

Klasifikasi Topik terhadap Judul Berita Kasus Covid-19 dengan Multilayer Perceptron

Topic Classification of Covid-19 Case News Headline using Multilayer Perceptron

Faradhiba Salsabila¹, Afrida Helen², Susi Yuliatwati³

^{1,3}Linguistik, Universitas Padjadjaran

²Teknik Informatika, Universitas Padjadjaran

E-mail: ¹faradhiba18001@mail.unpad.ac.id, ²helen@unpad.ac.id, ³susi.yuliatwati@unpad.ac.id

Abstrak

Peran media massa berpengaruh dalam meningkatkan kesadaran masyarakat terhadap penyebaran Covid-19. Berdasarkan laporan *Reuters Institute Digital News Report 2022*, media daring cenderung dikonsumsi oleh masyarakat Indonesia sebagai sumber berita dengan persentase 88%. Hal tersebut menunjukkan bahwa media daring merupakan tempat penyebaran informasi yang penting. Penelitian ini bertujuan untuk mengklasifikasikan topik yang ada dalam berita terkait kasus Covid-19 dalam media massa Kompas dengan menggunakan *multilayer perceptron*. Berdasarkan hasil penelitian, berita kasus Covid-19 dapat dikategorikan menjadi empat label, yaitu kebijakan pemerintah, pemberitahuan informasi, internasional, dan masyarakat umum. Tingkat akurasi yang didapat dari pemodelan dengan *multilayer perceptron* adalah 75%. Kemiripan pada kata-kata dalam data menyebabkan adanya kesalahan dalam membedakan antara satu topik dengan topik lainnya.

Kata kunci: Klasifikasi Topik, Covid-19, Multilayer Perceptron

Abstract

The role of mass media is influential in increasing public awareness of the spread of Covid-19. Based on Reuters Institute Digital News Report 2022 report, online media tend to be favored by the Indonesian people as a news source with a percentage of 88%. This shows that online media is an important place for disseminating information. This study aims to classify the topics in the news related to the Covid-19 case in the Kompas mass media using the multilayer perceptron. Based on the results of the study, news of Covid-19 cases can be categorized into four labels, namely government policy, information notification, international, and public. The level of accuracy obtained from modeling with multilayer perceptron is 75%. The similarity of the words in the data causes errors in distinguishing one topic from another.

Keywords: Topic Classification, Covid-19, Multilayer Perceptron

1. PENDAHULUAN

Covid-19 adalah penyakit menular yang berasal dari virus SARS-CoV-2. Masuknya Covid-19 ke Indonesia berdampak pada berbagai aspek kehidupan. Munculnya berbagai kebijakan untuk mengatasi Covid-19 tidak terelakkan. *Work From Home* (WFH), Pembelajaran Jarak Jauh (PJJ), PPKM, dan sebagainya adalah kebijakan yang dibuat untuk mengatasi pandemi Covid-19.

Kesadaran masyarakat terhadap kebijakan pemerintah dibutuhkan agar penurunan penyebaran Covid-19 dapat tercapai. Kesadaran tersebut dapat tumbuh jika masyarakat memiliki informasi yang memadai mengenai Covid-19. Media massa ikut berperan dalam hal ini. Hal tersebut disebabkan media massa merupakan wadah komunikasi dan informasi yang mampu meraih masyarakat dari berbagai kalangan [1]. Adanya *gadget* mempermudah penyebaran informasi tersebut secara masif.

Berita daring sebagai bagian dari media massa tak lepas dari kehidupan masyarakat. Berdasarkan laporan *Reuters Institute Digital News Report 2022*, masyarakat Indonesia cenderung mengonsumsi media daring sebagai sumber berita dengan persentase 88% [2]. Hal tersebut menunjukkan media daring merupakan tempat penyebaran informasi yang penting.

Kompas merupakan satu di antara media massa yang aktif menyebarkan berita terkini, termasuk kasus Covid-19. Klasifikasi topik pada berita terkait kasus Covid-19 dibutuhkan untuk mempermudah pengelompokan berita yang tersebar dengan begitu cepat. Hal tersebut dapat membantu masyarakat untuk mencari informasi yang dibutuhkan mengenai kasus Covid-19.

Linguistik komputasional adalah bidang interdisipliner antara linguistik dan komputer. Linguistik komputasional membahas pemrosesan bahasa alami (*Natural Language Processing*). Linguistik komputasional berhubungan dengan membuat model komputasi agar mesin dapat mempelajari, memahami, dan memproduksi konten mengenai bahasa manusia [3]. Bahasa manusia diubah menjadi data dengan format yang dapat dipahami oleh mesin. Klasifikasi merupakan teknik *data mining* yang digunakan untuk memberikan label kelas sehingga model algoritma dapat memprediksi label kelas [4]. Klasifikasi topik merupakan bagian dari linguistik komputasional.

Topik adalah inti pembahasan dalam teks. Topik dari berita dapat tecermin dari penggunaan judul. Hal tersebut berkaitan dengan fungsi judul teks. Judul yang baik dapat menggambarkan isi teks [5]. Judul membantu pembaca ketika memilih teks, termasuk teks berita. Pembaca dapat mencari judul yang dibutuhkan dengan melihat kata kunci dalam judul. Dengan kata lain, judul dapat dianggap mewakili topik teks. Oleh sebab itu, penelitian ini menjadikan judul berita sebagai data untuk mengklasifikasi topik pada berita.

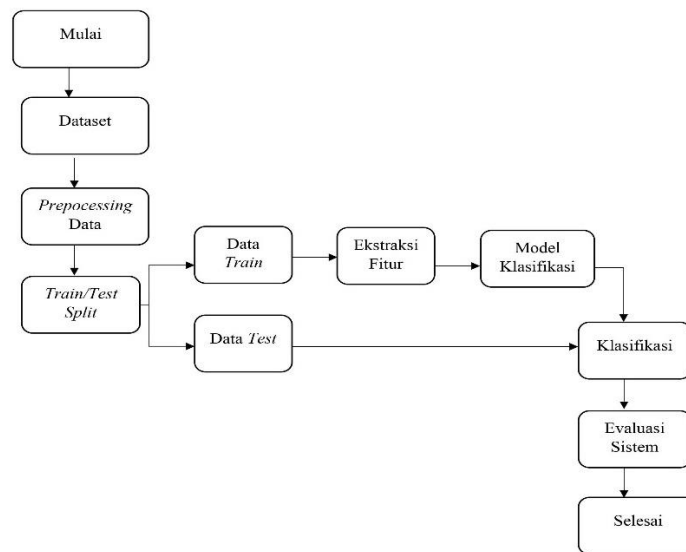
Beberapa penelitian terdahulu membahas klasifikasi topik pada berita terkait kasus Covid-19. Gotama dkk. mengklasifikasikan berita hoaks Covid-19 dengan menggunakan Klasifikasi *Rocchio* dan *Cosine Similarity* [6]. Tingkat akurasi yang didapat dalam penelitian tersebut adalah 87,6%. Palma dkk. menjadikan K-Nearest Neighbor sebagai algoritma untuk mengklasifikasikan berita hoaks Covid-19 [7]. Tingkat akurasi yang dicapai dalam penelitian tersebut adalah 48% dengan nilai $k = 5$. Prasetya dan Ferdiansyah menerapkan algoritma *naive bayes* dan *cross validation* dalam pengklasifikasian berita hoaks Covid-19 [8]. Tingkat akurasi penerapan algoritma tersebut adalah 86,3%.

Penggunaan *multilayer perceptron* sudah dilakukan pada beberapa penelitian terdahulu. Tika dan Adiwijaya membuat klasifikasi topik berita dengan *multilayer perceptron* dan tingkat yang didapat adalah akurasi 77, 44% [9]. Berita yang dijadikan data dalam penelitian tersebut adalah berita dari berbagai media massa berbahasa Indonesia. Sudianto dkk. menggunakan algoritma *support vector machine* dan *multilayer perceptron* untuk mengklasifikasi topik berita berbahasa Indonesia dalam berbagai situs web dari Juli sampai Desember 2015 [10]. Dengan menggunakan *support vector machine*, tingkat akurasi yang didapat adalah 74%, sedangkan dengan menggunakan *multilayer perceptron*, tingkat akurasi yang dicapai adalah 78%.

Berdasarkan penelitian terdahulu tersebut, belum ada yang mengklasifikasikan topik pada berita terkait kasus Covid-19 dengan menggunakan *multilayer perceptron*. Oleh sebab itu, penelitian ini bertujuan untuk mengklasifikasikan topik yang ada dalam berita terkait kasus Covid-19 dalam media massa Kompas dengan menggunakan *multilayer perceptron*. Penelitian ini diharapkan dapat dijadikan referensi bagi penelitian terkait klasifikasi topik berita.

2. METODE PENELITIAN

Penelitian ini terdiri beberapa tahap. Untuk membantu dalam memberikan gambaran, tahapan penelitian ditampilkan dalam *flowchart*. Berikut ini tahapan penelitian dalam Gambar 1.



Gambar 1 *Flowchart* Penelitian

2.1 Pengumpulan Data

Penelitian ini menggunakan 500 berita berbahasa Indonesia dengan tag #KASUS COVID-19. Data diambil dengan melakukan *crawling* pada website Kompas. Data dikumpulkan dari sampai 18 Maret 2020 sampai 4 Mei 2020. Dataset terbagi menjadi dua, yaitu *data training* dan *data testing*. Data uji digunakan untuk mengevaluasi tingkat akurasi data saat menggunakan *confusion matrix*.

Langkah pertama dalam mengumpulkan data adalah memasukkan *script* di Google Colab. Peneliti kemudian memasukkan tautan URL yang sudah dipilih dan nomor halaman dari tautan tersebut. Peneliti kemudian menekan *runtime* kemudian *run all*. Proses *crawling* pun berjalan. Ketika proses *crawling* berhenti karena sinyal internet terputus, peneliti menekan *save and reload* untuk menjalankan *crawling* kembali. Setelah proses *crawling* sudah selesai, *file* korpus dapat diunduh dalam format csv. Data kemudian dirapikan agar mudah untuk dibaca.

2.2 Preprocessing

Preprocessing adalah proses mengolah data agar dapat diklasifikasikan. Pada tahap *preprocessing*, *noise* dalam data dihilangkan. Berikut ini pemaparannya.

a. Case folding

Semua huruf kapital pada data diubah menjadi huruf kecil (*lowercase*). Pada tahap ini, data diseragamkan dalam bentuk huruf kecil.

b. Cleansing

Tanda baca yang tidak diperlukan dalam data dihapus. Spasi berlebihan pada data dihilangkan. *Enter* dihilangkan sehingga satu baris untuk satu data.

c. Stemming

Pada tahap ini, imbuhan yang terdapat dalam kata dihapus. Imbuhan yang dihapus adalah prefiks, sufiks, infiks, dan konfiks. Hal ini membuat kata berimbuhan kembali ke bentuk dasarnya. Contohnya adalah kata membaca menjadi baca. *Stemming* dilakukan dengan *library* Sastrawi.

d. Stopword removal

Pada tahap *stopword removal*, kata-kata yang termasuk dalam *stopword list* akan dihapus dari data. *Content word* akan dipertahankan, sedangkan *function word* akan dihapus. Hal ini disebabkan *content word* berisi informasi. Contoh kelas kata yang termasuk dalam *content word* adalah nomina, verba, dan adjektiva. Di sisi lain, *function word* berperan untuk

membuat suatu kalimat sesuai dengan tata bahasa. Contoh *function word* adalah preposisi, konjungsi, dan pronomina.

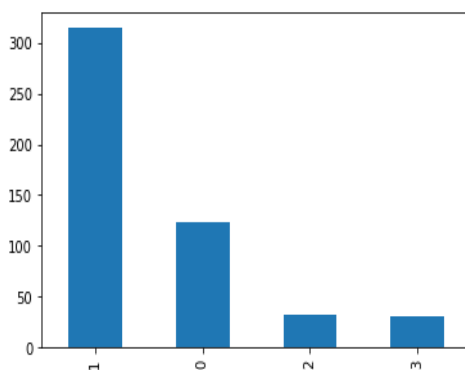
Tabel 1 Hasil Preprocessing

No.	Input	Output
1.	Anies Minta Masyarakat Patuhi "Social Distancing" untuk Tekan Kasus Covid-19 di Jakarta	anies masyarakat patuh social distancing tekan covid 19 jakarta
2.	Kasus Covid-19 di DKI Jakarta dan Penerapan PSBB yang Telah Disetujui Menkes	covid 19 dki jakarta terap psbb tuju menkes
3.	UPDATE Kasus Covid-19 di Jakarta hingga 17 April: 2.819 Positif, 204 Sembuh, 248 Meninggal	update covid 19 jakarta 17 april 2 819 positif 204 sembuh 248 tinggal
4.	UPDATE: Bertambah 106, Total Ada 2.092 Kasus Covid-19 di Indonesia	update 106 total 2 092 covid 19 indonesia
5.	Virus Corona, Korea Selatan dan Klaim Penurunan Kasus Covid-19	virus corona korea selatan klaim turun covid 19
6.	Italia, Spanyol, dan Perancis Laporkan Tren Penurunan Kasus Covid-19	italia spanyol perancis lapor tren turun covid 19
7.	Dokter Muda FK UI Akan Dikerahkan Tangani Kasus Covid-19 Ringan	dokter muda fk ui kerah tangan covid 19 ringan
8.	Cerita Dokter RSUD Wangaya Buat Pelindung Wajah yang Langka di Pasaran	cerita dokter rsud wangaya lindung wajah langka pasar

Tabel 1 menampilkan hasil dari *preprocessing* data. Pada tabel tersebut, semua huruf kapital menjadi huruf kecil pada data yang sudah mengalami *preprocessing*. Berbagai tanda baca, seperti titik dua (:), tanda hubung (-), tanda kutip (“ ”), dan tanda koma (,) sudah dihilangkan pada data. Data pada kolom *output* hanya memuat *content word*. *Function word* seperti *untuk, dan, yang, telah, di,* dan lain-lain dihapus. Imbuhan seperti prefiks -ber dan sufiks -i dihilangkan. Kata berimbuhan dikembalikan ke bentuk dasar, seperti kata penerapan, yang disisipi -pe -an berubah menjadi terap.

2.3 Pelabelan Data

Dari 500 berita yang sudah dikumpulkan, data dibagi menjadi empat kategori. Kategori tersebut ditentukan berdasarkan fokus dari berita kasus Covid-19. Kategori dengan label 0 adalah kebijakan pemerintah. Kategori tersebut berkaitan dengan kebijakan pemerintah dalam mengatasi Covid-19, seperti pemberlakuan PSBB, isolasi, karantina wilayah, dan lain-lain. Kategori dengan label 1 adalah pemberitahuan informasi. Judul dalam kategori tersebut berhubungan dengan pemberitahuan peningkatan atau penurunan kasus Covid-19. Kategori dengan label 2 adalah internasional. Judul yang dimuat dalam kategori tersebut membahas berita seputar Covid-19 dari luar negeri. Kategori dengan label 3 adalah masyarakat umum. Judul dengan kategori tersebut membahas Covid-19 dari sudut pandang masyarakat umum, seperti dokter, atlet, dan lain-lain.



Gambar 2 Kategori Data

Label 0 berjumlah 124 data, label 1 berjumlah 314 data, label 2 berjumlah 32 data, dan label 3 berjumlah 31 data. Data dengan jumlah paling banyak adalah label 1, yaitu pemberitahuan informasi. Label 3, yaitu masyarakat umum, merupakan jumlah data paling sedikit dengan memiliki selisih 1 data dari label 2.

Tabel 2 Pelabelan Judul

No.	Judul	Kategori	Topik
1.	anies masyarakat patuh social distancing tekan covid 19 jakarta	0	Kebijakan Pemerintah
2.	covid 19 surabaya pemkot bahas rencana psbb	0	Kebijakan Pemerintah
3.	update covid 19 jakarta 17 april 2 819 positif 204 sembuh 248 tinggal	1	Pemberitahuan Informasi
4.	update 106 total 2 092 covid 19 indonesia	1	Pemberitahuan Informasi
5.	virus corona korea selatan klaim turun covid 19	2	Internasional
6.	italia spanyol perancis lapor tren turun covid 19	2	Internasional
7.	dokter muda fk ui kerah tangan covid 19 ringan	3	Masyarakat Umum
8.	cerita dokter rsud wangaya lindung wajah langka pasar	3	Masyarakat Umum

Berdasarkan Tabel 2, judul dalam setiap kategori memiliki karakteristik masing-masing. Label 0 identik dengan nomina yang membahas kata benda, seperti *social distancing* dan PSBB. Label 0 juga dapat membuat nama *stakeholder* dan lembaga. Label 1 memiliki kecenderungan dengan kata *update* dan angka. Label 2 dapat diidentifikasi melalui adanya negara asing. Label 3 identik dengan munculnya nomina yang berkaitan dengan profesi masyarakat umum.

2.4 Pembagian Data dengan Train/Test Split

Train/Test Split merupakan metode untuk mengetahui performa model *machine learning*. Metode ini terdiri atas dua bagian, yaitu *training set* dan *test set*. *Training set* adalah dataset yang dilatih agar dapat menjalankan fungsi dari algoritma. *Test set* adalah dataset yang dites agar dapat mengetahui performanya. Dengan metode ini, *new data* atau data yang belum dilatih akan mendapat hasil prediksi yang lebih akurat. Data dibagi dengan perbandingan 80:20 sehingga data latih berjumlah 400, sedangkan data uji berjumlah 100.

2.5 Ekstraksi Fitur

Langkah berikutnya adalah *CountVectorizer*. Data diubah menjadi representasi vektor dengan menggunakan *Countvectorizer*. Kata-kata dialihkan menjadi kode bilangan bulat agar dapat diinput ke model. Hasil yang didapat ditampilkan dalam Gambar 3.

```
Vocabulary:
{'karantina': 459, 'laku': 503, 'jalan': 433, 'kunci': 495, 'sukses': 736, 'pemda': 623, 'ttu': 786, 'tekan': 760, 'covid': 334, '19': 50, 'no1': 576.
Full vectorizer:
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 1]
 ...
 [1 0 0 ... 0 0 0]
 [1 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

Gambar 3 Vector Array

Penelitian ini menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF) pada tahap ekstraksi dataset. Metode ini dapat menghitung nilai TF dan IDF dari setiap kata di setiap dokumen. Metode TF dapat menentukan umum atau tidak umumnya suatu kata

berdasarkan frekuensi kemunculannya, sedangkan metode IDF dapat menentukan distribusi kata secara luas pada kumpulan dokumen. TF-IDF adalah proses pembobotan kata dalam data berdasarkan perhitungannya terhadap frekuensi kemunculan kata. Gambar 4 menunjukkan hasil yang didapat dari langkah TF-IDF.

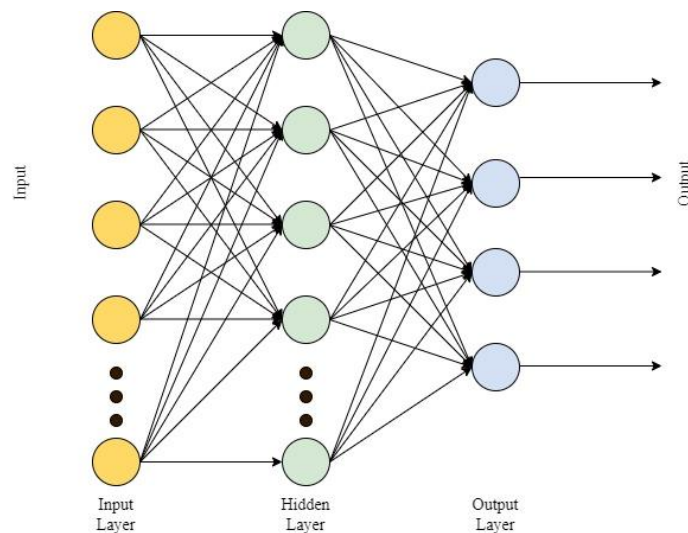
$$\begin{bmatrix}
 [0. & 0. & 0. & \dots & 0. & 0. &] \\
 [0. & 0. & 0. & \dots & 0. & 0. &] \\
 [0. & 0. & 0. & \dots & 0. & 0.28630262] \\
 \dots & & & & & & \\
 [0.30945753 & 0. & 0. & \dots & 0. & 0. &] \\
 [0.32571644 & 0. & 0. & \dots & 0. & 0. &] \\
 [0. & 0. & 0. & \dots & 0. & 0. &]
 \end{bmatrix}$$

Gambar 4 TF-IDF Array

2.6 Penerapan Algoritma

Setelah ekstraksi fitur, langkah berikutnya adalah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Metode tersebut membuat data sintesis untuk menambah jumlah kelas minor agar dataset menjadi seimbang [11]. Pembuatan data sintesis tersebut mempertimbangkan *k-nearest neighbor*. Pada penelitian ini, jumlah data label 1 jauh lebih banyak daripada label 2 dan 3. Jumlah label tertentu yang lebih banyak dari label lainnya dapat membuat akurasi algoritma cenderung lebih tinggi pada kelas mayoritas [12]. Oleh sebab itu, penggunaan SMOTE dibutuhkan untuk membuat proses klasifikasi menjadi lebih seimbang.

Multilayer perceptron dapat digunakan untuk mengklasifikasi topik pada berita. *multilayer perceptron* adalah turunan perceptron yang berasal dari sejumlah neuron yang dihubungkan oleh bobot penghubung. Algoritma ini mampu melakukan proses pembelajaran secara *back propagation* sehingga dapat menetapkan bobot tertentu untuk memprediksi dengan tepat [13]. *Multilayer perceptron* adalah model yang dapat mengolah data multivariat yang tidak linear karena memiliki beberapa *layer* [14]. Pada umumnya, *multilayer perceptron* memiliki *layer* masukan (*input*), *layer* tersembunyi (*hidden layer*), dan *layer* keluar (*output*). Setiap *layer* memiliki fungsi tertentu. Struktur *multilayer perceptron* ditampilkan dalam Gambar 5.



Gambar 5 Struktur *Multilayer Perceptron*

Pada tahap pemodelan klasifikasi, input algoritma *multilayer perceptron* adalah kata-kata hasil dari tahap preprocessing. Setiap node input dikalikan dengan bobot yang dibangkitkan secara random dengan rentang nilai 0 sampai dengan 1. Hasil perkalian node dengan bobot dijumlahkan untuk dijadikan input pada node berikut. Penelitian ini menggunakan satu hidden layer dan empat node output sesuai dengan jumlah kelas yang diusulkan. SMOTE

digunakan untuk membantu proses klasifikasi pada data yang tidak seimbang. Penerapan algoritma ditunjukkan dalam gambar 6.

```
[ ] textclassifier =Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('smote', SMOTE(random_state=12)),
    ('MLP', MLPClassifier(alpha=1, max_iter=1000))
])
```

Gambar 6 Penerapan Algoritma

3. HASIL DAN PEMBAHASAN

3.1 Evaluasi Model

Setelah menerapkan algoritma, tahap berikutnya adalah evaluasi sistem. Hasil yang didapat dihitung *precision*, *recall*, *F1-score*, dan akurasi. *Precision* merupakan perbandingan antara kasus diprediksi positif dengan prediksi seluruh hasil yang positif. *Recall* adalah perbandingan antara diprediksi positif dengan kasus positif yang dengan tepat diprediksi positif. *F1-Score* adalah perbandingan rata-rata presisi antara *precision* dan *recall*. *Accuracy* adalah perbandingan prediksi benar dengan seluruh data. Berikut ini rinciannya.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Accuracy} = \text{TP} + \text{FN} / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

$$\text{F1-Score} = 2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision})$$

Hasil yang didapat dievaluasi dengan menggunakan *confusion matrix*. *Confusion matrix* merupakan metode yang digunakan untuk menilai model klasifikasi *machine learning*. Pada tabel *confusion matrix*, jumlah data yang diklasifikasikan ditampilkan dengan detail baik salah maupun benar.

Tabel 3 *Confusion Matrix* pada Data Latih

		Topik Prediksi			
		Kebijakan Pemerintah	Pemberitahuan Informasi	Internasional	Masyarakat Umum
Topik Aktual	Kebijakan Pemerintah	76	16	2	3
	Pemberitahuan Informasi	14	229	5	3
	Internasional	2	0	23	0
	Masyarakat Umum	5	0	0	22

Tabel 3 menunjukkan hasil *confusion matrix* pada 400 *data train*. Data pada topik pemberitahuan statistik paling banyak ditemukan, diikuti oleh topik kebijakan pemerintah, topik internasional, dan topik masyarakat umum. Hal tersebut dikarenakan jumlah data pada topik pemberitahuan statistik berjumlah lebih banyak dari topik lainnya.

Tabel 4 *Confusion Matrix* pada Data Uji

		Topik Prediksi			
		Kebijakan Pemerintah	Pemberitahuan Informasi	Internasional	Masyarakat Umum
Topik Aktual	Kebijakan Pemerintah	16	8	1	2
	Pemberitahuan Informasi	2	54	5	2
	Internasional	0	2	5	0
	Masyarakat Umum	3	0	0	1

Pada Tabel 3 dan Tabel 4, topik kebijakan pemerintah diklasifikasikan menjadi topik pemberitahuan Informasi sebanyak 16 data latih dan 8 data uji. Hal ini tidak terlepas dari kemiripan kata-kata dalam data menyebabkan adanya kesalahan dalam membedakan antara satu topik dengan topik lainnya.

Tabel 5 Evaluasi Model

Topik	Precision	Recall	F1-Score
Kebijakan Pemerintah	0,76	0,59	0,67
Pemberitahuan Informasi	0,84	0,86	0,85
Internasional	0,45	0,71	0,56
Masyarakat Umum	0,20	0,25	0,22
Akurasi	0,75		
Rata-Rata	0,56	0,60	0,58

Berdasarkan Tabel 3, akurasi dari pengujian klasifikasi adalah 0,75. Nilai rata-rata *precision* adalah 0,56, nilai rata-rata *recall* adalah 0,60, dan nilai rata-rata *F1-Score* adalah 0,58. Topik yang memiliki *precision*, *recall*, dan *F1-score* paling besar adalah topik pemberitahuan informasi. Di sisi lain, topik dengan nilai *precision*, *recall*, dan *F1-Score* paling rendah adalah topik masyarakat umum.

4. KESIMPULAN DAN SARAN

Berita kasus Covid-19 terdiri atas 4 kategori, yaitu kebijakan pemerintah, pemberitahuan informasi, internasional, dan masyarakat umum. Berdasarkan hasil penelitian, simpulan dari penelitian ini adalah berita kasus Covid-19 dapat dikategorikan menjadi empat label, yaitu kebijakan pemerintah, pemberitahuan informasi, internasional, dan masyarakat umum. Tingkat akurasi yang didapat dari pemodelan dengan *multilayer perceptron* adalah 75%. Topik dalam berita Kasus Covid-19 memiliki kemiripan pada kata-kata dalam judul antara satu topik dengan topik lainnya. Judul yang membahas kebijakan pemerintah bisa saja memiliki kata yang sama dengan pemberitahuan informasi, seperti nama daerah atau nama provinsi. Adanya kemiripan pada kata-kata antarjudul menjadi penyebab kesalahan dalam membedakan antara satu topik dengan topik lainnya.

Penelitian berikutnya disarankan mencoba berbagai metode untuk meningkatkan akurasi dengan algoritma *multiplayer perceptron*. *K-Fold* dapat digunakan untuk pembagian data. Metode *up-down sampling* dapat digunakan untuk mengatasi data yang tidak seimbang. Untuk meningkatkan akurasi prediksi model *machine learning*, *Hyperparameter Tuning* menjadi opsi yang menarik untuk dicoba.

DAFTAR PUSTAKA

- [1] B. A. Santosa, "Peran Media Massa dalam Mencegah Konflik," *J. ASPIKOM*, vol. 3, no. 2, pp. 199–214, 2017, [Online]. Available: <https://jurnalaspikom.org/index.php/aspikom/article/view/128>.
- [2] N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, and R. K. Nielsen, "Reuters Institute Digital News Report 2022," Oxford, 2022. [Online]. Available: <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022>.
- [3] J. Hirschberg and C. D. Manning, "Advances in Natural Language Processing," *Science* (80-.), vol. 349, no. 6245, pp. 261–266, 2015, [Online]. Available: <https://www.science.org/doi/10.1126/science.aaa8685>.
- [4] S. Defiyanti, "Integrasi Metode Clustering dan Klasifikasi untuk Data Numerik," in

- Conference on Information Technology and Electrical Engineering*, 2017, pp. 256–261, [Online]. Available: https://www.researchgate.net/publication/319165131_Integrasi_Metode_Clustering_dan_Klasifikasi_untuk_Data_Numerik.
- [5] Musyawir, Susiati, and A. Irmawati, “Strategi Pemilihan Judul Penelitian Kebahasaan bagi Pemula; Mahasiswa Program Studi Pendidikan Bahasa Indonesia dan Sastra Indonesia Universitas Iqra Buru,” *J. Pengabd. Mandiri*, vol. 1, no. 3, pp. 517–524, 2022, [Online]. Available: <https://bajangjournal.com/index.php/JPM/article/view/1817>.
- [6] I. Gotama, S. Hariyanto, and H. Wijaya, “Klasifikasi Berita Hoaks Topik Covid-19 dengan Klasifikasi Rocchio dan Cosine Similarity,” *J. Algor*, vol. 2, no. 1, pp. 84–92, 2020, [Online]. Available: <https://jurnal.ubd.ac.id/index.php/algor/article/view/455>.
- [7] B. K. Palma, D. T. Murdiansyah, and W. Astuti, “Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algoritma K-nearest Neighbor,” *eProceedings Eng.*, vol. 8, no. 5, pp. 10637–10649, 2021, [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15672>.
- [8] F. Prasetya and Ferdiansyah, “Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes,” *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, pp. 132–139, 2022, [Online]. Available: <http://dx.doi.org/10.30865/json.v4i1.4852>.
- [9] G. Tika and Adiwijaya, “Klasifikasi Topik Berita Berbahasa Indonesia menggunakan Multilayer Perceptron,” *e-Proceeding Eng.*, vol. 6, no. 1, pp. 2137–2143, 2019, [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/8503>.
- [10] Sudianto, A. D. Sripamuji, and I. R. Ramadhanti, “Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasifikasi Topik Berita,” *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 11, no. 2, 2022, [Online]. Available: <https://ejournal.undiksha.ac.id/index.php/janapati/article/view/44151>.
- [11] R. Siringoringo, “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote dan K-Nearest Neighbor,” *J. Inf. Syst. Dev.*, vol. 3, no. 1, pp. 44–49, 2018, [Online]. Available: <https://ejournal-medan.uph.edu/index.php/isd/article/view/177>.
- [12] V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, [Online]. Available: <https://arxiv.org/abs/1106.1813>.
- [13] A. Setiadi, “Penerapan Algoritma Multilayer Perceptron untuk Deteksi Dini Penyakit Diabetes,” *Paradigma*, vol. 14, no. 1, pp. 46–59, 2012, [Online]. Available: <https://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/3378>.
- [14] T. Beysolow II, *Applied Natural Language Processing with Python*. Berkeley: Apress, 2018.