ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

TEXT MINING DAN KLASIFIKASI MULTI LABEL MENGGUNAKAN XGBOOST

Rimbun Siringoringo✉, Jamaluddin, Resianta Perangin-angin
Universitas Methodist Indonesia, Medan, Indonesia
Email: ringo.rimbun@methodist.ac.id

DOI: https://doi.org/10.46880/jmika.Vol6No2.pp234-238

ABSTRACT

The conventional classification process is applied to find a single criterion or label. The multi-label classification process is more complex because a large number of labels results in more classes. Another aspect that must be considered in multi-label classification is the existence of mutual dependencies between data labels. In traditional binary classification, classification analysis only aims to determine the label in the text, whether positive or negative. This method is sub-optimal because the relationship between labels cannot be determined. To overcome the weaknesses of these traditional methods, multi-label classification is one of the solutions in data labeling. With multi-label text classification, it allows the existence of many labels in a document and there is a semantic correlation between these labels. This research performs multi-label classification on research article texts using the ensemble classifier approach, namely XGBoost. Classification performance evaluation is based on several metrics criteria of confusion matrix, accuracy, and f1 score. Model evaluation is also carried out by comparing the performance of XGBoost with Logistic Regression. The results of the study using the train test split and cross-validation obtained an average accuracy of training and testing for Regression Logistics of 0.81, and an average f1 score of 0.47. The average accuracy for XGBoost is 0.88, and the average f1 score is 0.78. The results show that the XGBoost classifier model can be applied to produce a good classification performance.
Keyword: Extreme Gradient Boosting, Logistic Regression, Multi-Label Classification.

ABSTRAK

Proses klasifikasi konvensional diterapkan untuk menemukan satu kriteria atau label. Proses klasifikasi multi-label tergolong lebih kompleks karena jumlah label yang banyak menghasilkan lebih banyak lagi kelas. Aspek lain yang harus dipertibangkan pada lasifikasi multi-label adalah adanya dependesi secara mutual diantara label-label data. Pada klasifikasi tradisional yang bersifat biner, analisis klasifikasi hanya bertujuan untuk menentukan label pada teks, apakah positif atau negatif. Metode ini bersifat sub-optimal karena keterkaitan di antara label tidak dapat ditentukan. Untuk mengatasi kelemahan metode tradisional tersebut, klasifikasi multi label menjadi salah satu solusi dalam pelabelan data. Dengan klasifikasi teks multi label memungkinkan eksistensi dari banyak label sebuah dokumen serta adanya korelasi semantik di antara label-label tersebut. Penelitian ini melakukan klasifikasi multi label pada teks *research article* menggunakan pedekatan *ensemble classifier* yaitu XGBoost. Evaluasi kinerja klasifikasi dididasarkan pada beberapa kriteria metrik *confution matrix*, *accuracy*, dan *f1 score*. Evaluasi model juga dilakukan dengan membandingkan kinerja XGBoost dengan Logistic Regression. Hasil penelitian dengan menggunakan *train test split* dan *cross validation* diperoleh rata-rata accuracy training dan testing untuk Logistik Regresi sebesar 0,81, dan rata-rata f1 score 0,47. Rata-rata accuracy untuk XGBoost adalah 0,88, dan rata-rata f1 score adalah 0,78. Hasil penelitian menunjukkan bahwa model pengklasifikasi XGBoost dapat diterapkan sehingga menghasilkan kinerja klasifikasi yang baik.
***Kata Kunci: Extreme Gradient Boosting, Logistic Regresi,* Klasifikasi Multi-Label.**

PENDAHULUAN

Pemerintah senantiasa melalui badan riset dan inovasi nasional (BRIN) senantiasa berbenah dalam rangkan meningkatkan kualitas dan kuantitas penelitian nasional. Peneliti dan pemangku kepentingan riset berlomba menyediakan sarana publikasi riset dan jurnal dan perhelatan seminar nasional dan internasional. Terdapat ratusan seminar internasional dengan ratusan bidang keilmuan terkait. Salah satu hal penting yang dibutuhkan oleh peneliti adalah memilih bidang Hal ini tertu saja menimbulkan kesulitan bagi peneliti untuk menemukan bidang riset yang sesai dengan bidangnya. Disis lain, riset secara interdisiplin digandrungi oleh peneliti. Sebuah *paper* penelitian dapat memiliki lebih dari satu kriteria. Sebagai contoh, penelitian dengan judul "*The Triple Jump in Problem-Based Learning:*

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Unpacking Principles and Practices in Designing Assessment for Curriculum Alignment" adalah penelitian pedagogik dan manajemen. Penelitian non interdisipliner memiliki hanya satu disipin keilmuan. Misalnya penelitian dengan judul "*Deep Learning for Extreme Multi-label Text Classification*" merupakan topik penelitian untuk ilmu komputer.

PENELITIAN TERKAIT

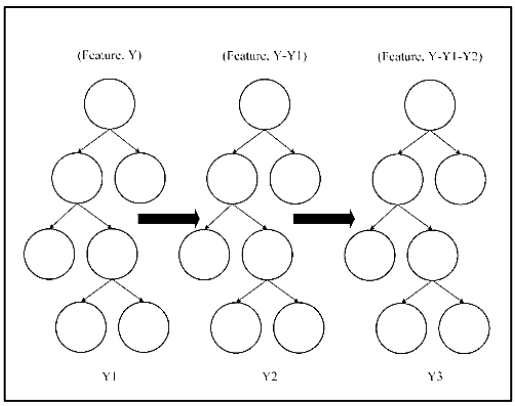
Proses klasifikasi konvensional hanya ditargetkan untuk menemukan satu kriteria atau label. Proses klasifikasi untuk klasifikasi multi-label tergolong lebih kompleks. Jika sebuah data memiliki L label, maka akan ada sebanyak 2^L kemungkinan kelas muti-label yang dihasilkan. Aspek lain yang harus dipertibangkan pada lasifikasi multi-label adalah adanya dependesi secara mutual diantara label-label data. Misalnya ungkapan "*deep learning*" memiliki probabilitas yang besar untuk kemunculan ungkapan "*artificial neural network*"

Pada klasifikasi binary tradisional, analisis klasifikasi hanya bertujuan untuk menentukan label pada teks, apakah positif atau negatif. Metode ini bersifat sub-optimal karena keterkaitan di antara label tidak dapat ditentukan (Chang & Yang, 2017). Untuk mengatasi kelemahan metode tradisional tersebut, klasifikasi multi label menjadi salah satu solusi dalam pelabelan data. Dengan klasifikasi teks multi label memungkinkan eksistensi dari banyak label sebuah dokumen serta adanya korelasi semantik di antara label-label tersebut (Xiao, Huang, Chen, & Jing, 2019).

Saat ini, terdapat beragam pendekatan yang dilakukan oleh para peneliti dalam mengembangkan metode klasifikasi teks multi label pada *research article,* diantaranya adalah dengan pendekatan klasifikasi *ensemble k-nearest* (Wu, Han, Chen, Li, & Zhang, 2022) , *multilayer neural network* (Liu, Wen, Gao, Zheng, & Zheng, 2020)*, graph neural network* (Pal, Selvakumar, & Sankarasubbu, 2020)

Penelitian ini melakukan klasifikasi multi label pada teks *toxic comments* menggunakan pedekatan *ensemble classifier* yaitu dengan mengkombinasikan banyak basis pengklasifikasi. Pendekatan *ensemble* telah terbukti mampu meningkatkan performa klasifikasi pengkalsifikasi tunggal

Terdapat tiga metode utama pada *ensemble classifier*, yaitu teknik *bagging*, teknik *boosting*, dan teknik *stacking*. Salah satu metode yang sangat populer pada teknik *boosting* adalah *Extreme Gradien Boosting* atau disingkat XGBoost. XGBoost merupakan peningkatan dari pendahulunya yaitu *gradient boosting*. Salah satu peningkatan yang dihasilkan pada XGBoost adalah peningkatan konvergensi dan generalisasi (Jiang, Tong, Yin, & Xiong, 2019), (Wang & Guo, 2020).

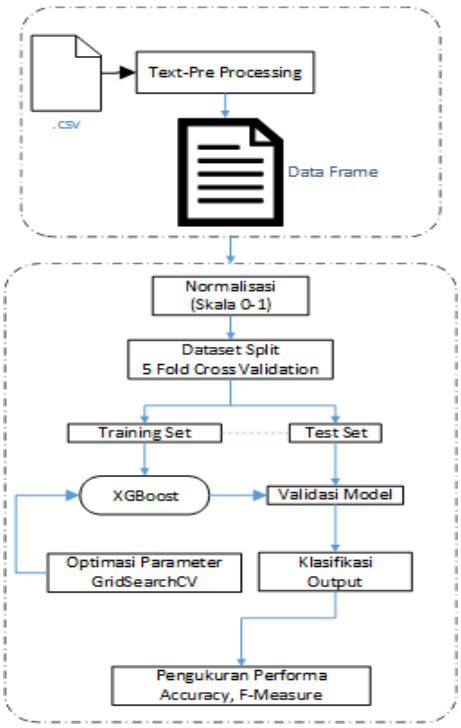


Gambar 1. Pohon Keputusan Pembentuk XGBoost

XGBoost merupakan teknik ensembel dengan menggabungkan beberapa pohon keputusan dasar (Jiang et al., 2019). Dengan teknik ensemble, nilai galat yang dihasilkan pada pohon pertama (*feature*, Y) dapat direduksi ke pohon kedua, dan seterusnya sampai pke pohon selanjutnya (*feature*, |Y-Y1|). Konsep ini dapat ditampilkan pada gambar 1.

METODOLOGI

Tahapan penelitian pada klasifikasi multi label teks menggunakan XGBoost dapat digambarkan melalui gambar 2 berikut ini.



Gambar 2. Prosedur Penyelesaian Masalah

Data

Data yang digunakan pada penelitian ini adalah *research article* data. Data tersebut terdiri dari TITLE dan ABSTRACT artikel ilmiah berbahasa Inggris untuk kategori Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, dan Quantitative Finance. Atribut title dan abstrak merupakan data teks di padukan menjadi satu atribut yaitu text. Pada tabel berikut ini merupakan sampel data yang tampilkan pada data *research article.* Jumlah data per kategori adalah Computer Science (8594), Physics (6013), Mathematics (5618), Statistics (5206), Quantitative Biology (587), Quantitative Finance (249).

Tabel 1. Data Multi Kriteria *Research Article*

Com	Phy	Mat	Sta	Bio	Fin	Text
1	0	0	0	0	0	Reconstructing Subject-Specific Effect Maps
1	0	0	0	0	0	Rotation Invariance Neural Network Rotation
0	0	1	0	0	0	Spherical polyharmonics and Poisson kernels
0	0	1	0	0	0	A finite element approximation for the stochastic
1	0	0	1	0	0	Comparative study of Discrete Wavelet Transformation

Pemrosesan awal

Tahapan pemrosesan awal yang dilakukan adalah Case Folding, Tokenizing, Stop words removal, Stemming dan Lemmatization. Tahapan stemming menerapkan algoritma porter stemmer dan tahapan lemmatization menerapkan wordnet lemmatizer. Hasil akhir pemrosesan awal teks menghasilkan data corpus yang di buat dalam data frame.

TF-IDF

Tahapan ini digunakan untuk menentukan bobot setiap kata pada data teks komentar. Hasil penerapan TF-IDF menghasilkan matriks data dengan dimensi data yang besar. Parameter TF-IDF yang diatur adalah:

Tabel 2. Parameter TF-IDF

Parameter	Nilai	Ket
min_df	3	Jumlah kata minimal
max_features	30000	Jumlah maksimum kata ditampilkan
ngram_range	(1,3)	Batas bawah dan batas range n-values

Parameter XGBoost

Parameter awal untuk Xgboost ditampilkan pada tabel 3. Terdapat tujuh jenis parameter yang di gunakan. Tabel tersebut juga menampilkan batas bawah dan batas atas nilai parameter yang di berikan.

Tabel 3. Format Data Parameter XGBoost

No	Parameter	Nilai (min, max), step
1	*Learning Rate*	(0.01, 1), 2
2	*N Estimators*	(10, 1500), 25
3	*Max Depth*	(1, 10), 1
4	*Min Child Weight*	(0.01, 10.0), 2
5	*Gamma Value*	(0.01, 10.0), 2
6	*Sub Sample*	(0.01, 1.0), 2
7	*Col Sample By Tree*	(0.01, 1.0), 2

HASIL DAN PEMBAHASAN

Pada tabel 4 ditampilkan hasil pemrosesan awal serta label data yang dihasilkan oleh model. Pada data baris ke-lima, Judul "*Comparative study of Discrete Wavelet Transformation"* menghasilkan token [comparative, study, discrete, wavelet, transformation], dan memiliki label sebagai computer science dan mathematic. Hasil premosesan awal terhadap menghasilkan sebanyak 45593 kosa kata unik. Hasil pemrosesan awal teks untuk sampel data dapat disajikan pada gambar 3.

Tabel 3. Hasil Klasifikasi Multi Kriteria

No	Com	Phy	Mat	Sta	Bio	Fin	Teks	Token teks	Target
1	1	0	0	0	0	0	Reconstructing Subject-Specific Effect Maps	[reconstruct, subject, specific, effect, maps]	0
$2	1	0	0	0	0	0	Rotation Invariance Neural Network Rotation	[rotation, invariance, neural, network, rotation]	0

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

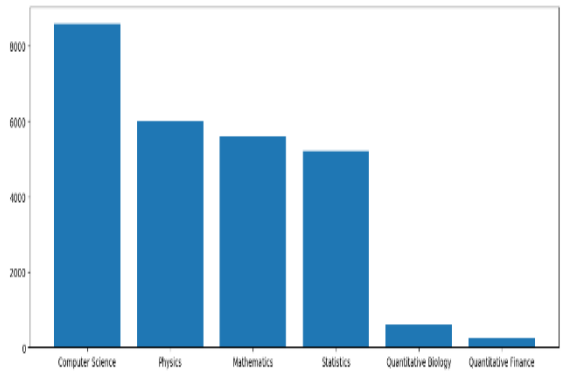
3	0	0	1	0	0	0	Spherical polyharmonics and Poisson kernels	[spherical, polyharmonic, poisson, kernel]	2
4	0	0	1	0	0	0	A finite element approximation for the stochastic	[finite, element, approximation, stochastic]	2
5	1	0	0	1	0	0	Comparative study of Discrete Wavelet Transformation	[comparative, study, discrete, wavelet, transformation]	0

Berikut ini merupakan hasil tokenisasi pada hasil pemrosesan awal data teks:

'spherical', 'polyharmonic', 'poisson', 'kernel',
'polyharmonic', 'function', 'introduce', 'develop', 'notion', 'spherical', 'polyharmonic', 'natural',
'generalisation', 'spherical', 'harmonic', 'particular', 'study', 'theory', 'zonal', 'polyharmonic',
'allow', 'analogously', 'zonal', 'harmonic',
'construct', 'poisson', 'kernel', 'polyharmonic',
'function', 'union', 'rotate', 'ball', 'find', 'representation', 'poisson', 'kernel', 'zonal', 'polyharmonic', 'term',
'gegenbauer', 'polynomial', 'connection', 'classical', 'poisson', 'kernel', 'harmonic',
'function', 'ball', 'poisson', 'kernel', 'polyharmonic', 'function', 'union', 'rotate', 'ball', 'cauchy',
'hua', 'kernel', 'holomorphic', 'function', 'lie'.

Gambar 3. Kosa Kata Unik Pada Teks

Data training terdiri dari sebanyak 20972 data dan data testing 8989 data. Data tersebut tidak mengandung missing value sehingga tidak perlu di lakukan methode penanganan khusus missing value.

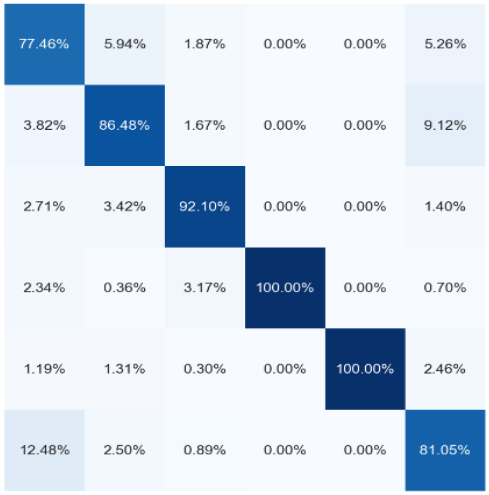


Gambar 4. Grafik Jumlah Data Per Kategori.

Evaluasi model

Klasifikasi teks komentar pada teks artikel di evaluasi menggunakan beberapa model yaitu Logistic Regression, Xgboost, Support Verctor machine. Evalusi di didasarkan pada beberapa kriteria metrik *confution matrix*, *accuracy*, dan *f1 score*. Evaluasi model juga dilakukan dengan membandingkan kinerja XGBoost dengan Logistic Regression dan Support Vector Machine. Pada gambar ditampilkan *confution matrix* untuk model XGBoost

Gambar 5. *Confution Matrix* XGBoost



Gambar 6. *Confution Matrix* Logistic Regression

Selanjutnya hasil *confution matrix* di atas di gunakan untuk menampilkan accuracy dan F1 Score pada tabel berikut.

Tabel 5. Perbandingan Performa Model

		Accuracy	**F1-Score**
LogR	Train	0.83	0.49
	Test	0.80	0.45
XGBoost	Train	0.96	0.95
	Test	0.80	0.60

Pada tabel diatas ditampilkan kinerja model Logistik Regresi dan XGBoost pada klasifikasi. Dengan menggunakan *train test split* dan *cross validation* diperoleh rata-rata accuracy training dan testing untuk Logistik Regresi sebesar 0,81, dan rata-rata f1 score 0,47. Rata-rata accuracy untuk XGBoost adalah 0,88, dan rata-rata f1 score adalah 0,78

KESIMPULAN

Penelitian ini menguji kinerja pengklasifikasi XGBoost dalam mengklasifikasi multi-label data *research article*. Tahapan pemrosesan awal dapat dilakukan dengan baik pada teks berbahasa Inggris sehingga dihasilkan data untuk menguji kinerja model XGBoost. Hasil penelitian menunjukkan bahwa model pengklasifikasi XGBoost dapat diterapkan sehingga menghasilkan kinerja klasifikasi yang baik.

DAFTAR PUSTAKA

Chang, W., & Yang, Y. (2017). Deep Learning for Extreme Multi-label Text Classification, 115–124.

Jiang, Y., Tong, G., Yin, H., & Xiong, N. (2019). A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters. *IEEE Access*, *7*, 118310–118321. https://doi.org/10.1109/access.2019.2936454

Liu, W., Wen, B., Gao, S., Zheng, J., & Zheng, Y. (2020). A multi-label text classification model based on ELMo and attention. *MATEC Web of Conferences*, *309*, 03015. https://doi.org/10.1051/matecconf/202030903015

Pal, A., Selvakumar, M., & Sankarasubbu, M. (2020). Magnet: Multi-label text classification using attention-based graph neural network. *ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, *2*, 494–505. https://doi.org/10.5220/0008940304940505

Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*, *17*(3), 205–221. https://doi.org/10.23919/JCC.2020.03.017

Wu, H., Han, M., Chen, Z., Li, M., & Zhang, X. (2022). A Weighted Ensemble Classification Algorithm Based on Nearest Neighbors for Multi-Label Data Stream. *ACM Trans. Knowl. Discov. Data*. https://doi.org/10.1145/3570960

Xiao, L., Huang, X., Chen, B., & Jing, L. (2019). Label-Specific Document Representation for Multi-Label Text Classification, 466–475.