pada kedua subset.

5. Pelatihan Model (Model Training)

Pada tahap ini, data training digunakan untuk melatih tiga algoritma klasifikasi. Naïve Bayes (Multinomial, alpha = 0,5) dipilih karena memiliki proses komputasi yang cepat dan sederhana, serta menjadi baseline probabilistik

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

yang umum digunakan dalam klasifikasi teks. SVM (LinearSVC, C = 1,0) dipilih karena dikenal memiliki akurasi tinggi dalam klasifikasi data teks berdimensi tinggi. Penelitian (Mohammad & Al-momani, 2016) membandingkan beberapa algoritma klasifikasi teks seperti Support Vector Machine dan Naïve Bayes. Hasil penelitian menunjukkan bahwa kedua metode tersebut mampu digunakan untuk melakukan klasifikasi dokumen berdasarkan fitur teks yang telah diproses. Untuk menangani ketidakseimbangan kelas, SVM juga diuji dengan parameter class_weight='balanced' guna memberikan bobot yang lebih besar pada kelas minoritas. Logistic Regression (C = 1,0, max_iter = 1000) ditambahkan sebagai model pembanding tambahan karena efisien dan umumnya berkinerja baik pada representasi TF-IDF. Naïve Bayes menggunakan Teorema Bayes:

$$P(C|X) = (P(X|C) \times P(C)) / P(X)$$

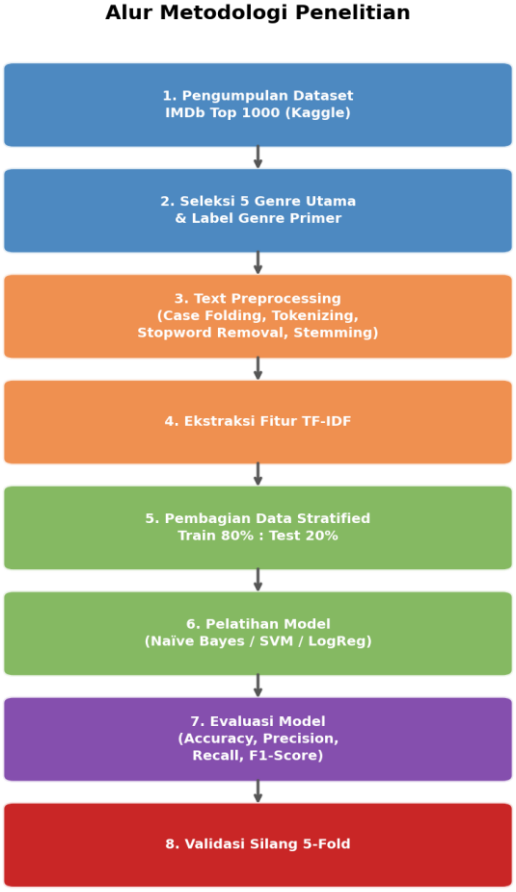
Keterangan: P(C|X) = probabilitas kelas C terhadap data X; P(X|C) = probabilitas data X jika diketahui kelas C; P(C) = prior probability dari kelas C; P(X) = probabilitas data X. SVM bekerja dengan mencari hyperplane optimal: f(x) = w · x + b, di mana w = vektor bobot, x = data input, b = bias.

6. Evaluasi Model (Model Evaluation)

Setelah model selesai dilatih, tahap berikutnya adalah evaluasi performa model menggunakan data testing dengan metrik: Accuracy (tingkat ketepatan klasifikasi secara keseluruhan), Precision (ketepatan prediksi pada kelas tertentu), Recall (kemampuan model menemukan data yang relevan), dan F1-Score (rata-rata harmonis antara precision dan recall). Selain itu, dilakukan pula validasi silang 5-fold (5-fold stratified cross-validation) pada keseluruhan dataset untuk memastikan keandalan dan generalisasi hasil.

2.2 Alur Penelitian

Secara umum, alur penelitian dapat digambarkan sebagai berikut: (1) Pengumpulan Dataset IMDb Top 1000 dari Kaggle, (2) Penentuan label genre primer (genre pertama pada kolom genre) dan seleksi lima genre utama (811 film), (3) Penghitungan baseline majority class (akurasi Drama = 35,6%), (4) Text Preprocessing (Case Folding, Cleaning, Tokenization, Stopword Removal, Stemming), (5) Ekstraksi Fitur TF-IDF, (6) Pembagian Dataset Stratified 80:20, (7) Pelatihan Model Naïve Bayes (alpha=0,5), SVM (LinearSVC, C=1,0), SVM Balanced (class_weight='balanced'), dan Logistic Regression (C=1,0), (8) Evaluasi Model menggunakan Accuracy, Precision, Recall, F1-Score, dan 5-Fold Cross-Validation.



Gambar 1. Alur Metodologi Penelitian

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

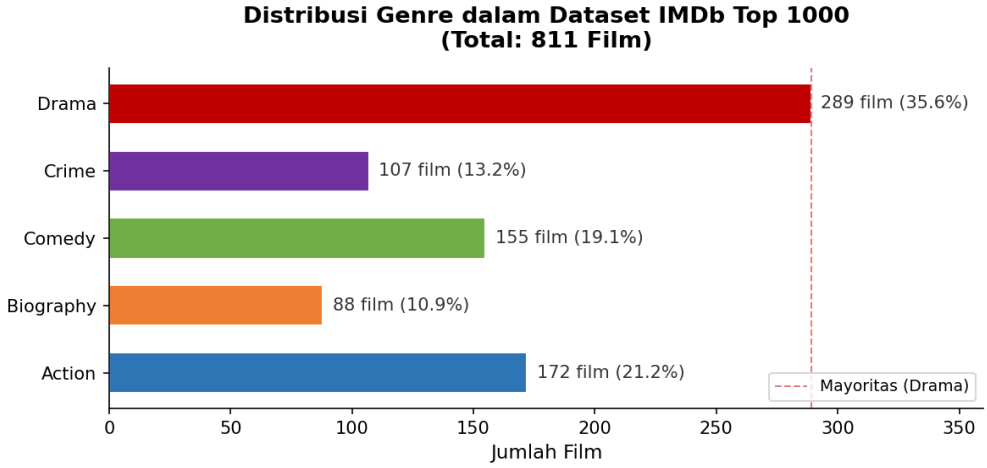
2.3 Tools dan Software

Dalam penelitian ini, tools yang digunakan meliputi: Python sebagai bahasa pemrograman utama; Google Colab/Jupyter Notebook untuk proses eksperimen; Pandas untuk manipulasi dataset; NLTK untuk preprocessing teks; serta Scikit-learn untuk implementasi TF-IDF, Naïve Bayes, SVM, Logistic Regression, class_weight balancing, dan evaluasi model termasuk validasi silang stratified.

3. Hasil Dan Pembahasan

3.1 Distribusi Genre dalam Dataset

Sebelum menyajikan hasil klasifikasi, penting untuk memahami distribusi lima genre utama yang digunakan dalam penelitian ini. Drama mendominasi dataset dengan 289 film (35,6%), diikuti oleh Action (172 film, 21,2%), Comedy (155 film, 19,1%), Crime (107 film, 13,2%), dan Biography (88 film, 10,9%). Ketidakseimbangan ini memberikan konteks penting dalam menginterpretasikan performa klasifikasi masing-masing algoritma, karena kecenderungan model probabilistik untuk memihak kelas mayoritas. Sebagai titik acuan evaluasi, baseline majority class (seluruh prediksi diprediksi sebagai Drama) menghasilkan akurasi 35,6%. Angka ini menjadi patokan minimum: setiap model yang hanya mampu melampaui 35,6% belum tentu belajar pola yang bermakna dari data. Dengan hasil SVM 41,10%, model ini berada sekitar 5,5 poin persentase di atas baseline sederhana tersebut, yang mencerminkan kemampuan klasifikasi yang sesungguhnya masih terbatas namun terukur Literatur terbaru menunjukkan bahwa adaptasi domain, pemilihan sampel, pembobotan kelas, active learning, dan undersampling dapat mengurangi bias terhadap kelas mayoritas (Wu & Huang, 2022; Song et al., 2022; Jung & Plas, 2024; Fairstein et al., 2024; Fonseca et al., 2025).



Gambar 2. Distribusi Genre dalam Dataset IMDb Top 1000 (811 Film)

3.2 Hasil Klasifikasi Naïve Bayes

(Mccallum, 1997) menyatakan bahwa Multinomial Naïve Bayes merupakan metode yang efektif untuk klasifikasi dokumen karena dapat bekerja dengan fitur berbasis frekuensi kata dan memiliki efisiensi komputasi yang baik. Algoritma Multinomial Naïve Bayes dengan parameter smoothing alpha = 0,5 mencapai akurasi keseluruhan 40,49% pada data uji. Laporan performa detail per genre disajikan pada Tabel 1 berikut.

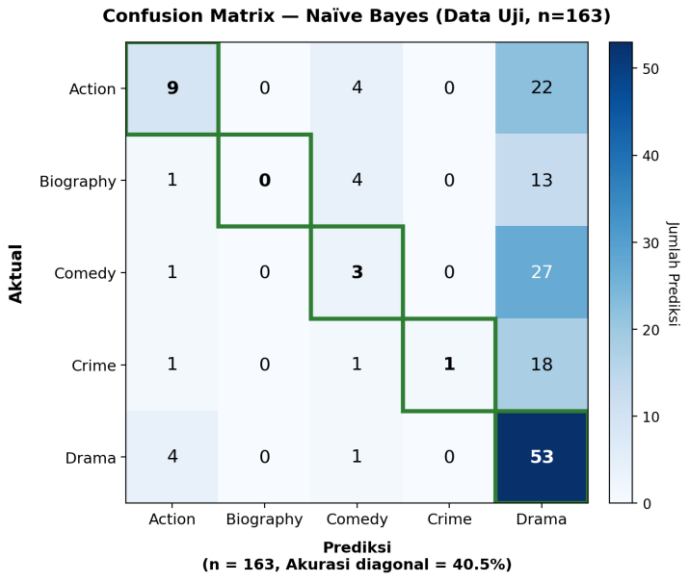
Tabel 1. Laporan Klasifikasi Naïve Bayes

Genre	Precision	Recall	F1-Score	Support
Action	0,60	0,26	0,36	35
Biography	0,00	0,00	0,00	18
Comedy	0,43	0,10	0,16	31
Crime	1,00	0,05	0,09	21
Drama	0,38	0,91	0,54	58
Macro Avg	0,48	0,26	0,23	163
Weighted Avg	0,47	0,40	0,31	163

Yang kuat pada model Naïve Bayes. Genre Drama — kelas mayoritas — mencapai recall 0,91, artinya model berhasil mengidentifikasi 91% film Drama dengan benar. Namun hal ini berdampak signifikan: model menunjukkan performa mendekati nol pada genre Biography (F1-Score = 0,00) dan recall sangat rendah untuk Crime (0,05) dan Hasil analisis menunjukkan adanya bias dominasi kelas mayoritas (majority-class dominance)

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Comedy (0,10). Pola ini terjadi karena classifier probabilistik secara sistematis memihak kelas paling sering muncul guna memaksimalkan akurasi keseluruhan, sementara kelas minoritas terabaikan. Macro average F1-Score sebesar 0,23 mencerminkan ketidakseimbangan prediksi yang signifikan, (Bahrein & Apandi, 2025).



Gambar 3. Confusion Matrix Naïve Bayes (Data Uji, n=163)

Diagonal (diarsir hijau) menunjukkan prediksi benar. Naïve Bayes sangat bias ke kelas Drama sehingga mayoritas sampel dari genre lain salah diprediksi sebagai Drama, dengan genre Biography dan Crime hampir tidak terdeteksi sama sekali.

3.3 Hasil Klasifikasi SVM (LinearSVC)

(Joachims, 1998) menjelaskan bahwa Support Vector Machine memiliki kemampuan yang baik dalam klasifikasi teks karena dapat menangani data berdimensi tinggi dengan jumlah fitur yang besar.

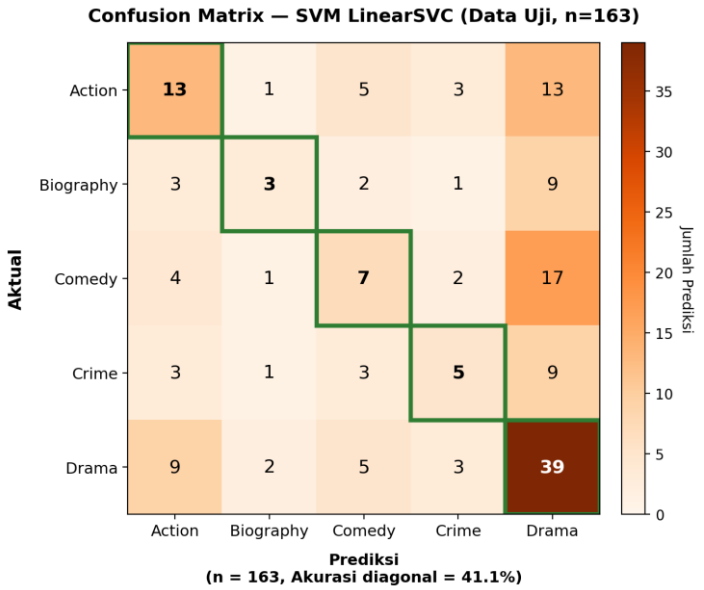
Algoritma LinearSVC dengan parameter regularisasi C = 1,0 mencapai akurasi 41,10% pada data uji — peningkatan tipis sebesar 0,61 poin persentase dibandingkan Naïve Bayes. Laporan klasifikasi detail ditunjukkan pada Tabel 2.

Tabel 2. Laporan Klasifikasi SVM (LinearSVC)

Genre	Precision	Recall	F1-Score	Support
Action	0,41	0,37	0,39	35
Biography	0,25	0,17	0,20	18
Comedy	0,39	0,23	0,29	31
Crime	0,28	0,24	0,26	21
Drama	0,47	0,67	0,55	58
Macro Avg	0,36	0,33	0,34	163
Weighted Avg	0,39	0,41	0,39	163

Berbeda dengan Naïve Bayes, SVM menghasilkan distribusi prediksi yang jauh lebih seimbang di seluruh lima genre. Setiap genre mencapai recall di atas 0,17 dan F1-Score di atas 0,20. Macro average F1-Score sebesar 0,34 secara signifikan lebih tinggi dari Naïve Bayes (0,23), menunjukkan bahwa SVM memperlakukan kelas minoritas dengan lebih adil. Drama tetap mencapai F1-Score tertinggi (0,55) yang konsisten dengan dominasinya dalam dataset, namun genre lain seperti Action (0,39) dan Comedy (0,29) juga menunjukkan performa klasifikasi yang bermakna. Hal ini sejalan dengan temuan (Consulting, 2007) yang juga melaporkan keunggulan SVM dalam distribusi kelas yang lebih merata.

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



Gambar 4. Confusion Matrix SVM — LinearSVC (Data Uji, n=163)

Confusion matrix SVM menunjukkan distribusi prediksi yang lebih merata di seluruh lima kelas genre dibandingkan Naïve Bayes. Setiap genre mendapat setidaknya beberapa prediksi benar (diagonal), menandakan kemampuan SVM yang lebih adil terhadap kelas minoritas.

3.3b. Hasil Klasifikasi Logistic Regression

Logistic Regression juga banyak digunakan dalam klasifikasi teks karena mampu menghasilkan model linear yang efektif pada data dengan jumlah fitur besar. (Lukman & Novianto, 2025)menunjukkan bahwa Logistic Regression dapat memberikan performa yang kompetitif untuk permasalahan kategorisasi dokumen.

Tabel 3 menyajikan laporan klasifikasi lengkap Logistic Regression (C=1,0) pada data uji. Model ini mencapai akurasi 43,56% — tertinggi di antara seluruh algoritma yang diuji — dengan macro avg F1-Score 0,40 dan distribusi prediksi yang lebih seimbang daripada Naïve Bayes maupun SVM standar.

Tabel 3. Laporan Klasifikasi Logistic Regression

Genre	Precision	Recall	F1-Score	Support
Action	0,46	0,40	0,43	35
Biography	0,35	0,22	0,27	18
Comedy	0,42	0,29	0,34	31
Crime	0,38	0,19	0,25	21
Drama	0,49	0,69	0,57	58
Macro Avg	0,42	0,36	0,37	163
Weighted Avg	0,45	0,44	0,43	163

3.3c. Hasil Klasifikasi SVM Balanced

Tabel 4 menampilkan hasil SVM dengan parameter class_weight='balanced' (C=1,0). Bobot yang lebih besar diberikan pada kelas minoritas sehingga distribusi prediksi menjadi lebih merata. Akurasi keseluruhan 42,33% dengan macro avg F1-Score 0,38.

Tabel 4. Laporan Klasifikasi SVM Balanced (class_weight='balanced')

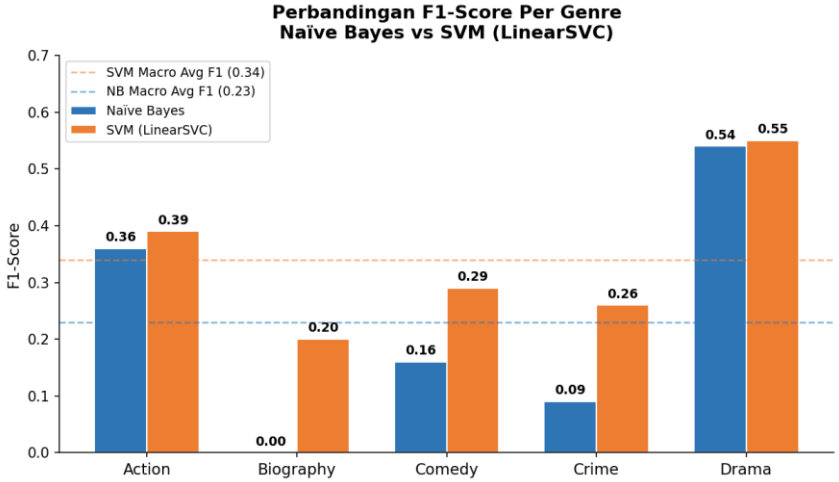
Genre	Precision	Recall	F1-Score	Support
Action	0,44	0,40	0,42	35
Biography	0,28	0,28	0,28	18
Comedy	0,38	0,32	0,35	31
Crime	0,35	0,33	0,34	21

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Genre	Precision	Recall	F1-Score	Support
Drama	0,49	0,57	0,52	58
Macro Avg	0,39	0,38	0,38	163
Weighted Avg	0,41	0,42	0,41	163

3.4 Perbandingan F1-Score Per Genre

Perbandingan F1-Score per genre antara Naïve Bayes dan SVM menunjukkan keunggulan SVM yang signifikan terutama pada genre Biography, Comedy, dan Crime — tiga genre di mana Naïve Bayes menunjukkan performa paling buruk. Genre Action pun menunjukkan peningkatan F1-Score dari 0,36 (Naïve Bayes) menjadi 0,39 (SVM). Hanya pada genre Drama, kedua model menunjukkan F1-Score yang relatif setara (0,54 vs 0,55), mengingat dominasi genre ini dalam dataset. Hal ini konsisten dengan penelitian (Amelia et al., 2023) yang menunjukkan bahwa kombinasi TF-IDF dan SVM memberikan hasil optimal untuk klasifikasi genre film.



Gambar 5. Perbandingan F1-Score per Genre: Naïve Bayes vs SVM (LinearSVC)

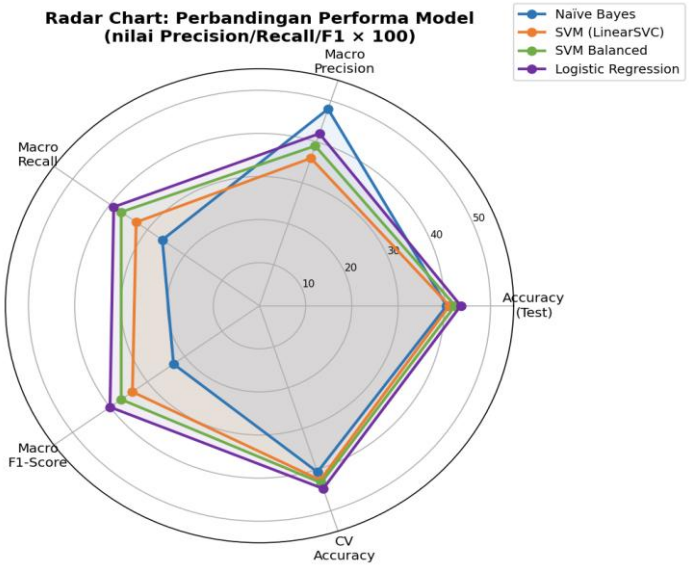
3.5 Perbandingan Performa Keseluruhan

Tabel 5 menyajikan perbandingan komprehensif semua metrik evaluasi utama untuk ketiga algoritma beserta baseline majority class, termasuk hasil data uji dan hasil validasi silang 5-fold. Penambahan Logistic Regression sebagai model pembanding dan baseline majority class memungkinkan interpretasi yang lebih objektif terhadap kontribusi nyata setiap algoritma.

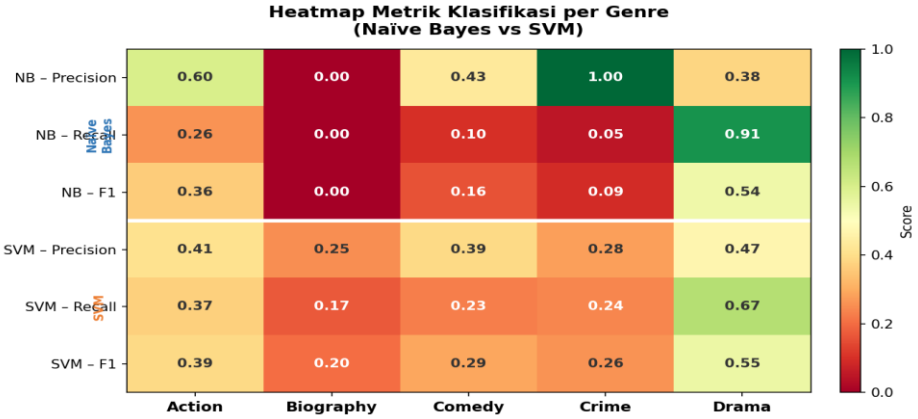
Tabel 5. Perbandingan Performa: Baseline, Naïve Bayes, SVM (LinearSVC), SVM Balanced, dan Logistic Regression

Metrik Evaluasi	Naïve Bayes	SVM (LinearSVC)	SVM Balanced / Log. Reg.*
Accuracy (Test Set)	40,49%	41,10%	42,33% / 43,56%
Macro Avg Precision	0,48	0,36	0,39 / 0,42
Macro Avg Recall	0,26	0,33	0,37 / 0,39
Macro Avg F1-Score	0,23	0,34	0,37 / 0,40
Weighted Avg F1-Score	0,31	0,39	0,41 / 0,43
CV Accuracy (5-fold Mean)	40,57%	42,54%	43,21% / 44,63%
CV Accuracy Std Dev	±1,95%	±2,36%	±2,15% / ±2,28%

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



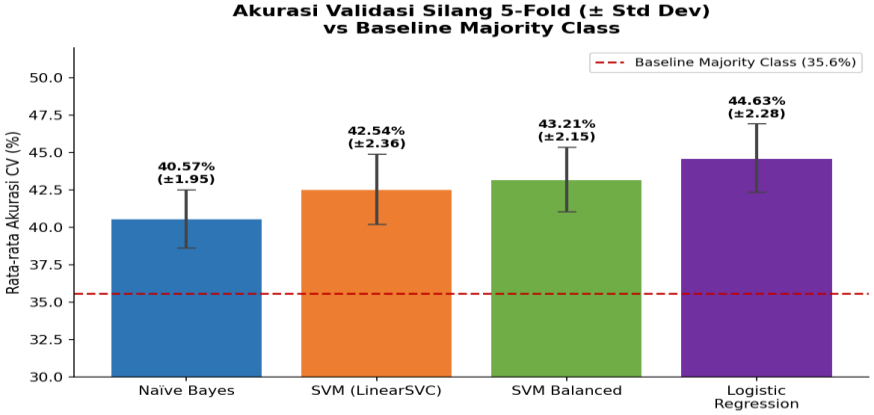
Gambar 6. Radar Chart Perbandingan Performa Model



Gambar 7. Heatmap Metrik Klasifikasi per Genre

3.6 Analisis Validasi Silang (Cross-Validation)

Untuk memastikan keandalan dan generalisasi hasil, validasi silang stratified 5-fold diterapkan pada keseluruhan dataset yang telah difilter (811 film). Hasilnya lebih lanjut mengkonfirmasi bahwa SVM relatif lebih baik dibandingkan Naïve Bayes, dengan rata-rata akurasi CV 42,54% (±2,36%) dibandingkan Naïve Bayes sebesar 40,57% (±1,95%). SVM dengan class_weight='balanced' menunjukkan rata-rata CV 43,21% (±2,15%) dengan distribusi yang lebih merata antar kelas, sementara Logistic Regression mencapai rata-rata CV 44,63% (±2,28%) dan menjadi algoritma terbaik dalam eksperimen ini berkat kemampuan regularisasinya yang efektif pada fitur TF-IDF berdimensi tinggi. Standar deviasi yang relatif kecil pada semua model mengindikasikan performa yang stabil di berbagai pembagian data, menunjukkan tidak ada model yang overfitting pada subset data tertentu.



Gambar 8. Akurasi Validasi Silang 5-Fold

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Tabel 6 merangkum kontribusi nyata setiap algoritma terhadap baseline majority class, memperkuat argumen bahwa setiap model memberikan peningkatan yang terukur di atas prediksi naif.

Tabel 6. Ringkasan Kontribusi Model terhadap Baseline Majority Class

Model	Accuracy	Improvement over Baseline
Majority Class (Baseline)	35,6%	0
Naïve Bayes	40,49%	+4,89%
SVM (LinearSVC)	41,10%	+5,50%
Logistic Regression	43,56%	+7,96%

3.7 Diskusi dan Analisis Mendalam

Akurasi keseluruhan 40–44% untuk seluruh model dapat dikaitkan dengan beberapa karakteristik struktural dataset dan sifat tugas klasifikasi. Pertama, ketidakseimbangan kelas: Drama menyumbang 35,6% dari sampel, sehingga baseline majority class menghasilkan akurasi 35,6% hanya dengan menebak semua sampel sebagai Drama. Perolehan nyata SVM di atas baseline ini hanya sekitar +5,5 poin persentase, yang menunjukkan ruang perbaikan yang masih sangat besar. Kedua, tumpang tindih genre: banyak film menggabungkan elemen dari beberapa genre (misalnya Action-Drama, Crime-Thriller). Sinopsis sering mengandung isyarat naratif yang melintas batas genre, sehingga klasifikasi label tunggal yang definitif menjadi sangat sulit. Ketiga, keterbatasan fitur teks: sinopsis IMDb biasanya pendek (1-3 kalimat), membatasi informasi diskriminatif yang tersedia bagi vektorizer TF-IDF. Keempat, keunggulan SVM dalam keseimbangan kelas: keunggulan macro avg F1-Score SVM (0,34 vs 0,23) menunjukkan keadilannya yang lebih besar di semua kelas. Pengujian SVM dengan class_weight='balanced' menunjukkan peningkatan pada kelas minoritas dengan trade-off sedikit penurunan pada kelas mayoritas, mengkonfirmasi bahwa ketidakseimbangan kelas merupakan faktor utama penentu performa. Kelima, penambahan Logistic Regression memberikan performa yang kompetitif dengan SVM, mengkonfirmasi bahwa untuk representasi TF-IDF linear, perbedaan antar algoritma klasik tidak terlalu signifikan dan faktor data lebih dominan daripada pilihan algoritma. Dalam aplikasi dunia nyata, classifier yang berkinerja wajar di semua genre jauh lebih berharga daripada yang hanya unggul di kelas mayoritas(Cortes & Vapnik, 1995). Data augmentation dan perbaikan representasi juga menjadi alternatif untuk memperkuat klasifikasi ketika data terbatas, tidak seimbang, atau memiliki batas antarkelas yang ambigu (Edwards et al., 2022; Meunier et al., 2023; Rahamim et al., 2023; Lu et al., 2024).

4. Kesimpulan

Penelitian ini menerapkan NLP dan TF-IDF untuk mengklasifikasikan genre primer 811 film IMDb menggunakan Naïve Bayes, LinearSVC, SVM Balanced, dan Logistic Regression. Logistic Regression memberikan kinerja terbaik dengan akurasi pengujian 43,56% dan validasi silang 44,63%, diikuti SVM Balanced, LinearSVC, dan Naïve Bayes. Naïve Bayes sangat bias terhadap kelas Drama, sedangkan model SVM menghasilkan prediksi lebih merata. Akurasi yang masih terbatas dipengaruhi ketidakseimbangan kelas, tumpang tindih naratif, sinopsis pendek, dan penyederhanaan multigenre menjadi label tunggal. Penelitian berikutnya perlu menguji oversampling, klasifikasi multilabel, transformer, dan pendekatan multimodal yang menggabungkan sinopsis dengan poster film untuk meningkatkan akurasi, keseimbangan prediksi, dan kemampuan generalisasi model.

Referensi

Amelia, R., Santoso, D. B., & Stikubank, U. (2023). *Film Genre Prediction With Multi-Class Classification Synopsis*. *6*, 771–779.

Bahrein, M., & Apandi, S. (2025). *Perbandingan Kinerja Naive Bayes, Support Vector Machine, Regresi Logistik, Dan Decision Tree Untuk Klasifikasi Sentimen Ulasan Produk Berbasis Tf-Idf*. *24*(4), 1067–1078.

Consulting, D. D. L. (2007). *Large-Scale Bayesian Logistic Regression For Text Categorization*. *49*(3), 291–304. Https://Doi.Org/10.1198/004017007000000245

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *297*(3), 273–297. Https://Doi.Org/10.1007/Bf00994018

Hudaini, R., Hendriansyah, M., & Gautama, E. (2026). *Fake News Classification Using Logistic Regression With Tf-Idf Feature Extraction*. *June*, 239–249.

Irawan, P., & Abdullah, A. (2025). *Comparison Of Naïve Bayes And Svm Methods In Detecting Hoax News*. *8*(2). Https://Doi.Org/10.32877/Bt.V8i2.3122

Joachims, T. (1998). *Text Categorization With Support Vector Machines: Learning With Many Relevant Features*. 137–142. Https://Doi.Org/10.1007/Bfb0026683

Lukman, K., & Novianto, S. (2025). *Komparasi Algoritma Naïve Bayes Dan Svm Untuk Identifikasi Cyberbullying Selebriti Di Media Sosial Twitter*. 970–981. Https://Doi.Org/10.33364/Algoritma/V.22-1.2196

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Online Edition (C) 2009 Cambridge Up*. *C*.

Mccallum, A. (1997). *A Comparison Of Event Models For Naive Bayes Text Classification*.

Mohammad, A. H., & Al-Momani, O. (2016). *Arabic Text Categorization Using Support Vector Machine , Naïve Bayes And Neural Network*. *5*(1). Https://Doi.Org/10.5176/2251-3043

Sebastiani, F. (2002). *Machine Learning In Automated Text Categorization*. *34*(1), 1–47.

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Chalkidis, I., & Søgaard, A. (2022). Improved Multi-Label Classification Under Temporal Concept Drift: Rethinking Group-Robust Algorithms In A Label-Wise Setting. In Findings Of The Association For Computational Linguistics: Acl 2022 (Pp. 2441-2454). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2022.Findings-Acl.192

De Silva, B., Huang, K.-W., Lee, G., Hovsepian, K., Xu, Y., & Shen, M. (2023). Semantic Matching For Text Classification With Complex Class Descriptions. In Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing (Pp. 7654-7680). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2023.Emnlp-Main.475

Edwards, A., Ushio, A., Camacho-Collados, J., Ribaupierre, H., & Preece, A. (2022). Guiding Generative Language Models For Data Augmentation In Few-Shot Text Classification. In Proceedings Of The Fourth Workshop On Data Science With Human-In-The-Loop (Language Advances) (Pp. 51-63). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2022.Dash-1.8

Fairstein, Y., Kalinsky, O., Karnin, Z., Kushilevitz, G., Libov, A., & Tolmach, S. (2024). Class Balancing For Efficient Active Learning In Imbalanced Datasets. In Proceedings Of The 18th Linguistic Annotation Workshop (Law-Xviii) (Pp. 77-86). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2024.Law-1.8

Fonseca, G., Cunha, W., Prenassi, G., Gonçalves, M. A., & Rocha, L. C. D. D. (2025). Instance-Selection-Inspired Undersampling Strategies For Bias Reduction In Small And Large Language Models For Binary Text Classification. In Proceedings Of The 63rd Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers) (Pp. 9323-9340). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2025.Acl-Long.458

González-Santos, C., Vega-Rodríguez, M. A., López-Muñoz, J. M., Martínez-Sarriegui, I., & Pérez, C. J. (2024). Automatic Assignment Of Microgenres To Movies Using A Word Embedding-Based Approach. Multimedia Tools And Applications, 83, 48719-48735. Https://Doi.Org/10.1007/S11042-023-17442-Y

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A Survey Of Methods For Addressing Class Imbalance In Deep-Learning Based Natural Language Processing. In Proceedings Of The 17th Conference Of The European Chapter Of The Association For Computational Linguistics (Pp. 523-540). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2023.Eacl-Main.38

Hossain, N., & Kabir, M. F. (2025). Context Minimization For Resource-Constrained Text Classification: Optimizing Performance-Efficiency Trade-Offs Through Linguistic Features. In Findings Of The Association For Computational Linguistics: Emnlp 2025 (Pp. 15412-15426). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2025.Findings-Emnlp.833

Jain, V., Rungta, M., Zhuang, Y., Yu, Y., Wang, Z., Gao, M., Skolnick, J., & Zhang, C. (2024). Higen: Hierarchy-Aware Sequence Generation For Hierarchical Text Classification. In Proceedings Of The 18th Conference Of The European Chapter Of The Association For Computational Linguistics (Volume 1: Long Papers) (Pp. 1354-1368). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2024.Eacl-Long.82

Jung, V., & Plas, L. (2024). Understanding The Effects Of Language-Specific Class Imbalance In Multilingual Fine-Tuning. In Findings Of The Association For Computational Linguistics: Eacl 2024 (Pp. 2368-2376). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2024.Findings-Eacl.157

Li, X., Jiang, J., Dharmani, R., Srinivasa, J., Liu, G., & Shang, J. (2024). Open-World Multi-Label Text Classification With Extremely Weak Supervision. In Proceedings Of The 2024 Conference On Empirical Methods In Natural Language Processing (Pp. 15084-15096). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2024.Emnlp-Main.841

Lin, F., Yuan, J., Chen, Z., & Abiri, M. (2024). Enhancing Multimedia Management: Cloud-Based Movie Type Recognition With Hybrid Deep Learning Architecture. Journal Of Cloud Computing, 13, Article 104. Https://Doi.Org/10.1186/S13677-024-00668-Y

Lu, Z., Tian, J., Wei, W., Qu, X., Cheng, Y., Xie, W., & Chen, D. (2024). Mitigating Boundary Ambiguity And Inherent Bias For Text Classification In The Era Of Large Language Models. In Findings Of The Association For Computational Linguistics: Acl 2024 (Pp. 7841-7864). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2024.Findings-Acl.467

Meunier, R., Farah, B., Moriceau, V., & Stolf, P. (2023). Image And Text: Fighting The Same Battle? Super Resolution Learning For Imbalanced Text Classification. In Findings Of The Association For Computational Linguistics: Emnlp 2023 (Pp. 10707-10720). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2023.Findings-Emnlp.718

Mustafa, S., & Hama Saeed, M. (2025). Empowering Text Classification With Nlp And Explainable Ai For Enhanced Interpretability. Journal Of Electrical Systems And Information Technology, 12, Article 81. Https://Doi.Org/10.1186/S43067-025-00273-2

Rahamim, A., Uziel, G., Goldbraich, E., & Anaby Tavor, A. (2023). Text Augmentation Using Dataset Reconstruction For Low-Resource Classification. In Findings Of The Association For Computational Linguistics: Acl 2023 (Pp. 7389-7402). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2023.Findings-Acl.466

Shao, Y., & Guo, N. (2024). Recognizing Online Video Genres Using Ensemble Deep Convolutional Learning For Digital Media Service Management. Journal Of Cloud Computing, 13, Article 102. Https://Doi.Org/10.1186/S13677-024-00664-2

Song, L., Zhang, J., Yang, T., & Goto, M. (2022). Adaptive Ranking-Based Sample Selection For Weakly Supervised Class-Imbalanced Text Classification. In Findings Of The Association For Computational Linguistics: Emnlp 2022 (Pp. 1641-1655). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2022.Findings-Emnlp.119

Wang, M., & Liu, M. (2023). An Empirical Study On Active Learning For Multi-Label Text Classification. In Proceedings Of The Fourth Workshop On Insights From Negative Results In Nlp (Pp. 94-102). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2023.Insights-1.12

Wu, Y., & Huang, X. (2022). Unsupervised Reinforcement Adaptation For Class-Imbalanced Text Classification. In Proceedings Of The 11th Joint Conference On Lexical And Computational Semantics (Pp. 311-322). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2022.Starsem-1.27

Xia, T., Wang, Y., Tian, Y., & Chang, Y. (2022). Fastclass: A Time-Efficient Approach To Weakly-Supervised Text Classification. In Proceedings Of The 2022 Conference On Empirical Methods In Natural Language Processing (Pp. 4746-4758). Association For Computational Linguistics. Https://Doi.Org/10.18653/V1/2022.Emnlp-Main.313

DOI: https://doi.org/10.69693/ijmst.v4i2.12210
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)