KLASIFIKASI TOPIK *TWEET* MENGENAI COVID MENGGUNAKAN METODE MULTINOMIAL NAIVE BAYES DENGAN PEMBOBOTAN TF-IDF

[1]Lydia Mayasari, [2]Dina Indarti
[1,2]Fakultas Teknologi Industri, Universitas Gunadarma
[1,2]Jl. Margonda Raya No. 100, Depok, Jawa Barat 16424
[1]lydiamayasari10@yahoo.com, [2]dina.indarti@staff.gunadarma.ac.id

Abstrak

COVID merupakan virus yang banyak menjangkiti masyarakat Indonesia, bahkan dunia saat ini. Upaya yang dilakukan oleh pemerintah yang tidak luput dari komentar masyarakat mulai dari komentar berupa pujian, kritik, serta saran yang diberikan melalui berbagai media sosial seperti Twitter. Banyak tweet yang dikirimkan mengenai COVID. Tujuan penelitian ini adalah mengklasifikasikan topik tweet mengenai COVID menggunakan Multinomial Naïve Bayes dengan pembobotan Term Frequency-Inverse Document Frequency (TF-IDF). Tahapan penelitian terdiri dari analisis masalah, pengumpulan data, pelabelan data, preprocessing, pembobotan TF-IDF, pelatihan menggunakan Multinomial Naïve Bayes, dan pengujian performa. Data tweet dikumpulkan dari 9 Juni 2021 sampai 9 Juli 2021 dengan kata kunci 'COVID'. Jumlah tweet yang digunakan dalam penelitian ini yaitu 4.909 yang terdiri dari 3.436 data pelatihan dan 1.473 data pengujian. Topik tweet dalam penelitian ini terdiri dari ekonomi, kesehatan, hiburan, sosial, dan hukum. Klasifikasi topik tweet dilakukan pada tweet bahasa Indonesia. Tweet yang telah dikumpulkan lalu melalui tahap preprocessing terdiri dari case folding, tokenizing, stopword removal, normalisasi, dan stemming. Berdasarkan hasil pengujian, akurasi klasifikasi topik tweet menggunakan Multinomial Naïve Bayes dengan pembobotan TF-IDF sebesar 61%.

***Kata Kunci**: Klasifikasi, Multinomial Naïve Bayes, TF-IDF, Topik Tweet, Twitter.*

Abstract

COVID is a virus that infects many people in Indonesia, even the world today. Efforts made by the government cannot miss from public comments in the form of compliments, critics, and suggestions through various social media such as Twitter. Many tweets were sent regarding COVID. The purpose of this study is to classify tweet topics regarding COVID using Multinomial Naïve Bayes with Term Frequency-Inverse Document Frequency (TF-IDF) weighting. The research stages consisted of problem analysis, data collection, data labeling, preprocessing, TF-IDF weighting, training using Multinomial Naïve Bayes, and performance testing. Tweet data was collected from 9 June 2021 to 9 July 2021 with the keyword 'COVID'. The number of tweets used in this research is 4.909 consisting of 3.436 training data and 1.473 testing data. Tweet topics in this study consist of economics, health, entertainment, social, and law. The classification of tweet topics is carried out on Indonesian tweets. Tweets that have been collected then go through a preprocessing stage consisting of case folding, tokenizing, stopword removal, normalization, and stemming. Based on the test results, the accuracy of tweet topic classification using Multinomial Naïve Bayes with the TF-IDF weighting is 61%.

***Keywords**: Classification, Multinomial Naïve Bayes, TF-IDF, Tweet Topics, Twitter.*

PENDAHULUAN

Media sosial merupakan sebuah media yang berguna untuk bersosialisasi satu dengan yang lain yang dapat dilakukan secara *online* melalui jaringan internet tanpa dibatasi ruang dan waktu. Salah satu contoh media sosial tersebut adalah Twitter. Twitter adalah sebuah layanan bagi teman, keluarga, dan rekan kerja untuk berkomunikasi dan tetap berhubung melalui pertukaran pesan yang cepat dan sering. Pengguna Twitter dapat melakukan pengiriman pesan disebut *tweet* yang dibatasi dengan 280. Pengguna memposting *tweet*, yang dapat berisi foto, video, tautan, dan teks [1]. *Tweet* dapat digunakan untuk menyampaikan pendapat atau opini tentang apapun yang terjadi secara langsung sehingga salah satu kegunaan Twitter yaitu untuk mengumpulkan informasi publik [2].

COVID-19 merupakan penyakit baru yang telah menjadi pandemi dan harus diwaspadai karena penularan yang relatif cepat serta memiliki tingkat mortalitas yang tidak dapat diabaikan [3]. COVID-19 ditetapkan secara resmi sebagai pandemik global oleh *World Health Organization* (WHO) pada tanggal 12 Maret 2020. Virus tersebut menyebar luas kepada seluruh masyarakat dunia, termasuk negara Indonesia. Jumlah kasus masyarakat di Indonesia yang terjangkit virus korona semakin bertambah setiap harinya. Berdasarkan data terbaru pada tanggal 3 Maret 2022, jumlah kasus positif di Indonesia mencapai 5.667.355 jiwa [4]. Adanya dampak dan bahaya yang muncul dari kasus COVID-19 maka dibutuhkan pencegahan dan penanganan. Upaya pencegahan meliputi 3M dan vaksinasi yang diharapkan berdampak pada angka konfirmasi positif COVID-19 di Indonesia [5]. Berbagai upaya penanganan juga dilakukan oleh pemerintah mulai dari bidang kesehatan, sosial, ekonomi, dan bidang lainnya. Upaya yang dilakukan oleh pemerintah tidak luput dari komentar masyarakat, mulai dari komentar berupa pujian, kritikan, serta saran yang diberikan melalui berbagai media sosial. Salah satu *platform* media sosial yang sering digunakan masyarakat dalam menyampaikan komentar adalah Twitter. Pengguna Twitter di Indonesia mengalami pertumbuhan yang tinggi [6]. Banyak *tweet* yang dikirimkan mengenai COVID. Klasifikasi topik *tweet* mengenai COVID dapat memudahkan masyarakat untuk memperoleh informasi tentang topik tertentu yang menjadi pembicaraan masyarakat. Pemerintah dapat memanfaatkan informasi klasifikasi topik *tweet* mengenai COVID untuk menindaklanjuti secara tepat terhadap masukkan, saran, serta kritik pada setiap topik.

Klasifikasi topik *tweet* merupakan salah satu permasalahan dalam *text mining* [7]. *Text mining* merupakan proses menemukan informasi dalam sekumpulan dokumen, dan mengidentifikasi secara otomatis pola yang terbentuk, dan berhubungan dengan informasi yang didapat dari kumpulan data yang tidak terstruktur. Klasifikasi teks merupakan proses

menemukan kesamaan dalam dokumen, korpus, maupun kelompok-kelompok dari dokumen yang telah dilabeli sebelumnya, berdasarkan topik, tema yang ditunjukkan oleh koleksi dokumen [8],[9].

Salah satu metode klasifikasi yang sederhana tetapi memiliki akurasi yang cukup tinggi yakni Naïve Bayes. Naïve Bayes merupakan metode yang bekerja sangat baik dan efektif dan efisien dibanding dengan metode klasifikasi lain. Metode Naïve Bayes menunjukkan akurasi dan kecepatan yang tinggi jika diimplementasikan pada basis data yang besar. Metode ini sering digunakan dalam menyelesaikan masalah dalam *machine learning* karena memiliki akurasi yang tinggi dengan perhitungan sederhana [10]. Salah satu model Naïve Bayes yang sering digunakan dalam klasifikasi teks adalah Multinomial Naïve Bayes [11]. Pengklasifikasian kelas dari suatu dokumen pada Multinomial Naïve Bayes tidak hanya ditentukan berdasarkan jumlah kata yang terdapat pada dokumen tersebut tetapi juga ditentukan oleh frekuensi kemunculan kata tersebut. Secara umum, Multinomial Naïve Bayes memiliki kinerja yang lebih baik dibandingkan Naïve Bayes dalam pengklasifikasian dokumen terutama pada jumlah data yang besar [12], [13]. Pembobotan fitur perlu dilakukan sebelum pengklasifikasian untuk meningkatkan akurasi klasifikasi, pembobotan yang umum digunakan adalah *Term Frequency-Inverse Document Frequency* (TF-IDF). Pembobotan TF-IDF mempertimbangkan parameter frekuensi kemunculan fitur dalam dokumen dan jumlah dokumen yang mengandung fitur tersebut [14].

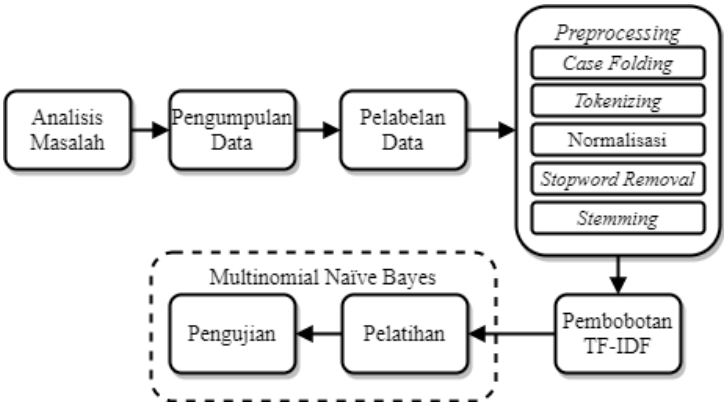
Beberapa penelitian telah membahas mengenai klasifikasi menggunakan metode Multinomial Naïve Bayes dengan pembobotan TF-IDF [12], [13], [15] – [19]. Penelitian yang dilakukan oleh Rahman, Wiranto, dan Doewes membahas mengenai klasifikasi berita *online* dalam teks bahasa Indonesia menggunakan Multinomial Naïve Bayes. Penelitian tersebut menggunakan seleksi fitur dengan *Document Frequency-Thresholding* (DF-*Thresholding*) dan pembobotan dengan TF-IDF. Hasil penelitian menunjukkan bahwa metode klasifikasi menggunakan Multinomial Naïve Bayes dengan pembobotan TF-IDF memiliki akurasi yang lebih baik dibandingkan Multinomial Naïve Bayes dengan seleksi fitur menggunakan DF-*Thresholding*. Hasil akurasi akhir dalam pengklasifikasian berita dengan pembobotan TF-IDF pada metode Multinomial Naïve Bayes sebesar 94,29% [12]. Multinomial Naïve Bayes dengan pembobotan TF-IDF juga digunakan dalam pengklasifikasian genre novel. Genre novel pada penelitian tersebut terdiri dari romantis, komedi, horor, dan misteri. Hasil pengujian pada penelitian tersebut dinyatakan dalam bentuk *confusion matrix* dan diperoleh akurasi sebesar 80,5% [13]. Penelitian lain membahas mengenai klasifikasi kategori surat keluar di Diskominfo Kabupaten Tangerang menggunakan Multinomial Naïve Bayes dengan pembobotan TF-IDF. Surat keluar

terdiri dari 4 kategori yaitu cuti, undangan, perintah, dan peminjaman ruang. Hasil pengujian menunjukkan bahwa implementasi Multinomial Naïve Bayes Classifier pada sistem klasifikasi surat keluar memiliki tingkat akurasi, *precision*, *recall*, dan F-*measure* berturut-turut sebesar 89,58%, 79,17%, 78,72%, dan 77,05% [15].Berdasarkan latar belakang di atas, pada penelitian ini dilakukan klasifikasi topik *tweet* mengenai COVID. Metode klasifikasi yang digunakan dalam penelitian ini yaitu Multinomial Naïve Bayes dengan pembobotan TF-IDF. Topik *tweet* pada penelitian ini terdiri dari kesehatan, hiburan, hukum, sosial, dan ekonomi.

METODE PENELITIAN

Tahapan klasifikasi topik *tweet* mengenai COVID menggunakan metode Multinomial Naïve Bayes dengan pembobotan TF-IDF dapat dilihat pada Gambar 1. Berdasarkan Gambar 1, penelitian ini dimulai dengan analisis masalah, lalu pengumpulan data *tweet* mengenai COVID.

Data *tweet* tersebut diambil yang mengandung kata kunci 'COVID' dalam bahasa Indonesia dari 9 Juni 2021 sampai 9 Juli 2021. Jumlah *tweet* yang digunakan dalam penelitian ini yaitu 4.909. Setiap *tweet* yang telah dikumpulkan lalu diberikan pelabelan secara manual. Topik *tweet* pada penelitian ini diklasifikasi menjadi lima yaitu hukum, kesehatan, sosial, ekonomi, dan hiburan. Langkah selanjutnya melakukan *preprocessing* meliputi *case folding*, *tokenizing*, normalisasi, *stopword removal*, dan *stemming*. Proses yang dibaca dalam *case folding* yang mengandung "a" sampai dengan "z". Langkah selanjutnya melakukan *tokenizing* di mana pada proses ini memisahkan kalimat yang terdapat pada masing-masing data *tweet* tersebut. Misalnya terdapat simbol "\" atau sebuah *whitespace* maka simbol tersebut akan dihilangkan pada tahap ini.



Gambar 1. Tahapan Penelitian

Proses selanjutnya adalah normalisasi di mana pada proses ini adalah penggantian kata yang tidak sesuai dengan bahasa Indonesia, seperti kata yang tidak baku seperti "rmh", "dmn", "kmn", "kpn", dan lainnya yang tidak sesuai dengan ejaan bahasa Indonesia yang baik dan benar.

Tahap selanjutnya adalah melakukan *stopword removal* yang merupakan tahap penghilangan kata yang tidak penting terdapat dalam topik tersebut seperti "di", "yang", "dan". *Stemming* merupakan penghapuskan kata yang tidak sesuai dengan kata dasar atau dengan kata lain menghilangkan kata imbuhan seperti "di", "nya", "ke". Proses *stemming* pada penelitian ini menggunakan *library* Sastrawi [20].

Pembobotan menggunakan TF-IDF dilakukan setelah tahap *preprocessing*. Pembobotan TF-IDF terdapat dua bagian proses yaitu TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*). TF adalah seberapa sering kata muncul dalam sebuah dokumen yang artinya semakin banyak kata yang muncul pada setiap dokumen maka akan semakin besar pula nilai TF. IDF adalah jumlah dokumen yang mengandung tiap kata tersebut [14]. Rumus pembobotan TF-IDF seperti pada Persamaan (1) sampai Persamaan (3).

$$\text{TF-IDF}\ (d,t) = \text{TF}\ (d,t) * \text{IDF}\ (t)\backslash \quad (1)$$

$$\text{TF-IDF}\ (d,t) = \frac{\text{jumlah kata } t \text{ pada dokumen } d}{\text{total kata pada dokumen } d} \quad (2)$$

$$\text{IDF}\ (t) = log\ \frac{D}{df} \quad (3)$$

dengan t: kata, d: dokumen, $df =$ jumlah dokumen yang mengandung kata t, da n $D =$ jumlah dokumen. Tahap selanjutnya yang dilakukan pada penelitian ini yaitu pelatihan menggunakan Multinomial Naïve Bayes. Data pelatihan sebanyak 3.436 *tweet* atau 70% dari keseluruhan data.

Tahap pelatihan dilakukan perhitungan *prior probability* pada setiap label atau topik menggunakan Persamaan (4).

$$P(c) = \frac{N_c}{N} \quad (4)$$

dengan $P(c)$: *prior probability* kelas *c*, N_c: jumlah kelas *c* pada seluruh dokumen, N: jumlah seluruh dokumen. Nilai dari *prior probability* tersebut digunakan pada tahap pengujian [11].

Tahap pengujian dilakukan perhitungan probabilitas kata di suatu label dengan mempertimbangkan bobot (TF-IDF) dari kata tersebut. Rumus Multinomial Naïve Bayes yang digunakan dengan pembobotan TF-IDF seperti pada Persamaan (5).

$$P(t_n \mid c) = \frac{W_{ct}+1}{(\sum W' \in VW'_{ct})+B'} \quad (5)$$

dengan W_{ct}: nilai pembobotan TF-IDF pada *term t* di kelas *c*, $\sum W' \in VW'_{ct}$: jumlah total bobot dari keseluruhan *term* yang berada di kelas *c*, dan B': jumlah bobot dari *term* yang unik pada seluruh dokumen [11].

Tahap selanjutnya dilakukan perhitungan probabilitas suatu *tweet* masuk ke dalam suatu label dengan menggunakan Persamaan (6).

$$P(c|\textit{term}\ \text{dokumen}\ d) = P(c) \times P(t_1|c) \times P(t_2|c) \times \cdots \times P(t_n|c) \quad (6)$$

dengan $P(c)$: probabilitas *prior* dari kelas *c*, t_n: *term* (kata) ke-*n* pada dokumen *d*, $P(c|term$ dokumen $d)$: probabilitas suatu dokumen *d* berada di kelas *c*, dan $P(t_n|c)$: probabilitas kata ke-*n* pada kelas *c*.

Dalam menentukan label *tweet* pada data pengujian dilihat label dengan nilai probabilitas terbesar [11]. Pengukuran terhadap akurasi klasifikasi topik *tweet* pada tahap pengujian menggunakan *confusion matrix*. Pengujian performa dilakukan untuk mengetahui seberapa akurat metode Multinomial Naïve Bayes dengan pembobotan TF-IDF dalam mengklasifikasikan topik *tweet* mengenai COVID.

HASIL DAN PEMBAHASAN

Pengumpulan data dilakukan melalui media sosial Twitter menggunakan bantuan API Twitter. Data *tweet* diambil menggunakan kata kunci COVID dari 9 Juni 2021 hingga 9 Juli 2021. Data *tweet* mengenai COVID yang telah terkumpul sebanyak 4.909 *tweet*. Contoh hasil dari pengumpulan data dapat dilihat pada Tabel 1. Data yang telah dikumpulkan selanjutnya dilakukan pelabelan secara manual yang terdiri dari 5 label yaitu sosial, hiburan, kesehatan, ekonomi, dan hukum. Contoh hasil dari pelabelan data dapat dilihat pada Tabel 2.

Tabel 1. Contoh Hasil Pengumpulan Data

Username	***Tweet***
Ustadz Bob | Day-455	Awas aja ni klo ga pake masker trus kumpul2 seenak jidat Jangan2 tu antek2nya covid
radiosmartfm959	Menteri Koperasi dan UKM Teten Masduki meminta para pelaku usaha UMKM untuk tidak ragu gabung ke koperasi. Menurutnya, untuk membangkitkan perekonomian Bali yang terdampak pandemi Covid-19, pemberdayaan koperasi dan UMKM harus dilakukan
HUFRIID	Satgas Covid-19 akan melakukan intervesi bagi daerah yg mengalami peningkatan BOR hal ini sebagai wujud Implementasi PPKM Mikro
Heriweb	Tujuh 7 Nakes di Kupang Terpapar Covid-19 Pasca Ikut Acara Perpisahan

Tabel 2. Contoh Hasil Pelabelan Data

Tweet	**Label**
Awas aja ni klo ga pake masker trus kumpul2 seenak jidat Jangan2 tu antek2nya covid	Kesehatan
Menteri Koperasi dan UKM Teten Masduki meminta para pelaku usaha UMKM untuk tidak ragu gabung ke koperasi. Menurutnya, untuk membangkitkan perekonomian Bali yang terdampak pandemi Covid-19, pemberdayaan koperasi dan UMKM harus dilakukan	Ekonomi
Satgas Covid-19 akan melakukan intervesi bagi daerah yg mengalami peningkatan BOR hal ini sebagai wujud Implementasi PPKM Mikro	Sosial
Tujuh 7 Nakes di Kupang Terpapar Covid-19 Pasca Ikut Acara Perpisahan	Kesehatan

Hasil *Preprocessing* Data

Tahap *preprocessing* meliputi *case folding*, *tokenizing*, *stopword removal*, normalisasi, dan *stemming*.

File hasil dari *preprocessing* tersebut disimpan dalam *file* baru .csv atau file Excel .xlsx di mana *file* tersebut digunakan sebagai *dataset* pada tahap pembobotan TF-IDF, pelatihan, dan pengujian. Contoh hasil akhir dari tahap *preprocessing* dapat dilihat pada Tabel 3.

Pembobotan TF-IDF

Tahap ini dilakukan pemberian bobot pada masing-masing kata (token) yang terdapat dalam *dataset* yang telah melalui tahap *preprocessing*. Contoh hasil pembobotan kata dengan TF-IDF dapat dilihat pada Gambar 2.

Tabel 3. Contoh Hasil *Prepocessing*

***Tweet* Asli**	***Tweet* Hasil *Prepocessing***
Awas aja ni klo ga pake masker trus kumpul2 seenak jidat Jangan2 tu antek2nya covid	awas tidak pakai masker terus kumpul enak jidat jangan antek covid
Menteri Koperasi dan UKM Teten Masduki meminta para pelaku usaha UMKM untuk tidak ragu gabung ke koperasi. Menurutnya, untuk membangkitkan perekonomian Bali yang terdampak pandemi Covid-19, pemberdayaan koperasi dan UMKM harus dilakukan	menteri koperasi ukm teten masduki minta pelaku usaha umkm tidak ragu gabung koperasi. turut bangkit ekonomi bali dampak pandemi covid19, daya koperasi umkm harus laku
Satgas Covid-19 akan melakukan intervesi bagi daerah yg mengalami peningkatan BOR hal ini sebagai wujud Implementasi PPKM Mikro	satgas covid19 laku intervensi bagi daerah alam tingkat bor sebagai wujud implementasi ppkm mikro
Tujuh 7 Nakes di Kupang Terpapar Covid-19 Pasca Ikut Acara Perpisahan	tujuh nakes kupang papar covid19 pasca ikut acara pisah

D1	wajib swab wajib vaksin kerja keluar kota
D2	mutasi covid menular laksana prokes
D3	lonjakan covid kumpul beli mcdonals bts

Token	TF			df	$\frac{D}{df}$	IDF = $log\frac{D}{df}$	TF * IDF		
	D1	D2	D3				D1	D2	D3
Wajib	$\frac{2}{7}$ = 0,285	$\frac{0}{5}$ = 0	$\frac{0}{6}$ = 0	1	$\frac{3}{1}$ = 3	log 3 = 0,477	0,135	0	0
Swab	$\frac{1}{7}$ = 0,142	$\frac{0}{5}$ = 0	$\frac{0}{6}$ = 0	1	$\frac{3}{1}$ = 3	log 3 = 0,477	0,067	0	0
Vaksin	$\frac{1}{7}$ = 0,142	$\frac{0}{5}$ = 0	$\frac{0}{6}$ = 0	1	$\frac{3}{1}$ = 3	log 3 = 0,477	0,067	0	0

Gambar 2. Contoh Pembobotan Kata dengan TF-IDF

Hasil Klasifikasi Menggunakan Multinomial Naive Bayes dengan Pembobotan TF-IDF

Contoh hasil klasifikasi menggunakan metode Multinomial Naïve Bayes dengan pembobotan TF-IDF dapat dilihat pada Tabel 4. Berdasarkan Tabel 4, *tweet* ke-1, 2, dan 3 diperoleh label prediksi dan label aktual sama. *Tweet* ke-4 merupakan contoh label prediksi dan label aktual berbeda. Hasil klasifikasi topik *tweet* ke-4 yaitu sosial, sedangkan label aktual yaitu kesehatan. Hasil pengujian klasifikasi topik *tweet* menggunakan Multinomial Naïve Bayes dengan pembobotan TF-IDF yang dinyatakan dalam bentuk *confusion matrix* dapat dilihat pada Tabel 5. Nilai-nilai pada Tabel 5 selanjutnya digunakan untuk menentukan nilai akurasi, *precision*, *recall*, dan F1-*score*. Nilai akurasi, *precision*, *recall*, dan F1-*score* dapat dilihat pada Tabel 6.

Tabel 4. Contoh Hasil Klasifikasi Topik *Tweet*

Tweet	**Aktual**	**Prediksi**
Awas aja ni klo ga pake masker trus kumpul2 seenak jidat Jangan2 tu antek2nya covid	Kesehatan	Kesehatan
Menteri Koperasi dan UKM Teten Masduki meminta para pelaku usaha UMKM untuk tidak ragu gabung ke koperasi. Menurutnya, untuk membangkitkan perekonomian Bali yang terdampak pandemi Covid-19, pemberdayaan koperasi dan UMKM harus dilakukan	Ekonomi	Ekonomi
Satgas Covid-19 akan melakukan intervesi bagi daerah yg mengalami peningkatan BOR hal ini sebagai wujud Implementasi PPKM Mikro	Sosial	Sosial
Tujuh 7 Nakes di Kupang Terpapar Covid-19 Pasca Ikut Acara Perpisahan	Kesehatan	Sosial

Tabel 5. *Confusion Matrix*

		Topik Prediksi				
		Ekonomi	**Hiburan**	**Hukum**	**Kesehatan**	**Sosial**
Topik Aktual	Ekonomi	15	3	1	13	31
	Hiburan	0	184	0	34	96
	Hukum	0	4	3	20	26
	Kesehatan	1	16	1	396	96
	Sosial	3	66	5	155	304

Tabel 6. Hasil Akurasi, *Precision*, *Recall*, dan F1-*score* Pengujian

	Precision	***Recall***	***F-Score***
Ekonomi	0,79	0,24	0,37
Hiburan	0,67	0,59	0,63
Hukum	0,30	0,06	0,10
Kesehatan	0,64	0,78	0,70
Sosial	0,55	0,57	0,56
Akurasi			**0,61**
Rata-Rata	**0,59**	**0,45**	**0,47**

Berdasarkan Tabel 6, akurasi pengujian sebesar 61%, nilai rata-rata *precision* sebesar 59%, nilai rata-rata *recall* sebesar 45%, dan nilai rata-rata F1-*score* sebesar 47%. Kesalahan dalam pengklasifikasian topik disebabkan adanya kemiripan antara kata-kata penyusun suatu topik dengan topik lainnya seperti susunan kata pada label kesehatan, sosial, dan hiburan yang sangat mirip. Selain itu, jumlah *tweet* pada setiap topik tidak seimbang sehingga dapat mempengaruhi hasil klasifikasi.

KESIMPULAN DAN SARAN

Klasifikasi topik *tweet* mengenai COVID menggunakan metode Multinomial Naïve Bayes dengan pembobotan TF-IDF telah berhasil dilakukan pada penelitian ini. Klasifikasi topik *tweet* dilakukan pada *tweet* dalam bahasa Indonesia. *Tweet* diklasifikasikan ke dalam lima topik yaitu kesehatan, sosial, ekonomi, hiburan, dan hukum. Data *tweet* dikumpulkan dari 9 Juni 2021 sampai 9 Juli 2021 dengan kata kunci 'COVID'. Jumlah *tweet* yang digunakan dalam penelitian ini yaitu 4.909 yang terdiri dari 3.436 data pelatihan dan 1.473 data pengujian. Hasil pengujian klasifikasi topik tweet menggunakan metode Multinomial Naïve Bayes dengan pembobotan TF-IDF diperoleh akurasi sebesar 61,23%, rata-rata *precision* sebesar 59%, rata-rata *recall* sebesar 45% dan rata-rata F1-*score* sebesar 47%. Kesalahan dalam pengklasifikasian topik disebabkan adanya kemiripan antara kata-kata penyusun suatu topik dengan topik lainnya seperti susunan kata pada label kesehatan, sosial, dan hiburan yang sangat mirip. Selain itu, jumlah *tweet* pada setiap topik tidak seimbang sehingga dapat mempengaruhi hasil klasifikasi.

Klasifikasi topik *tweet* pada penelitian lebih lanjut dapat menggunakan metode lain sehingga dapat mengetahui metode mana yang memiliki akurasi yang terbaik. Selain itu, *tweet* yang digunakan dalam klasifikasi dapat dikembangkan dengan bahasa lain, tidak hanya dibatasi pada *tweet* bahasa Indonesia. Pembobotan kata dapat menggunakan metode pembobotan lainnya yang dapat meningkatkan akurasi klasifikasi. Analisis lebih lanjut mengenai topik *tweet* tentang COVID yang banyak dibahas oleh masyarakat dapat dilakukan pada penelitian selanjutnya sehingga pemerintah dapat mengetahui upaya atau kebijakan penanganan COVID yang tepat. Pelabelan pada penelitian ini masih dilakukan secara manual dalam menentukan topik setiap *tweet* sehingga dalam penelitian lebih lanjut dapat dilakukan pembuatan korpus untuk menentukan topik dari suatu *tweet*.

DAFTAR PUSTAKA

[1] Twitter, "Pertanyaan umum pengguna baru," Twitter, 2021. [Online]. Available: https://help.twitter.com/id/new-user-faq. [Accessed: Jan. 2, 2022].

[2] N. A. Paramastri dan G. Gumilar, "Penggunaan twitter sebagai medium distribusi berita dan newsgathering oleh tirto.id," Kajian Jurnalisme, vol. 3, no. 1, pp. 18 – 38, 2019.

[3] A. Susilo, C. M. Rumende, C. W. Pitoyo, W. D. Santoso, M. Yulianti, Herikurniawan, R. Sinto, G. Singh, L. Nainggolan, E. J. Nelwan, L. K. Chen, A. Widhani, E. Wijaya, B. Wicaksana, M. Maksum, F. Annisa, C. O. Jasirwan, dan E. Yunihastuti, "Coronavirus disease 2019: tinjauan literatur terkini," Jurnal Penyakit Dalam Indonesia, vol. 7, no. 1, pp. 45 – 67, 2020.

[4] Satuan Tugas Penanganan COVID-19, "Situasi covid-19 di Indonesia (update per 3 Maret 2022)," Satuan Tugas Penanganan COVID-19, 2021. [Online]. Available: https://covid19.go.id/artikel/2022/03/03/situasi-covid-19-di-indonesia-update-3-maret-2022. [Accessed: Mar 15, 2022].

[5] A. N. A. Hiola, A. Asrifuddin, dan F. L. F. G. Langi, "Hubungan antara upaya pencegahan covid-19 dengan angka konfirmasi positif covid19 di Indonesia," Jurnal Kesmas, vol. 11, no. 2, pp. 135 – 142, 2022.

[6] I. Zukhrufillah, "Gejala media sosial twitter sebagai media sosial alternatif," Al-I'lam: Jurnal Komunikasi dan Penyiaran Islam, vol. 1, no. 2, pp. 102 – 109, 2018.

[7] V. Gupta dan G. S. Lehal, "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60 – 76, 2009.

[8] S. H. Liao, P. H. Chu, dan P. Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert systems with applications, vol. 39, no. 12, pp. 11303 – 11311, 2012.

[9] S. V. Gaikwad, A. Chaugule, dan P. Patil, "Text mining methods and techniques," International Journal of Computer Applications, vol. 85, no. 17, pp. 42 – 45, 2014.

[10] I. Rish, "An empirical study of the naive bayes classifier," International Joint Conference on Artificial Intelligence, California, 2006.

[11] C. D. Manning, P. Raghavan, dan H. Schütze, Introduction to information retrieval. New York: Cambridge University Press, 2008.

[12] Rahman, Wiranto, dan A. Doewes, "Online news classification using multinomial naïve bayes," ITSMART: Jurnal Ilmiah Teknologi dan Informasi, vol. 6, no. 1, pp. 32 – 38, 2017.

[13] V. R. S. Nastiti, S. Basuki, dan Hilman, "Klasifikasi sinopsis novel menggunakan metode naïve bayes classifier," Repositor, vol. 1, no. 2, pp. 125 – 130, 2019.

[14] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of Documentation, vol. 60, no. 5, pp. 503 – 520, 2004.

[15] D. H. Kalokasari, I. M. Shofi, dan A. H. Setyaningrum, "Implementasi algoritma multinomial naïve bayes classifier pada sistem klasifikasi surat keluar (studi kasus: diskominfo kabupaten tangerang)," Jurnal Teknik Informatika, vol. 10, no.2, pp. 109 – 118, 2017.

[16] C. S. Sriyano dan E. B. Setiawan, "Pendeteksian berita hoax menggunakan naïve bayes multinomial pada twitter dengan fitur pembobotan tf-idf," e-Proceeding of Engineering, vol. 8, no. 2, 2021, pp. 3396 – 3405.

[17] S. S. Ritonga, E. B. Setiawan, dan I. Kurniawan, "Analisis trending topik pada twitter menggunakan metode naïve bayes dengan pembobotan tf-idf," e-Proceeding of Engineering, vol. 7, No. 1, 2020, pp. 2806 – 2816.

[18] A. Sabrani, I. G. P. W. Wedashwara, dan F. Bimantoro, "Metode multinomial naïve bayes untuk klasifikasi artikel online tentang gempa di Indonesia," Jurnal Teknologi Informasi, Komputer dan Aplikasinya (JTIKA), vol. 2, no. 1, pp. 89 – 100, 2020.

[19] R. M. Furqon dan E. B. Setiawan, "Deteksi berita rumor pada sosial media twitter menggunakan metode naïve bayes multinomial dengan pembobotan tf-idf," e-Proceeding of Engineering, vol. 7, no. 2, 2020, pp. 7916 – 7925.

[20] A. Librian, "High quality stemmer library for Indonesian Language (Bahasa)," github, 2017. [Online]. Available: https://github.com/sastrawi/sastrawi [Accessed: Jan. 15, 2022].