1436 MULTI LABEL KLASIFIKASI GENRE FILM BERDASARKAN SINOPSIS MENGGUNAKAN METODE LONG SHORT-TERM MEMORY (LSTM)

By Jihadul Akbar

WORD COUNT 3798
TIME SUBMITTED 03-JAN-2025 02:22PM
PAPER ID 113844651

MULTI LABEL KLASIFIKASI GENRE FILM BERDASARKAN SINOPSIS MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY (LSTM)*

Jihadul Akbar[1],

Abstract

Film is a medium of entertainment that can be enjoyed by many people, not only as entertainment but also as a means of marketing, trade, and education. Genre is one of the important characteristics of a film. Therefore, genre classification is a way to identify relationships among films, making it easier for viewers to find films that suit their preferences. Genre classification can be very comprehensive or diverse based on certain criteria, as many films may include multiple genres within them. To address this issue, researchers propose a multi-label classification of film genres based on synopses using the Long Short Term Memory (LSTM) algorithm and comparing the performance of Word2Vec, GloVe, and FastText word embeddings. The proposed LSTM model architecture consists of several layers, namely Embedding, SpatialDropout1D, LSTM, and Dense layers. The values for each layer are 300, 0.5, 128, and 18, respectively. Testing was conducted using three scenarios, and the GloVe word embedding was found to perform better than Word2Vec and FastText in all three testing scenarios. The F1-score of the GloVe word embedding outperformed Word2Vec and FastText with scores of 0.603, 0.591, and 0.580, respectively.

Keywords : *Multilabel Classification, LSTM, Word Embedding, Optuna, TPE*

Abstrak

Film merupakan sarana hiburan yang dapat dinikmati oleh banyak orang, bukan hanya sebagai hiburan tetapi juga merupakan sarana pemasaran, perdagangan dan pendidikan. Genre merupakan salah satu karakteristik penting dari sebuah film. Oleh sebab itu klasifikasi genre merupakan cara untuk menemukan hubungan dari masing-masing film sehingga memudahkan penonton untuk menemukan film yang sesuai. Klasifikasi genre film mungkin sangat komprehensif atau beragam berdasarkan kriteria, ada banyak genre yang serupa dalam satu film mungkin termasuk beberapa genre di dalamnya. Untuk menyelesaikan masalah tersebut peneliti mengusulkan klasifikasi multilabel genre film berdasarkan sinopsis menggunakan algoritma *Long Short Term Memory* (LSTM) dan membandingkan kinerja *word embedding Word2Vec*, *GloVe* dan *FastText*. Arsitektur model *LSTM* yang diusulkan terdiri dari beberapa layer yakni layer *Embedding, SpastialDropout1d*, LSTM dan Dese. Nilai dari masing-masing layer yakni 300, 0.5, 128, dan 18. Pengujian dilakukan dengan tiga skenario pengujian didapatkan *word embedding GloVe* mendapatkan hasil lebih baik di banding Word2Vec dan FastText pada tiga skenario pengujian. Hasil *f1-score word embedding GloVe* mengungguli *Word2Vec* dan *FastText* dengan nilai 0.603, 0.591 dan 0.580.

Kata kunci : *Klasifikasi Multilabel, LSTM, Word Embedding, Optuna, TPE*

---

1. PENDAHULUAN

Kemajuan teknologi informasi khususnya perkembangan internet di Indonesia menciptakan era digital dimana informasi, komunikasi, hiburan, bahkan kebutuhan sehari-hari dapat diakses dengan mudah. Jumlah masyarakat Indonesia yang terhubung ke internet terus meningkat. Berdasarkan hasil Polling Indonesia yang bekerja sama dengan Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pada tahun 2024 jumlah pengguna internet di Indonesia sudah mencapai 221,56 juta jiwa atau setara dengan 79,5% dari total penduduk Indonesia. Kemajuan teknologi dapat berdampak pada industri hiburan termasuk industri film.

Film merupakan karya kreasi manusia yang berisikan unsur estetika tinggi sebagai media komunikasi yang menyalurkan dan menyebarkan pesan kepada publik. Film digunakan sebagai sarana hiburan yang menyajikan cerita, peristiwa, musik, drama, humor dan sajian teknis lainnya yang dapat dinikmati oleh banyak orang. Selain sebagai hiburan film juga merupakan sarana pemasaran, perdagangan dan dimanfaatkan juga dalam pendidikan[1]. Film dapat dibedakan menjadi beberapa kategori atau lebih dikenal

dengan sebutan genre film. Genre merupakan salah satu karakteristik paling penting dari sebuah film menceritakan keseluruhan isi film. Fungsi dari genre yakni untuk memudahkan dalam klasifikasi dan membagi film dari seluruh film yang pernah diproduksi [2]. Jenis genre film yang ditayangkan di platform *streaming* secara online.

Streaming merupakan cara tak terbatas untuk mendistribusikan dan mengonsumsi konten media. Layanan *streaming* digital semakin umum di berbagai segmen industri media, seperti Youtube, Netflix, dan Spotify. Netflix merupakan salah satu penyedia layanan *streaming* video yang populer dan pada tahun 2015 hampir 100 layanan *streaming* berbeda yang ada di Amerika Serikat [3]. Pencarian adalah masalah utama yang menjadi perhatian pengguna, karena sulitnya menemukan film yang sesuai dengan keinginan penonton. Oleh sebab itu klasifikasi genre atau pengelompokan film merupakan cara untuk menemukan hubungan dari masing-masing film sehingga memudahkan penonton untuk menemukan film yang sesuai harapan.

Klasifikasi genre film mungkin sangat komprehensif atau beragam berdasarkan kriteria. Selain itu, ada banyak genre yang serupa, dalam satu film mungkin termasuk beberapa genre didalam-Nya, membuat klasifikasi yang akurat menjadi sulit [4]. Untuk mengatasi masalah ini dan melakukan klasifikasi genre secara efisien, banyak penelitian sebelumnya menggunakan *Machine Learning* dan *Deep Learning* untuk melakukan klasifikasi genre film secara otomatis, berdasarkan berbagai data seperti poster film [5],[6], plot *summaries* [7], [8], sinopsis [9], [10], [11] dan trailer [12] yang digunakan secara terpisah atau dalam kombinasi [13]. Namun, penggunaan plot *summaries* film dibatasi oleh fakta bahwa plot *summaries* tersebut hanya mengungkapkan bagian pengantar dari sinopsis dan bukan keseluruhan isi film. Sebaliknya, sinopsis adalah merupakan ringkasan garis besar dari isi sebuah film. Trailer berisi berbagai jenis informasi, seperti bingkai gambar dan audio. Namun, trailer memerlukan kapasitas komputasi yang tinggi, karena ukuran datanya yang besar [6]. Begitu juga poster film merupakan gambar tunggal yang memiliki tingkat variabilitas yang tinggi dan kurangnya pembentukan pola [4].

Deep Learning telah mengungguli model *Machine Learning* pada banyak aplikasi terutama pada pendekatan analisis data tradisional [14]. *Long Short Term Memory* (LSTM) terbukti lebih unggul dari algoritma seperti *Multilayer Perceptron* (MLP), *Support Vector Machine* (SVM) dan *Decision Tree* (DT) [15] pada klasifikasi genre multilabel.

Berdasarkan studi literatur yang dilakukan maka peneliti menggunakan *Long Short Term Memory* (LSTM) yang terbukti unggul pada penelitian sebelumnya, membandingkan kinerja *word embedding* Word2Vec, GloVe dan *Fastext*. Hasil akhir dari penelitian adalah melihat ada pengeruh performa dari LSTM dari *word embedding*.

2. TINJAUAN PUSTAKA

Penelitian tentang klasifikasi genre telah banyak dilakukan sebelumnya, seperti melakukan klasifikasi genre multilabel berdasarkan fitur yang diekstraksi dari sinopsis film. Penelitian ini menggunakan 19 pendekatan ekstraksi fitur yang terpisah, 9 darinya menggunakan teknik *Term Frequency Inverse Document Frequency* (TF-IDF) dan sisanya menggunakan *word embedding* [10]Kumpulan data yang digunakan terdiri dari 12.094 sinopsis dalam bahasa Portugis Brazil yang dikumpulkan dari website TMDb serta dilabeli ke dalam 12 genre. Mereka menggunakan empat pengklasifikasian yang berbeda (*Multilayer Perceptron* (MLP), *Decision Tree*, *Random Forest*, dan *Extra Tree*) Hasil terbaik yakni 54,8 % (*f1-score*) menggunakan *classifier* MLP yang dilatih pada fitur TF-IDF dengan mempertimbangkan *Trigram* dengan dimensi 1.000.

Pada penelitian selanjutnya [11] menambah dataset menjadi 13.394 sinopsis Portugis yang dilabeli menjadi 18 genre. Mereka juga melakukan percobaan dengan versi dataset *over sampling*. Dari percobaan tersebut menghasilkan skor 0,478 (*f1-score*) dengan menggunakan *classifier* berbasis TF-IDF. Sedangkan hasil terbaik untuk dataset yang sudah dilakukan *over sampling* dengan beberapa kombinasi ekstraksi fitur menghasilkan skor rata-rata 0,611 (*f1-score*).

Penelitian lain juga melakukan prediksi genre dan rating film berdasarkan sinopsis dengan menggunakan berbagai bahasa yakni Hindi, Telugu, Tami, Malayalam, Korea, Perancis dan Jepang [16]. Dataset yang digunakan yakni data ulasan multi bahasa yang diambil dari tujuh situs web yang berbeda. Total dataset yang terkumpul sejumlah 14.991 yang dilabel menjadi 9 genre. Hasil terbaik dari percobaan yang dilakukan yakni 91,2% pada bahasa Telugu dengan menggunakan model hibrid dengan tiga imputan yakni *word embedding*, *char embedding* dan *sentence embedding*. *Word Embedding dan char embedding* dimasukkan ke dalam jaringan *Long-Short Term Memory* (LSTM) yang berbeda. Sentence dimasukkan ke jaringan padat yang terhubung penuh. *Input* dari masing-masing yakni 300 dimensi dan *output* 100 dimensi. Nilai *dropout* 0,4 di seluruh model dan melatih model pada 300 *epoch* dengan ukuran *batch* 512.

Penelitian multimodal dengan menggunakan kumpulan data trailer, subtitle, sinopsis dan poster film dari 152.622 judul film dari Movie Database (TMDb) yang dilabeli dengan 18 genre [13]. Evaluasi eksperim2 yang digunakan dalam penelitian ini yakni *Mel Frequency Cepstral Coefficients* (MFCCs), *Statistical Spectrum Descriptor* (SSD), *Local Binary Pattern* (LBP), *Long Short Term Memory* (LSTM), dan *Convolutional Neural Networks* (CNN). Hasil terbaik adalah dengan menggunakan LSTM yang dilatih pada dataset sinopsis dan subtitle film yakni 0,674 (*f1-score*) dan 0,725 (AUC-FR). Penelitian multilabel juga dilakukan menggunakan data poster film, trailer, plot dan menggunakan *deep network* [17]. Menggunakan dataset *Movi Scope* dengan total dataset 5.043 yang dilabeli dengan 13 genre. Ringkasan plot dilakukan pemotongan panjang hingga 3000 kata, kemudian diubah menjadi 1 x 3000 menggunakan *word embedding* GloVe 42B. Hasil terbaik menunjukkan bahwa ringkasan plot sangat informatif dalam memprediksi genre dengan skor 0,631 (*micro* AP).

Klasifikasi multilabel menggunakan pendekatan *ensemble learning* pada masalah klasifikasi genre film menggunakan data tekstual dalam bentuk synopsis [18]. Dataset yang digunakan merupakan dataset *CMU Movie Summary Corpus* dengan memberi 5 label genre. Mereka menggunakan tiga model klasifikasi yakni *Naïve Bayes, Random Forest*, dan *XGBoost*. Ekstraksi fitur pada *Naïve Bayes* menggunakan *bag-of-word* dan sisanya menggunakan *word2vec*. Hasil dari masing-masing model yakni *Naïve Bayes* 0,568 (*micro-f-score*), *Random Forest* 0,576 (*micro-f-score*) dan *XGBoost* 0,652 (*marco-f-score*). Adapun hasil untuk model *ansambel* gabungan dari tiga model menghasilkan nilai 0,673 (*macro-f-score*).

Memprediksi multilabel genre film dengan 19nggunakan tiga model pengklasifikasian yakni *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN) [19]. Menggunakan dataset *CMU Movie Summary Corpus* berisikan 42.306 plot film, karena kumpulan data tidak seimbang sehingga label genre yang paling relevan yang digunakan. Tujuan penelitian ini adalah mencari pengaturan parameter *Doc2Vec* optimal untuk menghasilkan klasifikasi. Metode ANN memberikan nilai terbaik yakni 0,88 (*f1-score*) dengan pengaturan *windows size* 5 dan *vektor size* 200.

Pendekatan klasifikasi genre menggunakan sinopsis film yang menggunakan model *supervised learning* (k-NN dan SVM) dan model *deep learning* (CNN dan RNN) [20]. Dataset yang digunakan adalah *Movie Database* (IMDb) yang terdiri dari 5000 film dari website *Kaggel* dan 10.000 sinopsis yang diunduh dari website *Rotten Tomatoes*, dataset dilabel menjadi 9 genre film. Skor terbaik 80,5 (*accuracy*) dengan menggunakan model *deep learning* RNN dengan LSTM layer dengan ekstraksi fitur menggunakan *Doc2Vec* dengan jumlah *epoch* 40.

Klasifikasi genre *multiclass* menggunakan sinopsis bah18 Indonesia dengan menggunakan algoritma *Support Vector Machine* (SVM), *Multinomial Naïve Bayes* (MNB), dan *Multilayer Perceptron* (MLP) [21]Ekstraksi fitur menggunakan *bag-of-word* dan TF-IDF. Model terbaik dengan skor 45% (*f1-score*) pada gabungan SVM+TF-IDF. Dataset yang digunakan di yakni data dari website IMDb dan website www.*filmindonesia.or.id* sejumlah 1.005 film yang di label menjadi 5 genre film.

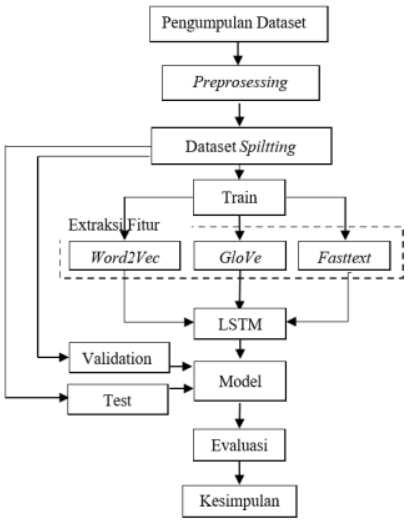
Peneliti melakukan klasifikasi genre berdasarkan plot dan membagi plot menjadi beberapa kalimat. Model klasifikasi yang digunakan yakni *Bidirectional LSTM* (Bi-LSTM) [8]. Total dataset yang dikumpulkan 6.360 plot film dengan 22.278 kalimat yang dilabel menjadi 4 genre film. Teknik ekstraksi fitur yang digunakan adalah *bag-of-word* dan TF-IDF. Hasil terbaik diperoleh dengan Bi-LSTM pada kalimat dengan skor 67,68 (*macro-f1-score*).

Penelitian saat ini adalah klasifikasi 10ltilabel genre film berdasarkan sinopsis menggunakan algoritma *Long Short Term Memory* (LSTM) de7gan tujuan mengetahui performa penerapan *word embedding Word2Vec*, *GloVe* dan *FastText* serta melakukan optimasi *hyperparameter* menggunakan metode *Tree-structured Parzen Estimators* (TPE) dengan *framework* Optuna. Arsitektur model *LSTM* yang diusulkan terdiri dari beberapa layer yakni layer *Embedding, SpastialDropout1d*, LSTM dan Dese.

3. METODOLOGI PENELITIAN

3.1. Tahapan Penelitian

Dalam penelitian ini terdapat beberapa langkah alur penelitian yang digambarkan pada Gambar 1 berikut:



Gambar 1. Alur Penelitian

3.2. Pengumpulan Data Set

Pada tahap ini dilakukan *scraping* data dari *website Internet Movie Databases* (IMDb) untuk mengumpulkan data film seperti *title, year,* genre dan sinopsis serta film yang di produksi antara tahun 1920 sampai dengan tahun 2022. Tabel 1 merupakan hasil scaraping data. Setelah data dikumpulkan selanjutnya disimpan dalam bentuk file *CSV*. Data selanjutnya dilakukan pembersihan data yakni proses untuk menghapus data yang duplikasi dan data yang kosong.

Tabel 1. Sampel Data

title	*year*	*genre*	*synopsis*
Das Cabinet des Dr. Caligari	1920	['Horror', 'Mystery', 'Thriller']	THE CABINET OF DR. CALIGARIFrancis (Friedrich ...
Algol - Tragödie der Macht	1921	['Fantasy', 'Sci-Fi']	Algol, the star devil of ancient mythology, of...
Blade af Satans bog	1921	['Drama']	Carl Theodor Dreyer's classic silent film tell...
Dangerous Days	1920	['Drama']	In the period before World War I, Clayton Spen...

3.3. Preprocessing Dataset

Tahap ini melakukan *case folding, removing punctuation, stopword removal, lemmatization* dan *one-hot-encode*. Proses *case folding* yaitu mengubah semua karakter huruf pada kalimat menjadi huruf kecil. Proses *removing punctuation* yakni menghapus semua karakter yang tidak *valid* seperti angka, tanda baca, simbol dan URL (*Uniform Resource Locator*). Proses *stopword removal* digunakan untuk menghilangkan kata-kata yang tidak penting dan sering muncul seperti *"i, me, my, we, our, you, your, him, his"*. Proses *lemmatization* merupakan proses mengubah kata menjadi kata dasar seperti "*running*" menjadi "*run*". *One hot encode* merupakan metode yang mempresentasikan data kategori menjadi biner yang bernilai 0 dan 1. Genre yang bernilai 0 merupakan genre yang bukan termasuk dalam sinopsis serta genre yang bernilai 1 merupakan genre yang termasuk dalam sinopsis tersebut. Total dataset setelah dilakukan preprosesing data yakni 19.407 baris data.

3.4. *Splitting* data

Splitting data merupakan proses membagi dataset menjadi tiga bagian yakni data *train* yang digunakan untuk melatih model, data *validation* digunakan untuk memvalidasi data train dan data *testing* digunakan untuk menguji seberapa baik model yang telah di latih. Perbandingan data *training, validate* dan *testing* pada penelitian ini yakni 60:20:20.

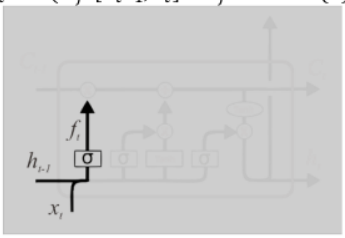
3.5. Ekstraksi Fitur

Salah satu dari teknik ekstraksi fitur adalah *word embedding* yang bekerja dengan cara mengonversi kata yang berupa karakter alfanumerik ke dalam bentuk vektor. Setiap kata dalam *vektor* mempresentasikan sebuah titik pada *space* dengan dimensi tertentu. Dengan menggunakan *word embedding* kata-kata yang sama memiliki makna yang sama akan berada berdekatan. Pada penelitian ini *word embedding* yang digunakan adalah *Word2vec*, *GloVe* dan *FastText*.

3.6. Algoritma LSTM

Langkah pertama dalam algoritma LSTM yakni menentukan informasi apa yang akan di hapus dalam sel saat ini. Informasi tersebut ditentukan oleh gerbang *forget* yang berisikan lapisan *sigmoid*. Dapat dilihat pada gambar 2 Gerbang *forget*, h_{t-1} dan X_t yang berisikan angka antara 0 dan 1 di setiap keadaan sel C_{t-1}. Nilai 1 berarti simpan informasi dalam sel dan nilai 0 hapus informasi dalam sel. Gerbang *forget* dapat dihitung dengan persamaan 1

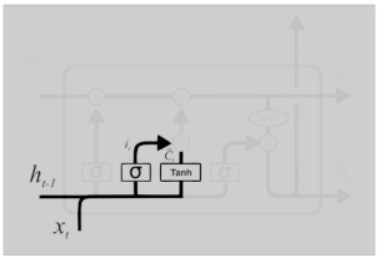
$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f \qquad (1)$$



Gambar 2. Gerbang *forget*

Langkah selanjutnya yakni memutuskan informasi baru apa yang akan disimpan ke dalam sel saat ini. Dalam gerbang *input* terdapat dua bagian yakni pertama berisikan lapisan *sigmoid* yang bertugas untuk menentukan nilai yang akan diperbaharui, kedua merupakan lapisan *than* yang membuat kandidat *vector* baru $\acute{C}_t$ yang akan di tambahkan dalam sel. Selanjutnya akan digabungkan kedua bagian untuk memperbaharui status *input*. Gambar 3 merupakan ilustrasi gerbang *input* dan dapat dihitung dengan persamaan 3.1

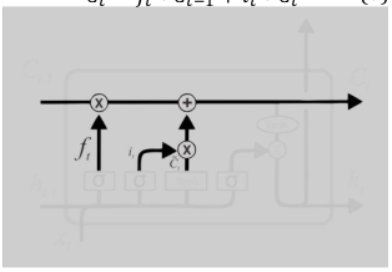
$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i \qquad (2)$$
$$\acute{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c$$



Gambar 3. Gerbang *Input*

Tahap selanjutnya merupakan memperbaharui status sel lama, C_{t-1} ke dalam sel baru C_t. Di mana dalam langkah sebelumnya sudah mendapatkan nilai, sel lama dikalikan dengan f_t serta menghapus informasi pada gerbang *forget* dan menambah $i_t * \acute{C}_t$. Hasil dari operasi tersebut di gunakan untuk memperbaharui nilai sel. Gambar 4 merupakan ilustrasi memperbaharui nilai sel baru dari gerbang *forget* dan *input*. Persamaan 3.2 adalah perhitungannya.

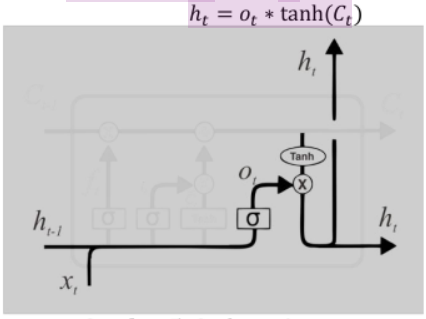
$$C_t = f_t * C_{t-1} + i_t * \acute{C}_t \qquad (3)$$



Gambar 4. Memperbaharui sel dari hasil gerbang *forget* dan *input*.

Gerbang terakhir yakni gerbang *output* yang menghasilkan sel baru (h_t). Gerbang *output* terdiri dari dua yakni lapisan *sigmoid* dan *tanh*. Lapisan *sigmoid* berfungsi untuk menentukan bagian dari sel yang akan dihasilkan. Lapisan *tanh* untuk menentukan nilai antara -1 dan 1. Hasil dari lapisan *sigmoid* dikalikan dengan lapisan *tanh* sehingga menghasilkan *output* baru. Gambar 5 ilustrasi gerbang *output* dan rumus untuk menghitungnya dapat dilihat pada persamaan 4.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o \qquad (4)$$
$$h_t = o_t * \tanh(C_t)$$



Gambar 5. Gerbang *Output*

3.7. Evaluasi

Confusion Matrix adalah salah satu cara paling populer untuk mengevaluasi model klasifikasi. *Confusion matrix* dibuat dengan membandingkan label kelas yang diprediksi dari titik data dengan label kelas yang sebenarnya. Perbandingan ini diulang untuk seluruh kumpulan data dan hasil perbandingan ini dikompilasi dalam format matriks atau tabular [22].

Tabel 2. *Confusion matrix*

		Label yang diprediksi	
		***Negative* (PN)**	***Positive* (PP)**
Kondisi sebenarnya	***Negative* (N)**	***True negative* (TN)**	***False positive* (FP)**
	***Positive* (P)**	***False negative* (FN)**	***True positive* (TP)**

Dari tabel 2 merupakan struktur *confusion matrix*, Secara umum, memiliki kelas positif dan kelas lainnya adalah kelas negatif. Berdasarkan struktur ini, dapat dengan jelas melihat empat istilah penting.

- **True Positive (TP):** Merupakan data positif yang diprediksi benar.
- **True Negative (FN):** Merupakan data negatif yang diprediksi benar.
- **False Positive (FP):** Merupakan data negatif namun diprediksi sebagai data positif.
- ***False Negative (FN):*** Merupakan data positif namun diprediksi sebagai data negatif.

3.7.1 Accuracy

Pada klasifikasi multilabel, *accuracy* didefinisikan dengan persamaan (5) proposisi antara jumlah label yang diprediksi dengan benar dan jumlah total label yang aktif, baik dalam set label asli maupun yang di prediksi. Metrik ini dihitung dengan menggunakan empat ukuran berbeda seperti pada tabel 2.2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

3.7.2 *Precision, Recall* dan *F1-Score*

Precision pada persamaan 6 dianggap sebagai salah satu metrik yang lebih intuitif untuk menilai kinerja prediktif multilabel. Ini dihitung sebagai proporsi antara jumlah label yang diprediksi dengan benar dan jumlah total label yang diprediksi. Dengan demikian, dapat diartikan sebagai persentase label prediksi yang benar-benar relevan untuk instance. Metrik ini biasanya digunakan bersama dengan *Recall* persamaan 7 yang mengembalikan persentase label yang diprediksi dengan benar di antara semua label yang benar-benar relevan. Artinya, rasio label benar diberikan sebagai output oleh *classifier*.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Penggunaan *Precision* dan *Recall* secara bersama-sama sangat umum di bidang pencarian informasi sehingga metrik yang menggabungkannya ditentukan. Ini dikenal sebagai *f1-secore* pada persamaan 8 dan dihitung sebagai rata-rata harmonik dari yang sebelumnya. Dengan cara ini diperoleh ukuran bobot dari berapa banyak label relevan yang diprediksi dan berapa banyak label yang diprediksi relevan.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

3.7.3 AUC-PR (*Area Under the Precision-Recall Curve*)

Area Under Curve (AUC) mengukur seberapa baik sebuah model membedakan antara dua kelas. Kurva dalam AUC adalah kurva ROC, yang merupakan kurva probabilitas yang diplot dengan TP pada sumbu *y* dan FP pada sumbu *x*. Ini adalah ukuran yang baik untuk masalah klasifikasi biner dengan distribusi miring. Rumus untuk AUC sama dengan akurasi Mikro, tetapi AUC mengambil probabilitas prediksi sebagai masukan alih-alih label yang diprediksi. Rumus untuk AUC adalah seperti pada persamaan 9 berikut:

$$AUC = \int_0^1 TPR(FPR^{-1})(x))dx \quad (9)$$

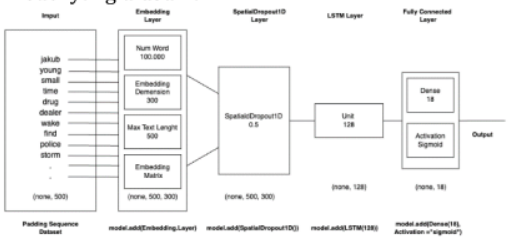
$$TPR(T) = \int_T^\infty f_1(x)dx$$

$$FPR(T) = \int_T^\infty f_0(x)dx$$

f_0 adalah fungsi kepadatan probabilitas untuk kelas negatif, dan f_1 adalah fungsi kepadatan probabilitas untuk kelas positif.

4. HASIL DAN PEMBAHASAN

Arsitektur model yang digunakan dalam penelitian ini diadopsi dari penelitian [23]. Arsitektur model terdiri dari *Embedding layer*, *SpastialDropout1d layer*, *LSTM layer* dan *Dese layer* pada lapisan terakhir. Gambar 6 merupakan detail arsitektur model yang di usulkan. Nilai dari input *sequence* 500, layer *embedding* dengan jumlah *token* 100.000, dimensi *embedding* yakni 300. Layer SpatialDropout1D dengan nilai 0.5 dan 1 layer LSTM berisikan 128 cell, serta layer terakhir yakni layer *output* dengan *dense* bernilai 18 sesuai dengan jumlah *output* dari total genre yang digunakan. *Loss function* menggunakan *binary crossentropy* dan *Optimizer Adam* dengan nilai *learning rate* default yakni 0.001. Ukuran *batch size* menggunakan 32 dan *epoch* sebanyak 200. Dengan menggunakan fungsi *early stopping* untuk menghindari *overfitting* dengan memonitoring *validation loss* serta mencari nilai minimal. Model akan berhenti apabila tidak ada peningkatan *validation loss* selama 5 kali. Selanjutnya diuji coba kan untuk masing-masing *word embedding* untuk menganalisis hasil dari model yang diusulkan.



Gambar 6. Arsitektur Model

4.1 Klasifikasi *word embedding Word2Vec, GloVe* dan *FastText*

1. LSTM *word embeddin* Word2Vec

Hasil model LSTM dengan *word embedding* Word2Vec dapat dilihat pada tabel 3 dalam bentuk *multilabel classification report*. Hasil rata-rata *precision, recall* dan *f1-score* yakni 0.598, 0,584 dan 0.691.

Tabel 3. *Classification report* LSTM Word2Vec

	Precision	Recall	F1-Score	Support
action	0.444	0.723	0.550	573
adventure	0.593	0.453	0.513	559
animation	0.429	0.320	0.367	103
biography	0.383	0.109	0.170	165
comedy	0.685	0.486	0.569	1050
crime	0.603	0.637	0.620	587
documentary	0.971	0.604	0.744	328
drama	0.665	0.787	0.721	1912
family	0.548	0.382	0.450	225
fantasy	0.577	0.322	0.414	301
history	0.338	0.158	0.215	165
horror	0.857	0.697	0.769	644
music	0.763	0.155	0.258	187
mystery	0.415	0.267	0.325	439
romance	0.541	0.443	0.487	723

sci-fi	0.848	0.621	0.717	422
thriller	0.423	0.840	0.563	779
war	0.772	0.602	0.677	304
micro avg	0.598	0.584	0.591	9466
macro avg	0.603	0.478	0.507	9466
weighted avg	0.623	0.584	0.581	9466
samples avg	0.615	0.625	0.578	9466

Hasil perhitungan *score* dari model LSTM *word embedding Word2Vec* dapat dilihat pada Gambar 7.

	LSTM-Word2Vec
Accuracy	0.150
Precision	0.598
Recall	0.584
F1	0.591
ROC AUC	0.761

Gambar 7. Hasil LSTM Word2Vec

2. LSTM *word embeddin Glove*

Hasil model LSTM dengan *word embedding* GloVe yang disajikan dalam *multilabel classification report* dapat dilihat pada tabel 4 dengan hasil rata-rata *precision, recall dan f1-score* yakni 0.627, 0,580 dan 0.603.

Tabel 4. *Classification report* LSTM *GloVe*

	Precision	Recall	F1-Score	Support
Action	0.488	0.646	0.556	573
adventure	0.582	0.485	0.529	559
animation	0.429	0.350	0.385	103
biography	0.370	0.103	0.161	165
comedy	0.659	0.563	0.607	1050
crime	0.662	0.591	0.625	587
documentary	0.948	0.558	0.702	328
drama	0.687	0.749	0.717	1912
family	0.485	0.493	0.489	225
fantasy	0.497	0.502	0.499	301
history	0.367	0.176	0.238	165
horror	0.841	0.775	0.807	644
music	0.778	0.225	0.349	187
mystery	0.486	0.235	0.316	439
romance	0.544	0.393	0.456	723
sci-fi	0.825	0.671	0.740	422
thriller	0.495	0.703	0.581	779
war	0.717	0.641	0.677	304
micro avg	0.627	0.580	0.603	9466
macro avg	0.603	0.492	0.524	9466
weighted avg	0.632	0.580	0.593	9466
samples avg	0.645	0.622	0.592	9466

Hasil perhitungan *get_score* dari model LSTM *word embedding* Glove dapat dilihat pada gambar 8, di mana nilai *accuracy, precision, recall, f1-score* dan ROC AUC yakni 0.171, 0.627, 0.580, 0.603, dan 0.763.

	LSTM-GloVe
Accuracy	0.171
Precision	0.627
Recall	0.580
F1	0.603
ROC AUC	0.763

Gambar 8. Hasil LSTM Glove

3. LSTM *word embeddin* FastText

Hasil model LSTM dengan *word embedding FastText* yang disajikan dalam *multilabel classification report* dapat dilihat pada tabel 5 dengan hasil rata-rata *precision, recall dan f1-score* yakni 0.613, 0,551 dan 0.580.

Tabel 5. *Classification report* LSTM *FastText*

	Precision	Recall	F1-Score	Support
action	0.494	0.656	0.564	573
adventure	0.690	0.295	0.414	559
animation	0.522	0.117	0.190	103
biography	0.438	0.085	0.142	165
comedy	0.679	0.434	0.530	1050
crime	0.590	0.675	0.630	587
documentary	0.858	0.720	0.783	328
drama	0.665	0.806	0.729	1912
family	0.776	0.231	0.356	225
fantasy	0.803	0.203	0.324	301
history	0.444	0.073	0.125	165
horror	0.848	0.756	0.800	644
music	0.730	0.144	0.241	187
mystery	0.491	0.194	0.278	439
romance	0.613	0.297	0.400	723
sci-fi	0.742	0.697	0.719	422
thriller	0.399	0.861	0.546	779
war	0.835	0.365	0.508	304
micro avg	0.613	0.551	0.580	9466
macro avg	0.645	0.423	0.460	9466
weighted avg	0.646	0.551	0.552	9466
samples avg	0.637	0.597	0.574	9466

Hasil perhitungan *get_score* dari model LSTM *word embedding FastText* dapat dilihat pada gambar 9, di mana nilai *accuracy, precision, recall, f1-score* dan ROC AUC yakni 0.157, 0.613, 0.551, 0.580, dan 0.768.

	LSTM-FastText
Accuracy	0.157
Precision	0.613
Recall	0.551
F1	0.580
ROC AUC	0.748

Gambar 9. Hasil LSTM *FastText*

Hasil pengujian dengan menggunakan 22 itektur model dan masing-masing *word embedding* dapat dilihat pada tabel 6. Nilai *f1-score* tertinggi didapatkan pada model yang menggunakan *word embedding* GloVe yakni 0.603 dan nilai AUC tertinggi yakni 0,762. Diikuti oleh *word embedding* Word2Vec dengan nilai *f1-sore* 0,591 dan nilai AUC 0.761. *Word embedding FastText* mendapatkan nilai terendah yakni f1-score 0,580 dan nilai AUC 0,748.

Tabel 6. Hasil model yang di usulkan pada tiga *word embedding*

	Accuracy	Precision	Recall	F1	AUC
LSTM Word2Vec	0,150	0,698	0,584	0,591	0,761
LSTM Glove	0,171	0,627	0,580	**0,603**	**0,763**
LSTM FastText	0,157	0,613	0,551	0,580	0,748

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, maka dapat disimpulkan bahwa hasil *f1-score word embedding GloVe* mengungguli *Word2Vec* dan *FastText* dengan nilai 0.578, 0.571 dan 0.548. Membuktikan bahwa dengan menggunakan *pre-trained word embedding* dapat meningkatkan hasil *f1-score*.

6. UCAPAN TERIMA KASIH

Terima kasih kami ucapkan kepada semua pihak yang telah memberikan kontribusi terhadapt penelitian ini.

1436 MULTI LABEL KLASIFIKASI GENRE FILM BERDASARKAN SINOPSIS MENGGUNAKAN METODE LONG SHORT-TERM MEMORY (LSTM)

ORIGINALITY REPORT

13%
SIMILARITY INDEX

PRIMARY SOURCES

1	Jihadul Akbar, Ema Utami, Ainul Yaqin. "Multi-Label Classification of Film Genres Based on Synopsis Using Support Vector Machine, Logistic Regression and Naïve Bayes Algorithms", 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022 Crossref	49 words — 1%
2	arxiv.org Internet	40 words — 1%
3	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publications	37 words — 1%
4	etheses.uin-malang.ac.id Internet	34 words — 1%
5	Muhammad Arief Rahman, Herman Budianto, Esther Irawati Setiawan. "Aspect Based Sentimen Analysis	33 words — 1%

Opini Publik Pada Instagram dengan Convolutional Neural Network", Journal of Intelligent System and Computation, 2019
Crossref

6	www.dedimulyadir.com Internet	33 words —	1%
7	Rohid Aji Sunan, Halif Fachrizal Erliawan K., Christian Sri Kusuma Aditya. "Klasifikasi Hoax Berita Politik Menggunakan Algoritma Long Short-Term Memory (LSTM) dengan Penambahan Fitur Embedding Global Vector (GloVe)", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2024 Crossref	32 words —	1%
8	medium.com Internet	30 words —	1%
9	Hao Su, Shiwu Yang, Chang Liu, Haiwei Liu. "Research on Named Entity Recognition in Fault Text of Railway Signal Equipment", 2022 14th International Conference on Advanced Computational Intelligence (ICACI), 2022 Crossref	23 words —	1%
10	jurnal.polibatam.ac.id Internet	17 words —	< 1%
11	adoc.pub Internet	16 words —	< 1%
12	docplayer.info Internet	14 words —	< 1%
13	etd.umy.ac.id Internet	14 words —	< 1%
14	repository.usd.ac.id Internet		

14 words — < 1%

15 Salwa Ziada Salsabiila, Helena Nurramdhani Irmanda, Artika Arista. "Comparison of Fasttext and Word2Vec Weighting Techniques for Classification of Multiclass Emotions Using the Conv-LSTM Method", 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), 2023
Crossref
12 words — < 1%

16 www.bartleby.com
Internet
12 words — < 1%

17 jurnal.yudharta.ac.id
Internet
11 words — < 1%

18 www.coursehero.com
Internet
11 words — < 1%

19 www.researchgate.net
Internet
11 words — < 1%

20 123dok.com
Internet
10 words — < 1%

21 Adhi Rahmadian. "Public Sentiment Towards Mandatory Halal Certification: A Large Language Model (LLM) Approach", Likuid Jurnal Ekonomi Industri Halal, 2024
Crossref
10 words — < 1%

22 Yudi Wibisono, Masayu Leylia Khodra. "Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin", INA-Rxiv, 2018
Publications
10 words — < 1%

23	ejurnal.stmik-budidarma.ac.id Internet	10 words — < 1%
24	Azy Mushofy Anwary, Asep ID Hadiana, Puspita Nurul Sabrina. "ANALISIS SENTIMENT PENGGUNAAN VAKSIN COVID-19 MENGGUNAKAN GEO-TAGGED TWEETS DAN ALGORITMA NAIVE BAYES", Informatics and Digital Expert (INDEX), 2021 Crossref	9 words — < 1%
25	Navid NaderiAlizadeh, Rohit Singh. "Aggregating Residue-Level Protein Language Model Embeddings with Optimal Transport", Cold Spring Harbor Laboratory, 2024 Crossref Posted Content	9 words — < 1%
26	ediss.sub.uni-hamburg.de Internet	9 words — < 1%
27	pubs.rsc.org Internet	9 words — < 1%
28	www.alexrichardo.com Internet	8 words — < 1%
29	www.neliti.com Internet	8 words — < 1%
30	Hendri Candra Mayana, Desmarita Leni. "Deteksi Kerusakan Ban Mobil Menggunakan Convolutional Neural Network dengan Arsitektur ResNet-34", Jurnal Surya Teknika, 2023 Crossref	6 words — < 1%
31	Melani Budianta, Manneke Budiman, Abidin Kusno, Mikihiro Moriyama. "Cultural Dynamics in a Globalized World", CRC Press, 2017	6 words — < 1%

Publications

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF