KLASIFIKASI TEKS BERITA BERBAHASA INDONESIA MENGGUNAKAN MACHINE LEARNING DAN DEEP LEARNING: STUDI LITERATUR

Alfando, Regiolina Hayami
Teknik Informatika, Universitas Muhammadiyah Riau
190401092@student.umri.ac.id, regiolinahayami@umri.ac.id

ABSTRAK

Berita merupakan suatu opini aktual yang menarik dan akurat serta dianggap penting bagi sejumlah besar pembaca, pendengar, maupun penonton. Sebuah dokumen berita seringkali mengacu pada lebih dari satu kategori, sehingga perlu menggunakan metode Klasifikasi yang tidak hanya cepat tetapi juga dapat mengelompokkan teks berita menjadi beberapa kategori. Data yang digunakan sudah terlabel dengan kategori-kategori yang relevan. Data yang digunakan berasal dari berbagai sumber berita dalam bahasa Indonesia. Data tersebut kemudian diproses untuk menghasilkan fitur-fitur yang dapat merepresentasikan teks berita secara numerik, seperti penghitungan kata-kata yang muncul pada teks. Beberapa algoritma *Machine Learning* dan *Deep Learning* yang digunakan dalam penelitian ini adalah *K-Nearest Neighbor (KNN), Multinominal Naive Bayes, Long Short-Term Memory, Multi Layer Perceptron,* dan *Support Vector Machine (SVM)* dan lainnya. Kinerja dari setiap algoritma dievaluasi menggunakan beberapa metrik evaluasi, seperti accuracy, precision, recall, hamming loss dan F1 score. Hasil penelitian menunjukkan bahwa *Support Vector Machine (SVM)* memberikan kinerja yang terbaik diantara algoritma *Machine Learning* lainnya dengan nilai accuracy sebesar 94,24%. Sedangkan pada algoritma *Deep Learning, Long Short-Term Memory* mendapatkan nilai accuracy sebesar 95%. Klasifikasi multi-label pada teks berita bahasa Indonesia dengan menggunakan *Machine Learning* dan *Deep Learning* memiliki potensi untuk diterapkan pada berbagai aplikasi seperti klasifikasi otomatis pada platform berita online atau monitoring isu-isu terkini yang berkaitan dengan topik tertentu.

Kata kunci: *Multi label, Bahasa Indonesia, Berita, Machine Learning, Deep Learning.*

1. PENDAHULUAN

Berita pada umumnya merupakan sebuah informasi yang dialirkan dari banyak sumber, definisinya pun sangat banyak namun dapat ditarik beberapa kesimpulan mengenai makna arti berita. Sering sekali ditemukan judul artikel berita dengan satu topik saja namun di dalam artikel tersebut bisa saja mengandung satu atau lebih topik berita[1]

Permasalahan yang sering muncul adalah penggunaan media digital yang banyak dalam penyampaian informasi menyebabkan jumlah berita digital yang dirilis oleh beberapa portal berita setiap harinya menjadi sangat banyak, dan menyebabkan berita memiliki keterkaitan lebih dari satu kategori.[2]

Dari banyaknya ketersediaan berita yang ada dan memiliki keterkaitan dengan banyak kategori atau multi-label berdampak pada banyaknya berita yang memiliki makna yang sama dan selalu terkait dengan kata yang ambigu, dimana satu obyek berita memiliki sejumlah kelas kategori yang berbeda[2]

Saat ini, metode atau algoritma yang sangat terkenal dalam text categorization adalah *Machine Learning* (Sebastiani dkk., 200). Beberapa algoritma machine learning yang dapat digunakan dalam text categorization diantaranya adalah algoritma *K-Nearest Neighbor*, *Naïve Bayes Classifier*, dan lainnya. Implementasi dari algoritma-algoritma tersebut pada text categorization berdasar pada kemuculan kata atau morfologi kata[3]

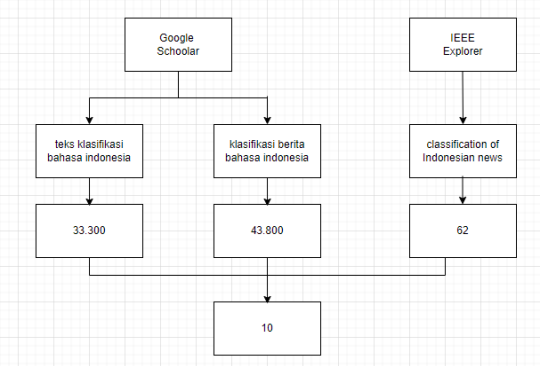
Konsep dari klasifikasi teks dengan menggunakan algoritma-algoritma tersebut adalah memasukkan teks baru yang belum diketahui kategorinya ke dalam kategori dengan melakukan pelatihan terhadap sekumpulan teks yang telah diketahui kategorinya.[3]

Oleh karena itu, dilakukan kajian literatur terhadap topik penelitian text categorization, dimana dikumpulkan metode atau algoritma untuk menganalisis teks berita dalam bahasa Indonesia. Bahasa Indonesia dipilih sebagai subjek tinjauan pustaka karena bahasa ini merupakan bahasa utama penulis dan studi kasusnya dari Bahasa Indonesia. Selain itu, kajian literatur ini dilakukan untuk mengkaji pemilihan dataset yang digunakan dalam studi dan kinerja masing-masing algoritma yang digunakan dalam studi.

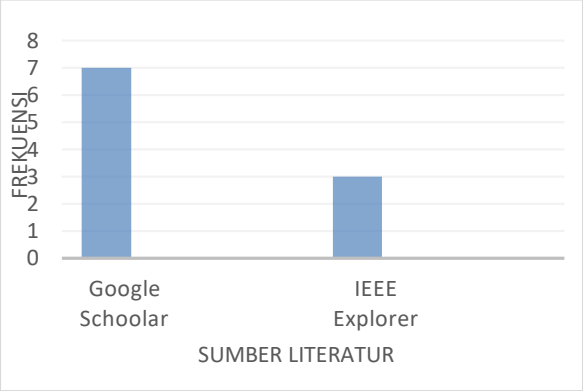
2. TINJAUAN PUSTAKA

Sebelum melakukan kajian literatur, terlebih dahulu disusun koleksi literatur. Untuk melihat studi sebelumnya, perlu mengumpulkan literatur untuk ditinjau. Selain itu, dapat dilihat dari literatur apa saja yang telah dilakukan dan bagaimana hasilnya, sehingga kedepannya dapat diketahui posisi penelitian tentang klasifikasi teks berita melalui metode *Machine Learning* dan *Deep Learning*.

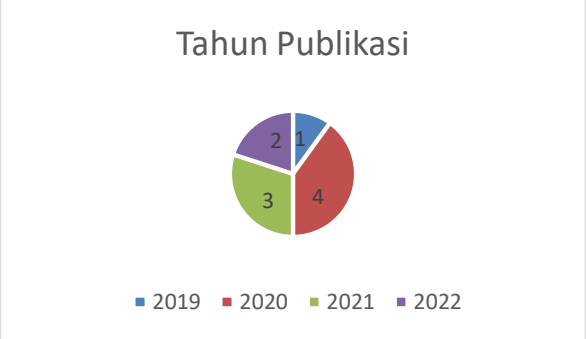
Strategi utama adalah menggunakan mesin pencari dari Google Scholar (https://scholar. google.co.id/) dan IEEE Xplore (https://ieeexplore.ieee.org/).



Gambar 1. Hasil Seleksi Literatur



Gambar 2. Perbandingan jumlah literatur yang diperoleh dari setiap mesin pencarian



Gambar 3. Perbandingan tahun publikasi pada literatur yang di-reviu.

Tabel 1. Literatur-literatur yang direview

Penulis	Judul
Lalu Gias Irham, dkk, [4]	Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine
Kristian Indradiarta G, dkk [5]	Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification pada Dokumen Berita Bahasa Indonesia
Sudianto, dkk [6]	Penerepan Algoritma Support Vector Machine dan Multi-Layer Perceptron Pada Klasifikasi topik Berita
Yudi Widhiyasana, dkk [7]	Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia
Muhammad Alif I, dkk [3]	Pengembangan Model Klasifikasi Dokumen Artikel Teks Berita
	Olahraga dan Bukan Olahraga dalam Bahasa Indonesia Menggunakan Algoritma Support Vector Machine
Winda Kurnia Sari, dkk[8]	Klasifikasi Teks Multilabel pada Artikel Berita Menggunakan Long Short-Term Memory dengan Word2Vec
Alfredo Gormantara, dkk [9]	Klasifikasi Kategori Dan Pelabelan Berita Bahasa Indonesia Menggunakan Mutual Information Dan K-Nearest Neighbors
Nur Ghaniaviyanto R [10]	Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor
Andi Yulia M, dkk [11]	Penerapan Algoritma K-Nearest Neighbor Pada Pengklasifikasian Dokumen Berita Online
Erwin Yudi H, dkk [12]	Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier

Dua kata kunci diuji di Google Scholar, yaitu "Teks Klasifikasi Berita Indonesia" dan "Klasifikasi Berita Indonesia". Maksud dari kedua kata kunci tersebut adalah agar hasil dari search engine yang diberikan mengarah pada literatur berbahasa Indonesia. Sebaliknya, IEEE Xplore menggunakan kata kunci tunggal, yaitu "Klasifikasi Berita Indonesia". Karena IEEE Xplore mengumpulkan literatur dalam bahasa Inggris. Kata kunci tersebut ditujukan agar hasil search engine berupa objek penelitian berupa teks bahasa Indonesia. Hasil pencarian ditunjukkan pada Gambar 1

Pada Gambar 1, dapat dilihat bahwa total makalah yang didapat dari kedua mesin pencarian adalah 77.162 literatur. Kemudian dilakukan filtering terhadap hasil return teratas. Filtering dilakukan dengan melihat objek dari penelitian pada literatur. Reviu dilakukan pada literatur dengan objek penelitian berupa teks pada berita berbahasa indonesia. Selain itu dilihat tahun publikasi dari literatur yang muncul pada mesin pencarian. Hasil dari filtering ini mendapatkan 10 literatur yang dirasa paling relevan untuk direviu. Adapun untuk perbandingan literatur yang diambil berdasarkan sumber mesin pencariannya dapat dilihat pada Gambar 2.

Grafik pada Gambar 2 memperlihatkan jumlah literatur yang diperoleh dari setiap mesin pencarian. Literatur dari hasil mesin pencarian pada IEEE Xplore dan Google Scholar memiliki jumlah 7 literatur dan IEEE Xpoler 3 literatur

Banyaknya hasil return dari mesin pencarian pada Google Scholar membuat filterisasi hanya dilakukan pada 10 halaman awal saja. Kemudian penentuan mengenai literatur mana saja yang akan direviu pada kedua mesin pencarian mempertimbangkan tahun publikasi. Literatur yang diambil adalah yang memiliki tahun publikasi di atas tahun 2019. Perbandingan jumlah tahun publikasi dari literatur yang digunakan dapat dilihat pada Gambar 3.

Pada Gambar 3 dapat dilihat bahwa keseluruhan literatur yang digunakan memiliki tahun publikasi 4 tahun ke belakang. Literatur yang paling tua didapatkan dengan tahun publikasi 2019. Sedangkan literatur yang dipublikasikan pada tahun 2020 adalah yang terbanyak dengan jumlah 4 literatur. Adapun mengenai detail judul dan penulis dari literatur yang akan digunakan sebagai bahan reviu dapat dilihat pada Tabel 1.

Selanjutnya literatur-literatur tersebut direviu untuk mendapatkan gambaran mengenai isi dari setiap literatur. Adapun fokus dari faktor yang akan diteliti kali ini ada 4 yaitu tujuan pembangunan model, penggunaan dataset, penerapan teknik pre-processing dan performa dari metode yang digunakan.

3. METODE PENELITIAN

Melihat isi dari setiap literatur yang diulas, penulis dapat membaginya menjadi 4 bagian. Pertama, menyangkut tujuan dari semua literatur. Kedua, menyangkut sumber dataset yang digunakan. Ketiga, menyangkut teknik pre-processing teks yang digunakan. Kemudian yang terakhir adalah tentang kinerja yang dihasilkan oleh masing-masing penelitian.

3.1. Tujuan Literatur

Berdasarkan penelitian yang dilakukan oleh peneliti, tujuan umum masing-masing literatur adalah:

1. Mengklasifikasikan teks berita yang memiliki lebih dari 1 kategori topik dengan metode algoritma yang ada di *Machine Learning* dan *Deep Learning*
2. Menganalisis dan meniliti hasil masing-masing pada tiap-tiap algoritma

3.2. Dataset

Reviu dilakukan dengan cara difokuskan pada literatur yang melakukan pengujian untuk melakukan membangun model dengan input atau bahan training dan testing berupa teks dari berita berbahasa indonesia. Setiap literatur mempunyai cara tersendiri untuk membangun dataset yang digunakan. Penulis secara garis besar dapat mengelompokkan 3 jenis dataset yang digunakan pada 10 literatur yang diperoleh. Adapun sebaran mengenai penggunaan jenis dataset disajikan dalam Tabel 3.

Tabel 3. Sumber dataset untuk penelitian

No	Jenis	Literatur
1	Portal Berita	[5][7][8][6][10][12][11]
2	Web Scraping	[9][3]
3	Dataset dari peneliti sebelumnya	[4]

Tabel 4. Sample data pada dataset multi label [2]

Data	Politik	Ekonomi	Teknologi
Ketua Tim Kampanye Nasional (TKN) Jokowi-Ma'ruf, Erick Thohir optimistis, capres petahana akan moncer hadapi Prabowo di debat keempat.	1	0	0
Perdagangan pasar saham Indonesia hari ini berpeluang untuk berbalik arah menguat	0	1	0
Perangkat storage Dell EMC Data Domain dan Integrated Data Protection Appliance (IDPA) mendapat pembaruan	0	0	1

3.3. Text Pre Processing

Text Preprocessing memiliki peran yang sangat penting dalam sistem Natural Language Processing (NLP) karena tahap ini bertanggung jawab untuk mengidentifikasi karakter, kata, dan kalimat yang menjadi unit dasar yang akan diolah pada seluruh tahap pemrosesan selanjutnya, mulai dari analisis morfologi hingga penandaan ucapan.[3] Kegiatan ini memungkinkan dokumen teks untuk diproses lebih lanjut, termasuk dalam aplikasi seperti pencarian informasi dan sistem terjemahan mesin. Dalam banyak kasus, dokumen teks juga memerlukan tahapan preprocessing untuk menghilangkan format khusus seperti angka dan tanggal, serta kata-kata umum seperti preposisi, artikel, dan kata benda yang tidak relevan untuk klasifikasi teks.

Preprocessing adalah langkah umum dalam klasifikasi teks[13]. Sasaran dalam fase ini adalah mengolah data yang hanya berupa teks menjadi informasi yang siap untuk diklasifikasi. Adapun preprocessing yang digunakan pada literatur ini adalah punctual removal, case folding, tokenization, stopword removal dan stemming.

Tabel 5. Teknik pre-processing pada teks

No	Jenis	Literatur
1	*Case Folding*	[4][5][6][7]][9][11][12]
2	*Stemming*	[4][5][7][3][9][11][12]
3	*Stop word Removal*	[4][5][7][13][9]
4	*Word Filtering*	[11][12]
5	*Tokenization*	[4][5][7][3][8][9][11][12]
6	*Cleaning*	[4][6][7][3][8][10]

Tabel 5 memperlihatkan bahwa *Tokenization* menjadi teknik yang paling sering digunakan yaitu pada 8 literatur. Kemudian untuk teknik yang dikombinasikan, *Tokenization, Case Folding, Cleaning, dan Stemming* menjadi kombinasi teknik yang paling sering digunakan.

3.3.1 Case Folding

Pada tahap ini, seluruh karakter pada dokumen akan diubah menjadi huruf kecil. Hal ini dilakukan dengan tujuan untuk mengurangi variasi kata dalam artikel berita.[6]

Tabel 6. Contoh proses Case Folding

Input	Garuda Indonesia
Output	garuda indonesia

3.4. *Stemming*

Teknik Stemming merupakan teknik yang diterapkan untuk mendapatkan kata dasar dari suatu kata. Dengan kata lain, teknik ini dilakukan untuk mendapatkan kata dasar dari teks yang berimbuhan. Tujuan dari dilakukannya teknik ini adalah agar ketika pembobotan dilakukan, kata-kata yang memiliki kata dasar sama akan mendapat bobot yang sama[14]

Tabel 7. Contoh proses Stemming Data

Input	garuda indonesia membeberkan alasan tarif tiket pesawat mahal penyebab utamanya perang harga tahun maskapai harga cost akibatnya kinerja keuangan berdarah darah alias rugi
Output	garuda indonesia beber alas tarif tiket pesawat mahal sebab utama perang harga tahun maskapai harga cost akibat kerja uang darah darah alias rugi

3.5. *Stop word removal*

Stop word merupakan kata yang sangat umum dan keberadaannya tidak mengubah makna dari kalimat. Sesuai dengan namanya yang mengandung kata "removal", teknik ini akan menghilangkan kata-kata yang tidak penting dalam kalimat [13].

Tabel 8. Contoh proses Stop word removal

Input	Garuda Indonesia maskapai dari Indonesia
Output	Garuda Indonesia maskapai Indonesia

3.6. *Tokenization*

Dalam tokenisasi, kalimat dipecah menjadi unit leksikal yang lebih kecil yang disebut token. Tujuan dari tokenisasi adalah eksplorasi kata-kata dalam sebuah kalimat.[10] Tokenisasi biasanya digunakan pada tahap awal berbagai proses Natural Language Processing (NLP) seperti pemodelan teks, pencocokan pola, dan klasifikasi. Tokenisasi mengenali token seperti kata, frasa, simbol, dan emotikon, yang kemudian dapat digunakan untuk membuat model teks.

Tabel 9. Contoh proses Tokenization

Input	Garuda Indonesia maskapai Indonesia
Output	[Garuda, Indonesia, maskapai, Indonesia]

Tabel 10. Jenis metode/algoritma Machine Learning dan Deep Learning yang digunakan

No	Metode/Algoritma	Literatur
1	Support Vector Machine	[4][5][6][3]
2	K-Nearest Neighbor	[9][10][11]
3	Naïve Bayes Classifier	[12]
4	Long Short-Term Memory	[8]
5	Convolutional Long Short-Term Memory	[7]
6	Multi Layer Perceptron	[6]

3.7. Metode yang digunakan

Dilihat dari algoritma yang digunakan pada setiap literatur, ada 2 hal yang dapat diambil dari literatur-literatur tersebut. Pertama, algoritma apa saja yang digunakan dalam penelitian di setiap literatur. Kedua, mengenai teknis yang dilakukan pada penelitian setiap literatur, apakah melakukan perbandingan terhadap lebih dari satu algoritma atau melakukan optimalisasi hanya pada satu algoritma saja. Adapun sebaran dari metode yang digunakan oleh tiap literatur disajikan dalam Tabel 10.

Dari Tabel 10, dapat dilihat bahwa algoritma yang banyak digunakan algoritma Support Vector Machine dan K-Nearest Neighbor. Satu literatur dapat memuat lebih dari satu metode/algoritma. Hal ini berkaitan dengan teknis dari penelitian yang dilakukan berupa perbandingan metode yang digunakan.

3.8. *Support Vector Machine* (SVM)

Penelitian menggunakan SVM mendapatkan akurasi dengan rentang 74% - 94.24% [4][5][6][3] . Menggunakan SVM dengan kernel linear sebagai metode klasifikasi dan menggunakan *Mutual Information* dan *TF IDF* sebagai seleksi fiturnya dengan akurasi 94.24% [4]

3.9. *K-Nearest Neighbor (KNN)*

Penelitian menggunakan *K-Nearest Neighbor (KNN)* mendapatkan akurasi 89,2%[10] dan 89,9[11] dengan proses klasifikasi dokumen K=7 dan menghitung jarak kedekatan dengan *Cosine Similarity*[11]. Dengan menggunakan mesin vektor pendukung (support vector machine) dengan margin besar (hyperplane) sebagai dasarnya dan menggunakan seleksi fitur *Mutual Information*[9] model ini dapat digunakan dengan hasil akurasi yang baik.

3.10. *Naïve Bayes Classifier*

Algoritma Enhanced Confix Stripping Stemmer dan Naive Bayes digunakan dalam proses klasifikasi penelitian ini. Sebanyak 600 data digunakan sebagai data pelatihan, dengan 150 data digunakan untuk kategori Olahraga, Teknologi, Ekonomi, dan lain-lain. Proses preprocessing data selama pelatihan dan pengujian memastikan bahwa dataset dapat diklasifikasikan. Prosedur preprocessing meliputi Case Folding, Tokenizing, Filtering, dan Stemming menggunakan algoritma Enhanced Confix Stipping Stemmer.

Klasifikasi dokumen berita yang dilakukan menghasilkan akurasi untuk kategori Olahraga, Teknologi, Ekonomi, dan Lain-lain berturut-turut sebesar 90%, 90%, 100%, dan 100%. Rata-rata akurasi dari empat kategori tersebut mencapai 95%. Nilai ini diperoleh berkat peran algoritma Enhanced Confix Stripping Stemmer yang berguna, dan metode klasifikasi Naïve Bayes Classifier yang terbukti handal.[12]

3.11. *Long Short-Term Memory*

Pada penelitian ini percobaan trial dan error dilakukan untuk klasifikasi teks menggunakan LSTM. Untuk mendapatkan hasil yang optimal klasifikasi teks menggunakan LSTM dengan fitur Word2Vec melakukan tuning hyper-parameter. Dalam pengujian ini, fitur Word2Vec dan Word2vec digunakan sebagai masukan untuk LSTM, dengan berbagai hyperparameter seperti fungsi aktivasi softmax dalam keluaran, fungsi aktivasi Relu dan Tanh, fungsi loss categorical crossentropy, learning rate 0,001 dan 0,0001, serta jumlah epoch 50[8]. Dari hasil pengujian tersebut, model keenam menggunakan fitur Word2Vec menghasilkan akurasi tertinggi sebesar 95,17, dengan rata-rata presisi, recall, dan F1-score mencapai 95.[8]

3.12. *Convolutional Long Short-Term Memory*

Algoritma kombinasi antara CNN dan LSTM ini memiliki kinerja lebih baik dari pendahulunya. Hal ini dapat dilihat dari nilai F1-score yang melebihi kedua metode lainnya. Dalam dataset berita yang diuji, C-LSTM mampu mencapai nilai F1-score sebesar 0,9327 atau 93,27%, yang lebih baik 2,4% dibandingkan dengan model LSTM dan 3,42% dibandingkan dengan model CNN yang telah diuji sebelumnya.[7]

Nilai F1-score dari pengklasifikasian berita dipengaruhi oleh ukuran batch dan learning rate. Hasilnya menunjukkan bahwa ukuran batch yang lebih kecil menghasilkan nilai F1-score yang lebih tinggi daripada ukuran batch yang besar. Sebagai hasilnya, dalam penelitian ini, ukuran batch yang dipilih sebagai hyperparameter adalah 16. Sementara itu, learning rate yang dipilih adalah learning rate default dari optimizer Adam yaitu 0,001.[7]

3.13. *Multi Layer Perceptron*

Dalam klasifikasi topik berita, SVM dan MLP dapat digunakan sebagai algoritma dengan melakukan feature selection menggunakan TF-IDF pada tahap preprocessing data. Feature selection dilakukan dengan mempersiapkan data pelatihan berdasarkan tingkat kepentingan korpus pada dataset. Algoritma SVM dan MLP terbukti lebih unggul dari lima algoritma lainnya dalam tugas klasifikasi topik berita. SVM dapat menghasilkan skor akurasi sebesar 74%, sementara MLP dapat mencapai skor akurasi sebesar 78%. Evaluasi lanjut dilakukan dengan menghitung rata-rata bobot untuk precision dan recall, yang menghasilkan skor berturut-turut sebesar 76% dan 74% pada SVM serta 79% dan 78% pada MLP. Selain menggunakan metode TF-IDF, vektorisasi pada tahap feature selection dapat dilakukan dengan menggunakan metode lain seperti FastText dan Word2Vec pada penelitian selanjutnya.[6]

4. PEMBAHASAN

Catatan pertama dalam melakukan review literatur tentang penelitian klasifikasi multi label teks berita bahasa Indonesia adalah pentingnya ketersediaan dataset yang dapat diakses secara publik, sehingga dataset tersebut mudah digunakan sebagai bahan penelitian.

Dalam melakukan klasifikasi multi-label pada teks berita bahasa Indonesia, terdapat beberapa tantangan yang perlu diperhatikan, seperti variasi bahasa dan penggunaan sinonim yang berbeda-beda. Oleh karena itu, penggunaan model yang baik dan data yang memadai sangat diperlukan untuk memperoleh hasil klasifikasi yang akurat.

Tabel 11. Perbandingan tingkat akurasi pada model dengan klasifikasi pada lebih dari 1 label

No	Metode	Literatur	Akurasi/Hamming los
1	LSTM+Word2Vec	[8]	95%
2	SVM+MI	[4]	94.24%
3	CLSTM	[7]	93,27%
4	SVM	[3]	92%
5	SVM+Doc2Vec	[5]	90%
6	NBC+ECSS	[12]	90%
7	KNN	[11]	89,9%
8	KNN+Word2Vec	[10]	89,2%
9	KNN+MI	[9]	0.0856
10	SVM+MLP	[6]	78%

Berbicara mengenai peninjauan model pada setiap literatur, penggunaan Deep Learning untuk mengklasifikasikan teks dengan label terbukti memiliki performa yang bagus. Terlihat bahwa beberapa literatur berhasil mencapai akurasi 95%[8]. Namun, ketika Machine Learning dan Deep Learning digabungkan, akurasi terendah yang diperoleh hanya 78% [6]. Namun, model-model yang dibuat dalam setiap literatur berhasil memberikan akurasi performa di atas 89% ketika menggunakan satu algoritma. Model yang terbaik dihasilkan dengan menggunakan metode Convolutional Long Short-Term Memory dengan akurasi 93,27% [7]. Sedangkan akurasi terendah diperoleh ketika menggunakan KNN, yaitu 89,9% [11].

Performa pemodelan yang buruk untuk data dengan lebih dari satu label disebabkan oleh masalah pada data. Ketidakseimbangan jumlah data pada beberapa label menyebabkan model kesulitan dalam mengenali teks pada label tertentu, dan ini mempengaruhi nilai akurasi model secara keseluruhan. Selain itu, tahap pre-processing juga harus diperhatikan. Teknik pre-processing yang digunakan sangat memengaruhi cara model belajar dari data yang diberikan

Beberapa saran yang dapat diajukan untuk penelitian selanjutnya ditemukan dari hasil reviu yang sudah dilakukan:

1. Membuat model Machine Learning atau Deep Learning yang dapat meningkatkan performa klasifikasi pada data multilabel
2. Menentukan model terbaik dan terakurat dengan parameter uji berupa teknik pre-processing yang diterapkan.
3. Meskipun algoritma Neural Network (RNN dan CNN) memiliki performa yang sangat baik dalam

melakukan klasifikasi data satu label, akan sangat menarik untuk menguji kemampuan algoritma tersebut dalam melakukan klasifikasi pada data multilabel

4. Melakukan pengujian dengan memanfaatkan data yang telah dipublikasikan untuk umum, dengan tujuan mempermudah perbandingan terhadap penelitian-penelitian sebelumnya.

5. KESIMPULAN

Berdasarkan reviu yang sudah dilakukan terhadap literatur-literatur yang diperoleh, dapat disimpulkan bahwa klasifikasi multi-label pada teks berita bahasa Indonesia menggunakan Machine Learning adalah suatu teknik yang digunakan untuk memprediksi label atau kategori dari sebuah dokumen teks berita bahasa Indonesia yang memiliki lebih dari satu label atau kategori. Metode ini dapat dilakukan dengan memanfaatkan teknik *Machine Learning* dan *Deep Learning* untuk membangun model klasifikasi. Hasil dari klasifikasi multi-label pada teks berita dapat digunakan untuk berbagai keperluan, seperti analisis sentimen, pengelompokkan dokumen, atau rekomendasi berita kepada pengguna. Namun, perlu diingat bahwa meskipun teknik *Machine Learning* dan *Deep Learning* dapat memberikan hasil yang akurat, namun model yang dibangun memerlukan dataset yang besar dan bervariasi serta pengaturan parameter yang tepat agar dapat menghasilkan prediksi yang optimal. Selain itu, interpretasi dari hasil klasifikasi juga perlu diperhatikan agar dapat memberikan pemahaman yang tepat mengenai kategori yang ditemukan.

DAFTAR PUSTAKA

[1] Made Riartha Prawira I, Adiwijaya, and Syahrul Mubarok M, "Klasifikasi Multi-Label Pada Topik Berita Berbahasa Indonesia Menggunakan Multinomial Naïve Bayes," vol. 5, no. 3, p. 7774, 2018.

[2] B. H. Mahendra, Adiwijaya, and U. N. Wisesty, "Kategorisasi Berita Multi-Label Berbahasa Indonesia Menggunakan Algoritma Random Forest," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9030–9041, 2019.

[3] M. Alif, I. Aulia, and Y. E. Kurniawati, "Pengembangan Model Klasifikasi Dokumen Artikel Teks Berita Olahraga dan Bukan Olahraga dalam Bahasa Indonesia Menggunakan Algoritma Support Vector Machine," vol. 8, no. 2, pp. 2279–2291, 2022.

[4] L. G. Irham, A. Adiwijaya, and U. N. Wisesty, "Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine," *J. Media Inform. Budidarma*, vol. 3, no. 4, p. 284, 2019, doi: 10.30865/mib.v3i4.1410.

[5] K. I. Gunawan and J. Santoso, "Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia," *J. Inf. Syst. Hosp. Technol.*, vol. 3, no. 01, pp. 29–38, 2021, doi: 10.37823/insight.v3i01.126.

[6] S. Sudianto, A. D. Sripamuji, I. Ramadhanti, R. R. Amalia, J. Saputra, and B. Prihatnowo, "Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasisifikasi Topik Berita," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 11, no. 2, pp. 84–91, 2022, [Online]. Available: https://ejournal.undiksha.ac.id/index.php/janapati/article/view/44151

[7] Y. Widhiyasana, T. Semiawan, I. Gibran, A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia (Convolutional Long Short-Term Memory Implementation for Indonesian News Classification)," *J. Nas. Tek. Elektro dan Teknol. Inf.* |, vol. 10, no. 4, pp. 354–361, 2021.

[8] W. Kurnia Sari, D. Palupi Rini, R. Firsandaya Malik, and I. B. Saladin Azhar, "Klasifikasi Teks Multilabel pada Artikel Berita Menggunakan Long Short-Term Memory dengan Word2Vec," *J. Resti*, vol. 1, no. 3, pp. 276–285, 2022.

[9] A. Gormantara and D. Boli Watomakin, "Klasifikasi Kategori dan Pelabelan Berita Bahasa Indonesia Menggunakan Mutual Information Dan K-Nearest Neighbor," *Tematika2*, vol. 8, pp. 75–82, 2020.

[10] Nur Ghaniaviyanto Ramadhan, "Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1083–1089, 2021, doi: 10.29207/resti.v5i6.3547.

[11] A. Y. Muniar, P. Pasnur, and K. R. Lestari, "Penerapan Algoritma K-Nearest Neighbor pada Pengklasifikasian Dokumen Berita Online," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 2, p. 137, 2020, doi: 10.35585/inspir.v10i2.2570.

[12] E. Y. Hidayat and M. A. Rizqi, "Klasifikasi Dokumen Berita Menggunakan Algoritma Enhanced Confix Stripping Stemmer dan Naïve Bayes Classifier," *J. Nas. Teknol. dan Sist. Inf.*, vol. 6, no. 2, pp. 90–99, 2020, doi: 10.25077/teknosi.v6i2.2020.90-99.

[13] I. K. Syuriadi, Adiwijaya, and W. Astuti, "Klasifikasi Teks Multi Label pada Hadis dalam Terjemahan Bahasa Indonesia Berdasarkan Anjuran, Larangan dan Informasi menggunakan TF-IDF dan KNN," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9121–9132, 2019.

[14] B. K. Palma, D. T. Murdiansyah, and W. Astuti, "Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K- Nearest Neighbor," *eProceedings …*, vol. 8, no. 5, pp. 10637–10649, 2021.