Analisis Sentimen Berdasarkan Topik Terkait Wabah Covid-19 di Twitter Menggunakan *Latent Dirichlet Allocation* (LDA) dan *Naive Bayes Classifier* (NBC)

Pangky Putra Aziztiya*[1], Muhammad Habibi[2], Netania Indi Kusumaningtyas[3]
[1,2]Informatika, FTTI Unjaya, Yogyakarta, Indonesia
[3]Sistem Informasi, FTTI Unjaya, Yogyakarta, Indonesia
e-mail: ***[1]putraazistiya@gmail.com**, [2]muhammadhabibi17@gmail.com, [3]netania0412@gmail.com

Abstract – In 2020 WHO determined that the Corona Virus (COVID-19) was a pandemic. The global spread of the COVID-19 outbreak has made Twitter one of the most widely used tools to publish and find information. This study aims to form a modeling of topics related to the COVID-19 outbreak on the Twitter social media platform and analyze positive and negative sentiments in each topic that has been obtained by combining the two Latent Dirichlet Allocation (LDA) and Naïve Bayes Classification (NBC) methods. Beginning with modeling the topic using the Latent Dirichlet Allocation so that the topics that have been obtained will be searched for the sentiment value of each topic using the Naïve Bayes Classifier method. This study succeeded in combining the two methods with a fairly good accuracy of 89%. In topic modeling, 5 ideal topics were obtained and it can be seen that the most discussed topic is booster vaccination. The results of the classification using NBC show that the topic of booster vaccination has more negative sentiments than positive sentiments.

Keywords *- COVID-19, Latent Dirichlet Allocation, Naïve Bayes Classifier, Text Mining, Topic Modelling*

Abstrak - Pada tahun 2020 WHO menetapkan adanya pandemi yang disebabkan oleh Virus Corona (COVID-19). Meluasnya wabah COVID-19 secara global menjadikan media sosial Twitter menjadi salah satu alat yang paling banyak digunakan untuk mempublikasikan dan mencari informasi. Penelitian ini bertujuan untuk membentuk pemodelan topik terkait wabah COVID-19 di platform media sosial Twitter dan menganalisis sentimen positif dan negatif di setiap topik yang sudah didapatkan dengan mengkombinasikan antara kedua metode Latent Dirichlet Allocation (LDA) dan Naïve Bayes Classification (NBC). Diawali dengan pemodelan topik menggunakan Latent Dirichlet Allocation sehingga topik yang telah didapatkan akan dicari nilai sentimen dari masing-masing topik dengan menggunakan metode Naïve Bayes Classifier. Penelitian ini berhasil mengombinasikan kedua metode dengan hasil akurasi yang cukup baik yaitu 89%. Pada pemodelan topik diperoleh 5 topik ideal dan dapat diketahui topik yang paling banyak dibahas adalah vaksinasi booster. Hasil klasifikasi menggunakan NBC dapat diketahui bahwa topik vaksinasi booster lebih banyak sentimen negatif dibandingkan sentimen positif.

Kata kunci *- COVID-19, Latent Dirichlet Allocation, Naïve Bayes Classifier, Text Mining, Topic Modelling*

I. PENDAHULUAN

Coronavirus Disease 2019 (COVID-19) adalah penyakit menular yang disebabkan oleh *Severe Acute Respiratory Syndrome Coronavirus* 2 (SARS-CoV-2) yang termasuk dalam kategori virus jenis terbaru. Terdapat dua jenis Virus Corona yang sudah diketahui dan dapat mengakibatkan gejala berat seperti *Middle East Respiratory Syndrome* (MERS) serta *Severe Acute Respiratory Syndrome* (SARS). Gejala awal yang sering muncul pada pasien yang terinfeksi meliputi gangguan pernafasan, demam, sakit tenggorokan, batuk, dan sesak nafas. Pada kasus terberat akan mengalami pneumonia, sindrom pernapasan akut, gagal ginjal bahkan kematian. Virus ini pertama kali teridentifikasi di Wuhan, China pada Desember 2019, dan dianggap sebagai pandemi oleh WHO pada Maret 2020. Selama pandemi, virus ini mengalami mutasi, menghasilkan varian-varian jenis baru dengan tipe dan pola transmisi berbeda-beda seperti Alpha, Beta, Gamma, dan Delta. Varian terbaru adalah Omicron yang memiliki tingkat transmisi lebih cepat sehingga penyebaran varian ini lebih cepat [1].

Twitter menjadi *platform* penting dalam menyebarkan informasi dan diskusi seputar COVID-19. Dari banyaknya *tweet* mengenai wabah COVID-19 dapat dilakukan analisis sentimen topik-topik [2] yang terdapat di dalam *tweet* tersebut menggunakan metode *Latent Dirichlet Allocation* untuk melakukan pemodelan topik dan *Naïve Bayes Classifier* untuk analisis sentimen. *Latent Dirichlet Allocation* adalah metode yang membantu dalam proses pengelompokan. Sedangakan *Naïve Bayes Classifier* merupakan metode pengklasifikasian sederhana yang sering digunakan dan mudah untuk diterapkan serta memiliki hasil yang baik pada banyak kasus [3] [4].

Penelitian sebelumnya mengenai COVID-19 telah dilakukan dengan fokus pada pemodelan topik dan analisis sentimen. Penelitian pertama [5] memusatkan pada pemodelan topik menggunakan metode *Naïve Bayes Classifier* dan *Latent Dirichlet Allocation* tanpa mengkombinasikan keduanya. Penelitian kedua [6] menganalisis *tweet* tentang vaksinasi dengan menggunakan *Naïve Bayes Classifier*, mencapai akurasi 93%. Penelitian ketiga [7] mengenai penanganan COVID-19 di Indonesia menggunakan *Naïve Bayes Classifier*, dengan akurasi sebesar 89.13% dan APER

sebesar 10.87%. Penelitian keempat [8] meneliti pengaruh *sampling* terhadap data *tweet* menggunakan *stratified random sampling*. Penelitian kelima [9] mengidentifikasi beragam topik yang dibahas oleh akun bot mengenai COVID-19 dengan menggunakan metode LDA.

Beberapa penelitian hanya menggunakan satu metode atau kombinasi kedua metode tersebut. Penelitian terbaru akan menggunakan kombinasi metode LDA dan NBC serta membuat *web dashboard* dengan fitur *import* data yang berbentuk *file* Microsoft Excel akan menampilkan data yang sudah benar-benar bersih dengan visualisasi yang jelas dan mudah dipahami.

II. METODE PENELITIAN

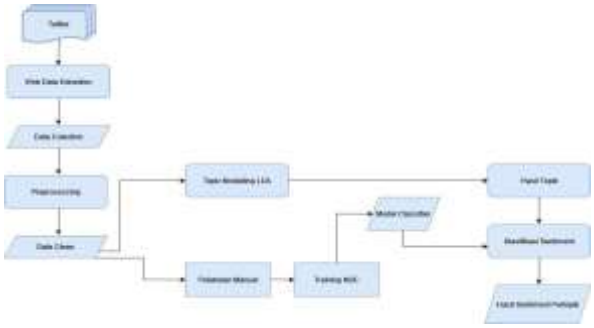
Penelitian ini adalah penelitian pemodelan topik dan dilakukan analisis sentimen pada data Twitter. Penelitian ini akan memakai algoritma *Latent Dirichlet Allocation* (LDA) dan *Naïve Bayes Classification* (NBC). Penelitian ini diawali dari latar belakang permasalahan, melakukan pemrosesan data yang telah diperoleh dan membentuk pemodelan topik serta melakukan pencarian nilai-nilai sentimen yang maksimal sehingga informasi diperoleh sesuai dengan apa yang diharapakan.

A. Bahan Penelitian

Bahan yang akan digunakan pada penelitian ini adalah data *tweet* maupun komentar di Twitter yang berfokus pada wabah Virus Corona (COVID-19).

B. Jalan Penelitian

Penelitian ini menggunakan Bahasa pemrograman Python, Anaconda 3 dan Jupyter Notebook untuk melakukan pengambilan data yang akan di tampilkan di Microsoft Office Excel dan akan dilakukan *Preprocessing* data, *feature extraction*, dan melakukan pemodelan topik serta proses klasifikasi sentimen dari setiap topik yang sudah didapatkan secara otomatis melalui *website dashboard* yang akan dibuat menggunakan bahasa pemrograman Python . Adapun tahap pada jalan penelitian ini terdapat pada Gambar 1.



Gambar 1. Alur penelitian

1) *Web Data Extraction*

Pada tahap ini akan mengambil data dan mengumpulkan data *tweet* dari Twitter mengenai wabah COVID-19 dengan menggunakan Anaconda Promt dan akan diproses di Jupyter Notebook yang akan menghasilkan kumpulan dokumen berbentuk *file* Microsoft Excel.

2) *Data Collection*

Data yang sudah terkumpulkan dengan 3 kata kunci: Vaksinasi, Omicron dan Delta memiliki jumlah data keseluruhan 15.004 dilakukan penghilangan *duplicate* data menjadi 10.300 yang akan digunakan untuk proses analisis dan belum dilakukan *preprocessing* data, serta tersimpan dalam bentuk *file* Excel sehingga data tersebut belum dapat digunakan untuk pengolahan data lebih lanjut, data tersebut dapat dilihat pada Tabel 1.

Tabel 1. *Data Collection*

No	*Text*
1.	Peneliti di India menemukan subvarian Omicron BA.2.75 yang merupakan turunan dari BA.2. Kecepatan penyebarannya diperkirakan 9 kali dari varian Delta. https://t.co/DfLomb9rI2
2.	Peneliti Kembangkan Rapid Test Identifikasi Varian Delta atau Omicron https://t.co/EZk9kxlM05 https://t.co/nPGS5dJWEF
3.	studi terbaru mrk yg mendapatkan vaksin booster pfizer & astrazeneca. peneliti menemukan penerima dua dosis sinovac dgn booster vaksin dari platform vektor (johnson & johnson n astrazeneca) serta mRNA (pfizer n moderna) dpt lebih kuat utk melawan varian seperti delta n omicron. https://t.co/01dhHNNhCK
4.	@dutaSherliana Hasil penelitian menunjukkan bahwa korona varian omicron lebih menular daripada Delta #BahanPokokStabil #BBMTerkendali
5.	#Paripurna44 Menkeu RI, Sri Mulyani: ...Kasus yang ekstrem, Pemerintah menerapkan kebijakan PPKM Darurat di sebagian besar wilayah NKRI. Untuk merespons dan mengantisipasi dampak varian Delta tersebut, Pemerintah menaikkan alokasi Program PC-PEN menjadi Rp744,8 triliun. (3)
6.	@TheMaximvs09 Kalau warga negaranya tetap disiplin prokes dan sudah full vaksin, harusnya ga terlalu ngeri dampaknya.. kayak kemaren omicron aja, menyebar cepat tapi tidak terlalu mematikan severti varian delta, karena sudah banyak yg vaksin
7.	@rasyilaa Itu kayaknya lo yg delta ya yg sampe anosmia?gue untungnya ga anosmia sih jadi kemungkinan ini yang varian baru. Thanks yaaaa. Surprise bgt ini org covid udh ilang eh malah gue kena setelah 2 tahun aman aman aja ðŸ¤£
8.	@swextdemon Bener-bener sekeluarga sakit, kita bingung mau ngurus bapak gimana karena kita juga sakit banget varian delta gak main-main. Inget banget di ugd saturasi bapakku cuma 59, dan subuhnya bener-bener ngedrop tapi masih ngigo. Tahun 2021 bener-bener berat banget.
9.	@hellodeobi Pas thn lalu aku kena juga sama sih kak efeknya sebulanan.. ðŸ˜žini kemungkinan adek kena varian yg delta ditambah sm yg varian baruðŸ˜ž
10.	Pada saat ini, hanya Omicron yang masuk daftar varian mengkhawatirkan WHO. Sebelumnya, alfa, beta, gamma dan delta juga termasuk. https://t.co/x37hAzr5uT

3) *Prepocessing*

Sebelum masuk ke tahap *preprocessing* data masukkan *library* serta modul yaitu nltk, emoji, nltk.tokenize, stopwords untuk melakukan *preprocessing.*

a) *Cleaning*

Tahapan untuk menghilangkan atribut yang tidak diperlukan seperti tag, emoji, dan tag url. Setelah dilakukan *cleaning* pada data yang tersimpan dalam data *collection* tadi data akan terlihat bersih dan dapat dilihat pada Tabel 2.

Tabel 2. *Data Cleaning*

No	***Cleaned_Text***
1.	hasil teliti corona varian omicron tular delta
2.	kayak pas varian delta
3.	potensi gelombang subvarian omicronn berat varian delta
4.	istilah varian virus covid
5.	covid varian delta omicron dari universe
6.	varian delta ancam anak muda
7.	semenjak covid varian delta tahun yang lalu poko sakit minum obat enggak bisa sembuh cuman istirahat doang
8.	varian delta pensiun kah min
9.	biar varian omicron sembuh kalau delta bahaya
10.	total covid varian omicron indonesia lampau delta

b) *Tokenizing*

Tahapan ini berfungsi untuk memisahkan sebuah kata dari kalimat agar menjadi kata tunggal agar kata tersebut dapat berdiri sendiri.

c) *Casefolding* dan *Stopwords*

Menghilangkan kata-kata yang tidak di perlukan, atribut-atribut yang sering muncul dan merubah semua huruf menjadi huruf kecil. Perintah *Casefolding* digunakan untuk merubah kata-kata menjadi huruf kecil sedangkan perintah *Stopwords* untuk menghilangkan kata-kata yang tidak perlu digunakan.

d) *Stemming*

Proses untuk merubah semua kata menjadi kata dasar. *Stemming* dilakukan dengan menghilangkan imbuhan yang terdapat pada setiap kata.

e) *Normalization*

Proses untuk merubah kata-kata yang disingkat menjadi sebuah kata yang utuh seperti : “yg” menjadi “yang”, “dng” menjadi “dengan” dan lain-lain.

4) *Data Clean*

Tahapan dimana data sudah benar-benar bersih dan sudah dapat digunakan untuk proses selanjutnya. Hasil dari data yang sudah di *preprocessing* akan di simpan atau di unduh dalam bentuk *file* Excel dan hasilnya dapat dilihat pada Tabel 3.

Tabel 3. Daftar *Data Clean*

No	***Cleaned_Text***
1.	hasil teliti corona varian omicron tular delta
2.	kayak pas varian delta
3.	potensi gelombang subvarian omicronn berat varian delta
4.	istilah varian virus covid
5.	covid varian delta omicron dari universe
6.	varian delta ancam anak muda
7.	semenjak covid varian delta tahun yang lalu poko sakit minum obat enggak bisa sembuh cuman istirahat doang
8.	varian delta pensiun kah min
9.	biar varian omicron sembuh kalau delta bahaya
10.	total covid varian omicron indonesia lampau delta

5) *Topic Modelling* Menggunakan LDA

Setelah melakukan *preprocessing* data akan dilakukan pemodelan topik menggunakan algoritma LDA dengan *library* memasukkan library gensim, lalu menyimpan modelLDA ke dalam *file* modelLDA 30.gensim, serta menampilkan model LDA. Pada tahap ini juga akan menampilkan topik yang telah didapatkan dari perulangan dari “for topic,words in topics_words:” dan “print(str(topic)+ "::"+ str(words))”.

a) *Visualisasi*

Topik-topik yang telah didapatkan dari tahapan *topic modelling* akan di visualisasikan dengan *library* dan *file* yang telah tersimpan. Pada tahap ini akan menggunakan modul *library* pyLDAvis.gensim, serta memuat *file* dengan dictionary_lda.gensim, corpus_lda.pkl dan modelLDA_5.gensim, akan ditampilkan *topic modelling* dari *file* yang telah dimuat. Selain itu, topik dapat divisualisasikan dengan kata-kata dalam topik, di mana ukuran setiap kata akan menampilkan frekuensi atau pentingnya kata dalam topik.

Untuk menampilkan *wordcloud* dari keseluruhan topik maka menggunakan modul *library path*, *image*, *wordCloud*, *stopwords*, *imagecolorgenerator*, serta matplotlib.pyplot serta *wordcloud* tersebut akan disimpan dengan nama *file* wordcloud_all.png. Sedangkan *wordcloud* masing-masing topik akan disimpan dengan nama *file* your_file_name.png.

b) Analisis

Pada tahap akhir, kualitas topik yang dihasilkan dari *topic modelling* diuji dengan menggunakan hasil *topic coherence* yang disajikan dalam grafik menggunakan *variable* ldatopics dan memanggil topik dari proses *topic modelling.*

Dalam tahap ini menggunakan *library* CoherenceModel, LdaModel, LsiModel, HdpModel serta membentuk fungsi untuk memproses *topic coherence* dari “def evaluate_graph(dictionary, corpus, texts, limit):” sampai “return lm_list, c_v”. Selain itu menampilkan grafik dari *topic coherence* sehingga dapat diketahui jumlah topik idealnya dengan jumlah grafik tertinggi.

6) Pelabelan Manual

Pelabelan manual adalah proses memberikan label terhadap kalimat dan kata yang terdapat pada sebuah dokumen sehingga nantinya dapat dilakukan analisis lebih lanjut mengenai sifat dari kalimat atau kata tersebut apakah memiliki sifat positif atau negatif. Dari data *tweet* yang sudah dilabeli pada tahap ini maka telah di dapatkan data *training* sebanyak 1017 data *tweet* dengan jumlah 518 data *tweet* positif dan 499 data *tweet* negatif. Hasil dari pelabelan manual dapat dilihat pada Gambar 2.

	No	Cleaned_Text	label	kelas
0	3	tenlivoyez kasih saran pake parfum delta varia...	negatif	0
1	23	did guys notice suara sirene ambulans frekuens...	negatif	0
2	29	swextdemon aa sorry for your loss kak jd ingat...	negatif	0
3	60	listy vaksin gejala flu doang coba kalau varia...	negatif	0
4	69	pansos status pandemi kampanye harap influence...	negatif	0
...	...	...	...	...
1012	10145	dapat belanja dapat vaksinasi covid gratis loh...	positif	1
1013	10192	duduk besar dunia gt juta persentase cakup vak...	positif	1
1014	10208	monitoring giat vaksinasi covid singkawang gra...	positif	1
1015	10265	pacpeers tarik project data topik analisis pro...	positif	1
1016	10266	covid world vaccination progress analysis prog...	positif	1

1017 rows × 4 columns

Gambar 2. Data Pelabelan Manual

Pelabelan manual ini dilakukan untuk memberi nilai sentimen terhadap kelas positif maupun negatif yang nantinya akan dihitung nilai akurasinya.

7) *Naïve Bayes Classifier*

Dalam pengolahan data menggunakan metode *Naïve Bayes Classifier* (NBC) yang melewati beberapa tahap seperti pelabelan manual, *feature extraction* TF-IDF, *training*, *testing*, dan klasifikasi data keseluruhan.

8) *Data Training*

Proses data *training* dimulai dengan ekstraksi pada data teks menggunakan TF-IDF, pada tahap ini digunakan pendekatan *Naive Bayes Classifier*, dan dilanjutkan dengan proses *training* data untuk membentuk model klasifikasi yang disimpan dengan format pickle yang dapat digunakan untuk mengklasifikasikan sentimen data keseluruhan secara otomatis. Perhitungan TF-IDF juga dapat dilakukan secara manual melalui Microsoft Office Excel. Klasifikasi pada penelitian ini menggunakan fitur ekstraksi TF-IDF yang memberikan hasil perhitungan secara otomatis pada pembobotan kata pada setiap dokumen. *Library* yang digunakan adalah sklearn.feature_extraction.text serta TfidfVectorizer untuk melakukan perhitungan secara otomatis. Dibantu dengan *library Multinomial Naïve Bayes* yang akan membantu untuk mengklasifikasi teks pada sebuah data di data *training*.

Setelah dilakukan perhitungan TF-IDF, selanjutnya dilakukan pencarian akurasi dari data *training* yang sebelumnya sudah dilakukan pelabelan manual tujuannya untuk mengetahui keakuratan dari sebuah dokumen atau data tersebut. *Cros-validation* adalah metode untuk mendapatkan hasil akurasi yang dihitung sampai beberapa kali dengan parameter yang sama, dengan membagi dua data yaitu data latih dan data uji memakai *library* from sklearn.model_selection serta *import* ShuffleSplit untuk melakukan perhitungan rata-rata dalam 10 kali.

Tahap selanjutnya adalah melakukan perhitungan keakuratan pemodelan yang digunakan untuk memprediksi label (kelas) dari data yang sudah tersedia. Dilanjutkan dengan melakukan pembuatan model klasifikasi dengan variabel X dan Y dengan data *training* yang sudah di proses sebelumnya. Model tersebut dibuatkan sebuah fungsi agar proses pemanggilannya lebih mudah dijalankan untuk tahap berikutnya sehingga akan lebih efektif dan efisien.

Dalam proses pembentukan model klasifikasi menggunakan *library* sklearn.pipeline dengan *import pipeline* yang fungsinya *testable* pada proses *cross-validation*, lalu *import* pickle untuk menyimpan serta membaca *file* berformat .pkl. Model yang diberikan nama text_classifier akan tersimpan ke dalam bentuk file.pickle sehingga nantinya dapat dibuka serta digunakan kembali. Bentuk *file* pickle selanjutnya akan digunakan untuk menjalankan *data testing* dengan jumlah 425 *tweet* sudah dilabeli secara manual dari jumlah data *training* 1017 *tweet* serta 425 *tweet data testing* dari total data 10.300 *tweet* yang sudah dilakukan filtering data di Microsoft Office Excel.

Setelah melakukan pemanggilan *dataframe* maka langkah selanjutnya adalah melakukan prediksi dari metode Naïve Bayes. Hasil prediksi yang sudah didapatkan akan disimpan didalam variabel *result_tweet* dalam sebuah bentuk data *list*.

9) *Testing*

Tahapan ini untuk menentukan akurasi dari model yang telah dibuat pada tahapan *training*, bertujuan untuk menentukan label (kelas) dari *data testing* yang telah disediakan. Maka akan ditampilkan penggabungan *data testing* yang dilabeli secara manual (*actual*) dan dari metode *Naïve Bayes* (*predicted*) yang terdapat pada Gambar 3.

	Cleaned_Text	actual	predicted
0	alhamdulillah varian delta tingkat	0	0
1	catat personal covid gelombang varian delta ta...	1	1
2	saya varian delta juli tahun beneran tidak ban...	1	0
3	alhamdulillah varian delta tingkat rumah sakit...	1	0
4	teliti india temu subvarian omicron ba turun b...	0	0
...	...	...	...
419	lindung masyarakat covid tingkat beri vaksinas...	1	1
420	dinas sehat dinkes purbalingga jawa tengah jat...	0	0
421	polisi sektor polsek cempaka putih gelar vaksi...	1	1
422	cepat vaksinasi covid wilayah koramil margorej...	1	1
423	iptu sukresno kapolsek donorojo camat donorojo...	1	1

Gambar 3. Penggabungan *Data Testing*

Selanjutnya akan dilakukan pencarian nilai perbandingan model dengan *confusion matrix* terdapat *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *True Negative* (TN). Nantinya hasil dari prediksi perbandingan tersebut akan dilakukan perhitungan menggunakan Microsoft Office Excel sehingga akan mendapatkan nilai akurasi dari *data testing* tersebut.

10) Klasifikasi

Setelah didapatkan nilai akurasi yang baik dari proses *training* dan proses *testing*, lalu dilakukan proses klasifikasi data keseluruhan dengan jumlah 15.004 data yang sudah dilakukan *filtering* di Microsoft Office Excel menjadi 10.300 data bersih yang digunakan. Data tersebut sebelumnya sudah di proses ke dalam *classification* metode *Latent Dirichlet Allocation* sehingga mendapatkan data yang sudah terdapat penyebaran topiknya serta nilai *scorenya*. Selanjutnya data tersebut akan dilakukan pencarian nilai sentimen

dari setiap topiknya menggunakan metode *Naïve Bayes Classifier* sehingga menghasilkan nilai sentimen pertopik.

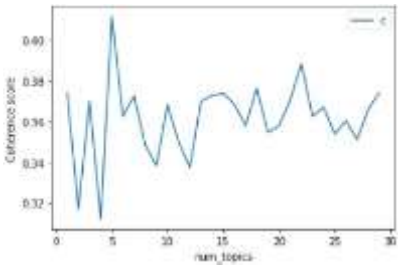
III. HASIL DAN PEMBAHASAN

A. Ringkasan Hasil Penelitian

Hasil penelitian dari penerapan dua metode yang dikombinasikan yaitu *topic modelling* menggunakan metode algoritma *Latent Dirichlet Allocation* (LDA) dengan analisis sentimen menggunakan metode *Naïve Bayes Classifier* (NBC) untuk menganalisis topik serta mencari nilai sentimen dari setiap topik yang sudah didapatkan mengenai wabah COVID-19 di Twitter berdasarkan tiga kata kunci yaitu Vaksinasi, Omicron dan Delta. Jangka waktu pengambilan data untuk vaksinasi di ambil dari tanggal 3 Januari 2021 - 4 Juli 2022 dengan jumlah data 5000 *tweet*, sedangkan untuk kata kunci delta dari tanggal 11 Oktober 2021 - 4 Juli 2022 dengan jumlah pengambilan datanya 5000 data *tweet* dan untuk kata kunci omicron dari tanggal 24 November 2021 - 4 Juli 2022 dengan jumlah 5000 data *tweet*. Dari ketiga kata kunci tersebut keseluruhan datanya dilakukan penggabungan sehingga jumlah datanya menjadi 15.000 data *tweet* dan dilakukan penghilangan *duplicate* data sehingga jumlah datanya menjadi 10.300 data *tweet*. Data tersebut digunakan untuk proses *training* dan *testing* pada metode *Naïve Bayes Classifier* dengan jumlah masing-masing data 1017 data untuk *training* yang sudah dilabeli secara manual dan 425 data untuk *testing* yang sudah dilabeli juga secara manual. Pada proses *training* dan *testing* NBC di dapatkan akurasi *training* 83.97% dan *testing* 89% dan didapatkan 5 topik ideal pada pembentukan *topic modelling* menggunakan metode LDA. Berikut ini pemaparan *topic modelling* dan analisis sentimen pada kasus COVID-19 berdasarkan pendapat masyarakat di Twitter menggunakan bahasa pemrogramman Python dengan mengkombinasikan kedua metode LDA dan NBC, dengan bantuan beberapa *library* yang sudah ada pada bahasa pemrogramman Python seperti pandas, numpy, nltk, emoji, sastrawi, sklearn, pickle, matplotlib, pyLDAvis, genism, corpora, PIL, os yang proses pengolahan datanya dilakukan di Jupyter Notebook.

B. Pembahasan Hasil *Topic Coherence*

Untuk menemukan jumlah topik yang optimal, maka akan menggunakan metode *topic coherence* untuk mengukur keterkaitan kata di dalam topik yang ada.

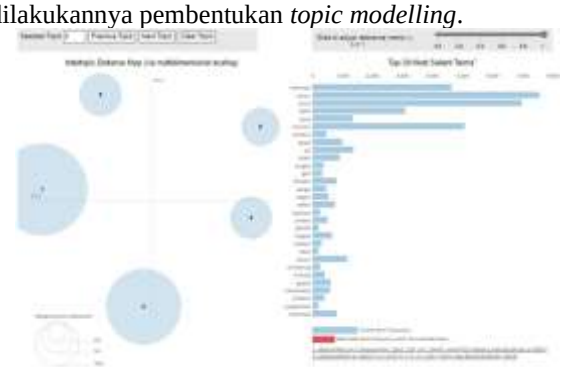


Gambar 4. Grafik Diagram Pada *Topic Coherence*

Grafik diagram pada Gambar 4 akan digunakan untuk menentukan jumlah topik dalam proses pemodelan topik dari data *tweet* yang sudah didapatkan dari Twitter. Dari jumlah topik yang sudah terlihat pada grafik nantinya akan dibandingkan berdasarkan nilai titik *coherence score* yang paling tinggi. Dapat dilihat *coherence score* yang paling tinggi terdapat pada nilai 0.41162116489168776 dengan jumlah topik ada 5.

C. Hasil *Topic Modelling*

Jumlah topik yang sudah didapatkan dari metode *topic coherence* terdapat lima topik ideal lalu akan diproses lagi menggunakan metode LDA. Hasil *topic modelling* dengan jumlah lima topik akan dilakukan visualiasi ke dalam bentuk *Inter Distance Map* yang dapat dilihat pada Gambar 5. Berdasarkan Gambar 5 pada jumlah topik lima hampir keseluruhan topik memiliki jarak sehingga jumlah topik tersebut dapat digunakan atau jumlah topik yang ideal untuk dilakukannya pembentukan *topic modelling*.



Gambar 5. *Topic Modelling* Berdasarkan 5 Topik

Berdasarkan Gambar 5 pada jumlah topik sepuluh sedikit topik yang memiliki jarak diantara satu sama lain dan pada jumlah topik lima hampir keseluruhan topik memiliki jarak. Maka data yang akan digunakan untuk proses selanjutnya adalah yang berjumlah lima topik dikarenakan data tersebut memiliki jarak diantara keseluruhannya.

D. Hasil *Wordcloud* Pertopik

Topik-topik yang sudah didapatkan melalui metode *topic coherence* dari setiap kata-kata yang terdapat dalam topik tersebut akan divisualisasikan kedalam bentuk *wordcloud* fungsinya untuk menampilkan frekuensi dari setiap kata yang sering mucul.

Gambar 6. *Wordcloud* Topik Ke-1

Berdasarkan Gambar 6, dapat dilihat kata yang sering muncul dalam topik 1 adalah "Varian", "Omicron", "Delta" dan "Covid". Maka dari itu dapat disimpulkan inti dari seluruh kata yang sering muncul adalah mengenai varian covid-19 seperti delta dan omicron.

Gambar 7. *Wordcloud* Topik Ke-2

Berdasarkan Gambar 7 dapat dilihat bahwa kata yang sering muncul dalam topik 2 adalah "Vaksinasi", "Covid", "Babinsa", "Warga", "Laksanakan", "Monitoring". Maka dapat disimpulkan segmen topik 2 ini membahas mengenai babinsa melaksanakan monitoring vaksinasi covid kepada warga.

Gambar 8. *Wordcloud* Topik Ke-3

Pada Gambar 8 dapat dilihat bahwa kata yang sering muncul dalam topik 3 adalah "Varian", "Covid", "Delta" dan "Langka". Maka dapat disimpulkan segmen topik 3 ini membahas mengenai pada saat varian covid delta dan omicron masker dan minyak goreng langka.

Gambar 9. *Wordcloud* Topik Ke-4

Pada Gambar 9 dapat dilihat bahwa kata yang sering muncul dalam topik 4 adalah "Covid", "Vaksinasi", "Jakarta" dan "Giat" serta "Kota". Maka dapat disimpulkan segmen topik 4 ini membahas bahwa polres mendampingi vaksinasi covid di kabupaten kota maupun desa.

Gambar 10. *Wordcloud* Topik Ke-5

Berdasarkan Gambar 10 dapat dilihat bahwa kata yang sering muncul dalam topik 5 adalah "Covid", "Vaksinasi", "Omicron" dan "Varian" serta "Indonesia". Maka dapat disimpulkan segmen topik 5 ini membahas bahwa Indonesia laksanakan vaksinasi covid varian omicron.

E. Hasil Analisis Pertopik

Dari hasil penelitian mengenai metode *topic modelling* untuk melakukan analisis topik dari setiap topik mengenai wabah COVID-19 yang pengambilan datanya sudah ditentukan dibagian latar belakang maka diperoleh topik dengan menggunakan lima jumlah topik.

Tabel 4. Kata dalam pertopik

No Topik	Kata-kata	Pembahasannya
1.	'Delta', 'Varian', 'Omicron,' 'Covid' 'Gejala', 'Bahaya', 'Tahun','Bahaya' 'Haji', 'Mati', 'puncak', 'orang-orang' 'tular', 'tinggi', 'sakit'	Mengenai bahaya varian Covid delta dan omicron
2.	'vaksinasi','monitoring','babinsa', 'laksanakan' 'gencar', 'polda', 'camat' 'polsek','hasil', 'puskesmas', 'jadwal', 'koramil', 'lurah' 'covid'	Babinsa gencar laksanakan monitoring vaksin covid
3.	'covid', 'delta', 'langka', 'minyak goreng' 'masker' 'omicron' 'alfa', 'varian', 'tuntas', 'oksigen' 'omicron', 'tahun'. 'polri', 'ekonomi' 'langka'	Pada saat varian covid delta dan omicron terjadi kelangkaan minyak goreng, masker dan oksigen
4.	'jakarta', 'kota', 'giat', 'damping', 'vaksinasi', 'kabupaten', 'juni', 'senin' 'polresta', 'desa', 'balai', 'sabtu'	Pelaksanaan vaksinasi disetiap kabupaten kota jakarta
5.	'vaksinasi', 'booster, 'varian', 'omicron' 'covid','sebar', 'sehat','tingkat' 'masyarakat','dosis,'cegah, 'indonesia','lindungi','ayo',' waspada', 'perintah','virus'	Pelaksaan vaksinasi tingkat booster di Indonesia lindungi masyarakat dari covid varian omicron

Berdasarkan dari jumlah lima topik pada Tabel 4 terdapat kesamaan bahkan ada juga yang saling berhubungan yaitu:

1. Pada topik 2 dan 4 membahas mengenai pelaksaan vaksinasi COVID-19
2. Pada topik 4 dan 5 membahas mengenai vaksinasi booster tingkat kabupaten kota untuk melindungi masyarakat Indonesia dari wabah COVID-19 varian delta maupun omicron
3. Pada topik 1 dan 5 membahas mengenai bahayanya varian delta dan omicron sehingga dilaksanakan vaksinasi tingkat 3 yaitu vaksinasi booster
4. Pada topik 1 dan 3 membahas mengenai varian delta dan omicron sehingga pada saat varian ini menyebar terjadi kelangkaan oksigen, masker bahkan minyak goreng

F. Hasil *Wordcloud* Keseluruhan Topik

Dari topik yang sudah ada, setiap kata yang berada didalam topik akan dikumpulkan menjadi satu lalu akan divisualisasikan dalam bentuk *wordcloud* untuk menampilkan susunan kata yang memiliki frekuensi yang sering muncul.

Gambar 11. *Wordcloud* Keseluruhan Topik

Dari *wordcloud* pada Gambar 11 dapat dilihat bahwa kata yang sering muncul dari semua segmen topik adalah "vaksin", "tahun", "covid", "delta", "omicron", "covid", "cepat", "gelombang", "sampai". Dapat disimpulkan bahwa inti dari keseluruhan kata yang sering muncul adalah pelaksaan cepat vaksin untuk mengatasi wabah covid varian omicron maupun delta.

G. Evaluasi dan Hasil Klasifikasi

Pada model klasifikasi *training* dibutuhkan jumlah 1017 data dimana masing-masing data sudah dibagi menjadi positif berjumlah 518 data dan negatif berjumlah 499 data. Selanjutnya menguji *k-fold cross validation* dapat dilakukan angka perulangan sebanyak 10 kali agar mendapatkan angka yang sesuai. Dari perhitungan tiap fold yang ada pada *k-fold cross validation* sudah terdapat nilai masing-masing dan hasil yang berbeda pada tiap foldnya. Dari perhitungan *cross validation* menghasilkan nilai rata-rata akurasi yang baik yaitu 83,98 untuk *accuracy* dan untuk *f-measure* 83,95.

Setelah melakukan penghitungan *cross validation* maka selanjutnya dapat menghitung *confusion matrix* untuk menguji keakuratan algoritma yang dibangun. Melalui *confusion matrix* dapat diketahui nilai data aktual dan prediksi yang dapat di lihat pada Tabel 5.

Tabel 5. *Confusion Matrix* pada *Cross Validation*

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	87	13
Negatif	19	84

Hasil *confusion matrix* pada Tabel 5 mendapatkan hasil dengan TP = 87, TN = 84, FP = 13 dan FN = 19. Setelah mendapatkan nilai *confusion matrix* serta nilai dari 10 kali *k-fold cross validation* maka hasil dari akurasi data *training* untuk *accuracy* mendapatkan nilai yang baik yaitu 83.82% dan untuk *f-measure* 83.95%.

Pada tahap selanjutnya akan dilakukan proses *data testing* dengan jumlah data 425 yang sudah dilabeli secara manual. Serta melakukan perhitungan *confusion matrix* untuk mengetahui nilai akurasi pada tahapan proses *data testing* dan untuk dapat membedakan nilai-nilai dari data *training* maupun *data testing*. Hasil perhitungan *confusion matrix* pada *data testing* dapat dilihat pada Tabel 6.

Tabel 6. *Confusion Matrix* pada *Data Testing*

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	211	26
Negatif	17	170

Maka akan dilakukan perhitungan *data testing* dengan data yang sudah diketahui TP = 211, TN = 170, FP = 26 dan FN = 17 dapat di hitung menggunakan rumus pada persamaan (3) untuk menghitung *accuracy* serta menghitung *f-measure* menggunakan rumus pada persamaan (6). Nilai dari *data testing* seperti *accuracy* dan *f-measure* didapat dari jumlah data *testing* yang berjumlah 425 data yang sudah dilabeli secara manual menghasilkan nilai *data testing* sebesar 89% dan 90%. Sehingga dapat dilihat pada proses data *training* dan *testing* memiliki nilai yang baik dan signifikan.

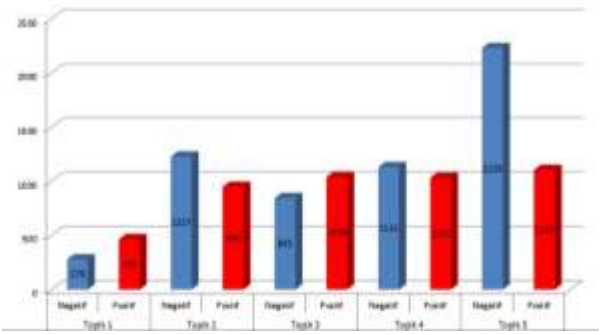
H. Hasil Klasifikasi Data Keseluruhan

Setelah melakukan proses *training* dan *testing* telah didapatkan hasil akurasi yang baik maka selanjutnya dilakukan proses klasifikasi dari data keseluruhan dengan jumlah 10.300 data. Data tersebut sudah di bagi untuk *training* sebesar 1017 data yang sudah terlabeli secara manual serta 425 data untuk proses *testing* yang juga sudah terlabeli secara manual. Pada proses *testing* dan *training* sebelumnya data yang digunakan belum terdapat *topic modelling* jadi data yang akan digunakan untuk proses klasifikasi keseluruhan yaitu data yang sudah di proses di metode LDA sehingga data tersebut sudah terbagi topik-topiknya dan akan di proses dengan model prediksi pada metode NBC sehingga dapat dilihat hasilnya pada Tabel 7.

Tabel 7. Hasil Klasifikasi Keseluruhan

No	Tweet	Topic	Class	Result_Nbc
1	covid varian delta india booming wakanda terima tangan buka pesawat charteran dari india sekarang pmk hewan ternak gara-gara buka import sapi india	0	1	positif
2	kemenkes lapor temu subvarian omicron ba ba indonesia dampak sehat hebat varian delta masyarakat harap berhatihati laksana protokol sehat ketat	1	0	negatif
3	anak covid tahun pas varian delta tidak yang bisa ngungkapin lihat bayi lemas tidak makan jadi kurus hrs maksa mum macem obatvit	2	0	negatif
4	kepala sub bidang dukung sehat satgas covid brigjen tni pur alexander ginting varian delta sirkulasi indonesia hadap gelombang tiga akibat omicron	3	1	positif
5	data sebar varian sarscov provinsi milik satgas covid varian delta sirkulasi tengah masyarakat	4	1	positif

Pada Tabel 7 dapat dilihat beberapa contoh *tweet* yang sudah diklasifikasikan berdasarkan topik dan sentimennya. Penentuan pemodelan topiknya dimulai dari angka 0 - 4 sehingga jumlah *topic modelling* terdapat 5 topik yang ideal untuk digunakan. Hasil dari keseluruhan data dapat di lihat pada Gambar 12.



Gambar 12. Hasil Diagram Dari Keseluruhan Data

Pada Gambar 12 dapat dilihat topik dan masing-masing jumlah sentimennya. Dari keseluruhan jumlah data yang sudah terlihat dan terbagi jika di jumlahkan akan menghasilkan jumlah data yang sesuai yaitu 10.300 data yang sudah terdapat hasil sentimen pertopik.

I. Web Dashboard

Web dashboard ini dibuat dengan menggunakan Python Flask [10] yang menampilkan hasil data yang sudah diolah di Jupyter Notebook berupa data *file* Exel.

Gambar 13. Tampilan awal dashboard

Gambar 13 merupakan tampilan awal *web dashboard*, selanjutnya mengupload data yang sudah didapatkan dari proses web *scraping* lalu melakukan *preprocessing* dan akan menampilkan hasil pada Gambar 14.

Gambar 14. Bagian Dataset

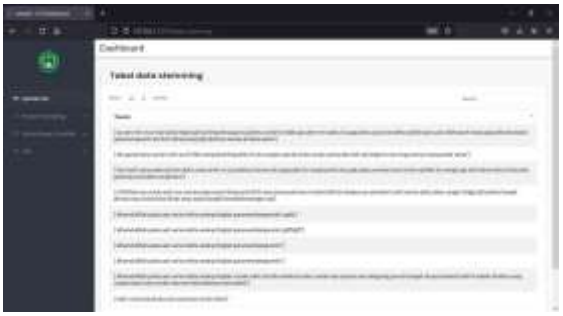
Pada Gambar 15 dapat dilihat Tabel *Tokenizing* yang berfungsi untuk membagi teks menjadi suatu kalimat yang bermakna.

Gambar 15. Tabel *tokenizing*

Pada Gambar 16 merupakan proses *stopwords* untuk filtering kata-kata yang tidak perlu digunakan untuk proses pengolahan data selanjutnya.

Gambar 16. Tabel data *stopwords*

Pada Gambar 17 dapat dilihat merupakan Tabel *Stemming* untuk mengubah kata menjadi kata dasar.



Gambar 17. Tabel data *stemming*

Gambar 18. Tabel data *clean*

Pada Gambar 18 merupakan proses akhir dari *preprocessing*, sehingga data sudah bersih dan sudah dapat digunakan untuk proses pengolahan data selanjutnya. Proses *data training* dapat dilihat pada Gambar 19.

Gambar 19. Tabel data *training*

Pada Gambar 19 merupakan proses pembentukan model pickle, nantinya model tersebut akan digunakan untuk membuka model untuk proses klasifikasi data keseluruhan.

Gambar 20. Pemodelan pickle

Pada Gambar 21 proses penampilan hasil *testing* dapat dilihat jumlah perhitungan nilainya yang sesuai dengan proses yang ada pada Jupyter Notebooknya, selanjutnya masuk ke tahapan LDA. Pada tahapan LDA ada beberapa tampilan yang akan diperlihatkan seperti pada Gambar 22 dan Gambar 23.

Gambar 21. Hasil *testing*

Gambar 22. Tabel *result* topik

Gambar 23. *Wordcloud* pertopik

Pada Gambar 24 proses akan di kembalikan ke metode Naïve Bayes Classifier yang dilakukan proses *training* dan *testing* terlebih dahulu pada metode NBC dan menghasilkan nilai akurasi yang baik. Selanjutnya memasukkan data yang sudah didapatkan di metode LDA yang sudah terdapat topiknya kedalam proses NBC sehingga menghasilkan nilai sentimen pertopik yang dapat dilihat pada Gambar 24.

Gambar 24. Hasil keseluruhan data

IV. KESIMPULAN

Penelitian ini berhasil mengklasifikasikan data *tweet* tentang wabah COVID-19 berdasarkan topik pembahasan dan sentimennya. Pemodelan topik dilakukan menggunakan metode LDA yang dapat menghasilkan topik yang sangat ideal karena memiliki jarak yang cukup jauh antara satu dengan yang lain. Hasil pengujian yang dilakukan memiliki nilai akurasi *training* dan *testing* dengan metode *Naïve Bayes Classifier* (NBC) dengan hasil data *training* 83% dan data *testing* 89% sehingga data tersebut dapat di kombinasikan dengan dua metode. Serta berhasil mengkombinasikan kedua metode *Latent Dirichlet Allocation* (LDA) dengan *Naïve Bayes Classifier* (NBC) sehingga menghasilkan topik yang ideal serta nilai-nilai sentimen pertopik.

DAFTAR PUSTAKA

[1] A. Susilo *et al.*, "Mutasi dan Varian Coronavirus Disease 2019 (COVID-19): Tinjauan Literatur Terkini," *Jurnal Penyakit Dalam Indonesia*, vol. 9, no. 1, pp. 59–81, 2022.

[2] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput*, vol. 8, no. 1, p. 147, 2021.

[3] T. K. Kurniasari, W. Maharani, and J. H. Husen, "Analisis Media Sosial Twitter Untuk Mengetahui Pengguna Berpengaruh Pada Portal Berita Di Indonesia Dengan Metode Tsim

(topic-based Social Influence Measurment)," *eProceedings of Engineering*, vol. 7, no. 3, 2020.

[4] N. C. Siregar, R. R. A. Siregar, and M. Y. D. Sudirman, "Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ)," *Jurnal Teknologia*, vol. 3, no. 1, 2020.

[5] H. A. Prakosa and S. Nasiroh, "Analisis Sentimen dan Pemodelan Topik Untuk Mengidentifikasi Topik Pandemi Covid-19 Pada Media Sosial Twitter menggunakan Naïve Bayes Classifier dan Latent Dirichleat Allocation," *JNANALOKA*, pp. 73–78, 2021.

[6] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Sentiment Analysis on Public Opinion About the Covid-19 Vaccine Using the Naïve Bayes Classifier Algorithm." Jdmsi, 2021.

[7] T. Yulianita, T. W. Utami, and M. Al Haris, "Analisis sentimen dalam penanganan covid-19 di indonesia menggunakan naive bayes classifier," in *Seminar Nasional Variansi*, 2020, pp. 235–243.

[8] K. F. Hakim, P. Silvianti, and A. M. Soleh, "Latent Dirichlet Allocation dalam Identifikasi Respon Masyarakat Indonesia Terhadap Covid-19 Tahun 2020-2021," *Xplore: Journal of Statistics*, vol. 10, no. 3, pp. 249–258, 2021.

[9] M. A. N. Febriansyach, F. Rashif, G. I. P. Nirvana, and N. A. Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Tweet Akun Bot Twitter bertagar# covid-19," *CogITo Smart Journal*, vol. 7, no. 1, pp. 170–181, 2021.

[10] R. K. Ngantung and M. A. I. Pakereng, "Model Pengembangan Sistem Informasi Akademik Berbasis User Centered Design Menerapkan Framework Flask Python," *Jurnal Media Informatika Budidarma*, vol. 5, no. 3, pp. 1052–1062, 2021.