Klasifikasi Sentimen Judul Berita Pemberitaan COVID-19 Tahun 2021 pada Media DetikHealth

Fahri Delfariyadi[1], Afrida Helen[2], Susi Yuliawati[3]

[1] Program Magister Linguistik Fakultas Ilmu Budaya Universitas Padjadjaran
[1]fahri18001@mail.unpad.ac.id

[2] Teknik Informatika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Padjadjaran
[2]helen@pens.ac.id

[3] Departemen Linguistik Fakultas Ilmu Budaya Universitas Padjadjaran
[3]susi.yuliawati@unpad.ac.id

***Abstrak*— Penelitian ini merupakan penelitian linguistik terapan yang mengombinasikan linguistik dan ilmu komputasi dan berfokus di bidang *natural language processing* (NLP). Fenomena yang dikaji adalah klasifikasi sentimen pada judul berita pemberitaan COVID-19 di media DetikHealth selama tahun 2021 sehingga orientasi penelitian ini adalah mengklasifikasikan sentimen pada fenomena tersebut. Pengumpulan data dilaksanakan dengan memanfaatkan fitur saring yang disediakan media tersebut dan analisis data dilakukan dalam dua tahap besar, yaitu *text preprocessing* dan klasifikasi sentimen. Algoritma yang diimplementasikan dalam penelitian ini adalah algoritma MultinomialNB yang merupakan bagian dari *naïve bayes classifier*. Hasil dari penelitian ini adalah diperolehnya tingkat akurasi prediksi sentimen sebesar 72.5%. Selain itu, uji coba dengan tanpa melakukan salah satu atau keseluruhan tahapan *preprocessing data* memberikan dampak terhadap tingkat akurasi mesin. Penurunan tingkat akurasi paling menonjol terlihat pada uji coba tanpa *stemming*. Uji coba tanpa *stemming* menunjukkan adanya pemahaman linguistik yang berbeda jika tahapan *stemming* dilakukan dan penurunan tingkat akurasi mesin. Temuan lain adalah label sentimen netral adalah label sentimen berita dengan prediksi benar tertinggi dan label positif adalah label yang relatif salah diprediksi mesin. Implikasi dari hal ini adalah label positif merupakan label yang berpotensi mengalami kekeliruan prediksi.**

***Kata Kunci*—klasifikasi sentimen, *text preprocessing*, NLP.**

I. PENDAHULUAN

Pada tahun 2020, sebuah virus ditemukan di salah satu kota di Tiongkok, yaitu Kota Wuhan. Virus yang ditemukan adalah *coronavirus* dan mengjangkiti umat manusia. Manusia yang terjangkiti virus ini mengalami persoalan pernasapan yang relatif pelik. Virus ini tidak hanya mengjangkiti manusia di suatu wilayan dan bersifat lokal, melainkan juga meluas hingga ke negara tetangga, seperti Jepang dan Korea Selatan. Kemudian, virus ini juga menyebar ke Indonesia hingga seluruh dunia. Akibatnya adalah virus ini dianggap sebagai pandemi global. Pandemi adalah wabah penyakit yang menyebar hingga ke beberapa wilayah [1].

Pandemi virus korona atau yang lebih dikenal dengan COVID-19 memberikan dampak yang sangat besar terhadap kehidupan manusia. Sekolah ditutup dan kegiatan belajar mengajar diubah menjadi daring adalah salah dua dari sekian banyak perubahan yang timbul di dunia pendidikan. Selain itu, pandemi ini juga memberikan pukulan terhadap kegiatan ekonomi, yaitu kegiatan ekonomi melemah, terjadinya pemutusan kerja, penurunan pendapatan, dan lain-lain.

Pemberitaan pandemi COVID-19 dapat dikaji melalui perpsektif linguistik multidisplin. Pada konteks ini, perspektif yang dimaksud adalah linguistik komputasional. Linguistik komputasional merupakan sebuah ilmu yang merupakan gabungan dari linguistik dan ilmu komputasi dan bertujuan untuk mengomputasikan proses berbahasa, baik berupa perbendaharaan kata maupun teks yang ada di bahasa alami (*natural language*) [2]. Melalui pendekatan ini, linguis dan ahli komputer dapat berkolaborasi kepakarannya sehingga dapat memberikan sumbangsih baru terhadap ilmu pengetahuan. Contoh sumbangsih yang dapat dimanfaatkan adalah fitur *Parts of Speech* (POS) *Tagging*. *POS Tagging* adalah fitur pemberian anotasi pada kelas kata pada ranah morfologi secara otomatis [3]. Implikasi dari fitur ini adalah linguis dapat langsung memberikan label pada data dan menghemat waktu pekerjaan.

Artikel ini mengangkat persoalan pemberitaan COVID-19 di indonesia. Komponen pemberitaan yang diteliti adalah judul berita yang tertera di kepala berita. Judul berita (*headline*) adalah label yang berada di atas atau kepala berita dan memberikan informasi yang relavan terhadap isi berita kepada pembaca [4]. Mengacu pada definisi ini, judul berita berfungsi sebagai perwakilan dari isi berita kepada para pembaca. Aspek yang diteliti dari komponen ini adalah sentimen yang terkandung di dalam judul berita. Analisis sentimen adalah sebuah set algoritma dan teknik yang digunakan untuk mendeteksi apakah suatu teks atau dokumen mengandung sentimen tertentu [5]. Secara umum, sentimen terbagi menjadi tiga, yaitu positif, negatif, dan netral. Sebagai contoh, kalimat *the movie was terrific* dideteksi oleh komputer sebagai kalimat yang mengandung sentimen negatif [6]. Hal ini disebabkan oleh adanya adjektiva *terrific* yang menunjukkan perasaan negatif penutur sehingga ekspresi negatif tersebut terdeteksi oleh komputer sebagai perasaan atau sentimen negatif.

Selanjutnya, peneliti menggunakan media pemberitaan daring DetikHealth sebagai sumber data pada penelitian ini. Pertimbangan pemilihan sumber data didasarkan pada perihal media ini merupakan salah satu media berita daring *mainstream* yang ada di Indonesia dan adanya fitur saring berita di media ini. Untuk membatasi masalah penelitian,

peneliti menetapkan batas penelitian berupa pembatasan interval waktu berita yang diambil. Artinya, berita yang diambil berasal dari kurun waktu tertentu, yaitu sepanjang tahun 2021. Hal ini didasarkan pada data statistik di [7] yang menyatakan bahwa tahun 2021 adalah tahun dengan jumlah kematian nakes paling banyak, yaitu sebanyak 1106 jiwa.

Berdasarkan pemaparan di atas, maka rumusan masalah pada penelitian ini adalah seperti apakah klasifikasi sentimen judul berita pemberitaan COVID-19 tahun 2021 pada media daring DetikHealth dan tujuan penelitian adalah untuk mengklaisifikasikan sentimen judul berita pemberitaan COVID-19 tahun 2021 pada media daring DetikHealth.

II. TINJAUAN PUSTAKA

A. Klasifikasi Sentimen

Analisis sentimen adalah sebuah set algoritma dan Teknik yang digunakan untuk mendeteksi sentimen yang terdapat di suatu teks dan mengklasifikasikannya ke dalam tiga jenis klasifikasi, yaitu positif, negatif, dan netral [5]. Untuk mengklasifikasikan sentimen, diperlukan adanya pemarkah lingual berupa leksikon yang terdapat di dalam teks. Sebagai contoh, leksion *good* 'bagus' merupakan pemarkah sentimen posistif dan leksikon *bad* 'buruk' merupakan pemarkah sentimen negatif [8]. Dari contoh tersebut terlihat dengan jelas bahwa leksikon berfungsi sebagai (1) pemarkah lingual dan (2) pemarkah sentimen di dalam teks.

B. Studi Terdahulu

Melalui penelusuran studi terdahulu yang relevan terhadap penelitian ini, peneliti menemukan tiga riset terdahulu yang beririsan dengan tulisan ini. Pertama, riset oleh [9] yang mengangkat topik pengaruh praproses terhadap sentimen komentar masyarakat pada media social Twitter. Hasil riset ini menunjukkan bahwa kombinasi *cleansing* dan *stemming*, serta normalisasi data menunjukkan kinerja terbaik. Hal ini terlihat dari tingkat akurasi sebesar 77.77%. Selain itu, mutual information memberikan nilai guna berupa seleksi fitur yang akan digunakan sehingga fitur yang dianggap kurang relevan dapat dihilangkan. Kedua, tulisan yang dibuat oleh [10] yang mengangkat isu klasifikasi sentimen para pengguna Twitter terhadap persoalan pengadaan vaksin COVID-19. Tulisan ini memapaparkan bahwa komentar para pengguna Twitter terhadap persoalan tersebut terbagi menjadi tiga klasifikasi, yaitu sentimen positif sebesar 48%, sentimen negatif sebesar 29%, dan sentimen netral sebesar23%. Penelitian ini berimplikasi bahwa kegiataan pengadaan vaksin COVID-19 menunjukkan dampak positif di kalangan para pengguna Twitter. Dan ketiga, penelitian oleh [11] yang mengkaji sentiment kegiatan pembelajaran daring dengan *Naïve Bayes Classifier* yang bersumber dari komentar orang tua peserta didik. Hasil penelitian ini menyatakan bahwa algoritma tersebut dapat melakukan perkiraan sentimen positif dan negatif dengan persentase nilai 62.5% sentimen positif, 37.5% sentimen negatif. Selain itu, penelitian ini juga memperoleh tingkat akurasi sebesar 65%.

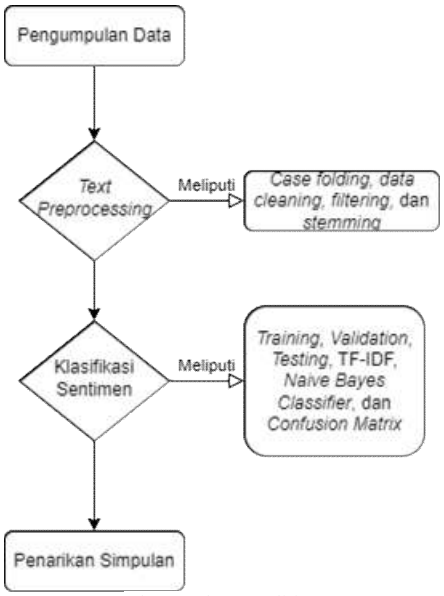
Memerhatikan tiga penelitian terdahulu tersebut, peneliti menemukan rumpang penelitian yang dapat diisi, yaitu belum adanya penelitian yang mengangkat topik klasifikasi sentimen COVID-19 dalam pemberitaan media daring di Indonesia. Atas dasar inilah, topik ini diusung dan diteliti dalam tulisan ini sehingga dapat mengisi rumpang penelitian.

III. METODOLOGI

Penelitian ini merupakan penelitian linguistik terapan karena menggabungkan dua keilmuan, yaitu linguistik dan ilmu komputasi sehingga dapat menghasilkan model komputasi dalam kajian bahasa. Melihat dari perspektif kedua keilmuan ini, penelitian ini merupakan penelitian linguistik komputasional atau dapat pula disebut penelitian *natural language processing* (NLP).

Sumber data penelitian ini berasal dari media daring DetikHealth dengan judul berita sebagai objek penelitiannya. Secara keseluruhan penelitian ini menggunakan metode *supervised machine learning*. Maksudnya adalah metode ini memanfaatkan dan melatih mesin terhadap suatu fenomena melalui pemasukan data. Metode ini berhubungan dengan *classification-based learning*. Pembelajaran berbasis klasifikasi menghasilkan sebuah klasifikasi berdasarkan data yang dipelajari mesin [12].

Untuk mengkaji hal ini, metode penelitian terbagi menjadi dua, yaitu metode pengumpulan data dan analisis data. Pada tahapan pengumpulan data, peneliti menetapkan batasan interval waktu yang telah ditentukan, yaitu sepanjang tahun 2021 batasan topik pemberitaan, yaitu terbatas pada pemberitaan COVID-19. Data diperoleh dengan cara memanfaatkan fitur saring berdasarkan waktu yang telah disediakan di media tersebut. Total data yang terkumpul sebesar 399 data. Setelah data terhimpun, peneliti membuat label sentimen dan memberikan label di tiap data. Lalu, tahapan analisis dilakukan dengan cara mengomputasi data bahasa tersebut di Google Collaborator. Tahapan analisis data dibagi menjadi dua, yaitu *text preprocessing* dan klasifikasi sentimen. Sebelum data memasuki tahapan klasifikasi sentimen, data terlebih dahulu diproses secara komputasi, seperti menghilangkan karakter yang tidak perlu, mengembalikan ke bentuk dasar, dan sebagainya. Setelah melalui tahapan ini, data siap untuk diklasifikasi berdasarkan sentimen yang terkandung. Secara keseluruhan, alur metode pada penelitian ini dapat divisualisasikan melalui Gambar 1 berikut.



Gbr. 1 Alur Penelitian

IV. HASIL DAN PEMBAHASAN

Adapun hasil temuan dari penelitian ini adalah terkumpulnya 399 judul berita mengenai pemberitaan COVID-19 sepanjang tahun 2021 di media daring DetikHealth. Lebih rinci, pembahasan dibagi menjadi dua sebagai berikut.

A. *Text Preprocessing*

Sebelum proses *text preprocessing* dijalankan, seluruh data telah dilabeli dengan tiga label sentimen, yaitu sentimen netral, positif, dan negative dengan total 399 data. Label netral disimbolkan dengan angka nol (0), positif dengan angka satu (1), dan negatif dengan angka dua (2). Selain itu, data berlabel netral berjumlah 147 data, data berlabel positif 114 data, dan data berlabel negatif berjumlah 138 data.

Text preprocessing merupakan tahapan data teks diolah dan disiapkan melalui pemograman komputasi dengan tujuan menghasilkan data yang siap pakai di tahapan analisis yang lebih lanjut. *Text preprocessing* merupakan hal yang krusial untuk dilakukan karena data penelitian ini adalah data teks (*text data*). Data teks adalah data yang berbentuk frasa atau kalimat, dapat ditemukan di berbagai media, dan set data teks disebut dengan istilah korpus [13]. Pada proses ini, *text preprocessing* terbagi menjadi empat bagian, yaitu *case folding*, *data cleaning*, *filtering*, dan *stemming*.

1) Case Folding

Proses pertama dalam *text preprocessing* adalah *case folding*. *Case folding* adalah proses mengubah kalimat yang mulanya ditulis dengan huruf kapital di tiap katanya kemudian diubah menjadi huruf kecil semua [5]. Proses ini dapat berjalan apabila set data telah diunggah ke mesin dalam format *Comma Seperated Value* (CSV) sehingga proses *case folding* ini dapat dijalankan. Cuplikan hasil *case folding* dapat dilihat melalui tabel 1 berikut.

TABEL I
CASE FOLDING

Pra *Case Folding*	**Pasca *Case Folding***
Ada Pasien Omicron Indonesia Sudah Vaksin Booster, Kemenkes Ungkap Kondisinya	ada pasien omicron indonesia sudah vaksin booster, kemenkes ungkap kondisinya

Mengacu pada Tabel I, judul berita yang mulanya diketik dengan huruf kapital di tiap katanya dikonversi menjadi judul dengan huruf kecil di semua katanya. Proses case folding merupakan hal yang esensial untuk dilakukan karena kata yang ditulis dengan huruf kecil dan kata yang ditulis dengan huruf kapital di bagian awal kata dianggap sebagai dua kata yang berbeda walaupun dua kata itu adalah sama. Sebagai caontoh, kata ‘Pasien’ dan ‘pasien’ diidentifikasi oleh mesin sebagai dua kata yang berbeda karena adanya diferensiasi huruf di posisi awal kata.

Meninjau secara linguistis, proses *case folding* memberikan pengaruh pada level morfologis kata. Pengaruh yang diberikan dapat dilihat dari kata-kata yang tergolong akronim. Akronim adalah proses morfologis berupa pemendekan kata sehingga terbentuklah kata yang ditulis dengan huruf awalnya [14]. Sebagai contoh, kata akronim yang terdapat dalam bahasa Indonesia seperti Kamus Besar Bahasa Indonesia (KBBI) dan Kartu Tanda Penduduk (KTP). Akibat dari *case folding* terhadap akronim adalah kata-kata akronim tidak ditulis sesuai dengan kaidah kebahasaan yang berlaku. Oleh sebab itu, terjadi penyimpangan secara kaidah kebahasaan. Untuk mengatasi hal ini, diperlukan pemahaman kebahasaan yang mumpuni guna mendeteksi dan memahami akronim yang mengalami proses *case folding*.

2) Data Cleaning

Setelah proses *case folding* selesai dijalankan, set data masuk ke tahapan yang disebut *data cleaning*. Proses *data cleaning* adalah proses yang melibatkan pembersihan dan penghapusan komponen-komponen yang dianggap tidak penting dan jika tidak dibersihkan dapat memengaruhi proses selanjutnya. *Data cleaning* yang dilakukan pada set data penelitian ini, antara lain yaitu, penghapusan tanda baca dan angka. Simbol numerik atau angka yang terdapat di judul berita pada penelitian ini dihapuskan sehingga judul berita tidak mengandung ekspresi numerik. Selain itu, tanda baca, seperti koma, tanda tanya, dan sebagainya juga turut dibersihkan dari data sehingga tidak ada data yang memiliki tanda baca. Berikut cuplikan data sebelum dan sesudah proses *data cleaning*.

TABEL III
DATA CLEANING

Pra *Data Cleaning*	**Pasca *Data Cleaning***
26 dari 27 kasus omicron ri impor dari negara-negara ini	dari kasus omicron ri impor dari negara-negara ini
vaksin booster gratis untuk lansia dan peserta bpjs, ini yang perlu diketahui	vaksin booster gratis untuk lansia dan peserta bpjs ini yang perlu diketahui

Memerhatikan Tabel II, terlihat bahwa angka 26 dan 27 pada kalimat pertama dibersihkan pada tahapan ini. Akibatnya adalah luaran yang didapatkan adalah tidak ditemukannya ekspresi numerik pada data. Selain itu, tanda baca berupa koma pada kalimat kedua juga dibersihkan pada tahapan ini yang mengakibatkan hasil yang didapatkan adalah tidak ditemukannya tanda baca pada data.

Berbeda dengan *case folding* yang memberikan dampak pada ranah morfologis, *data cleaning* memberikan dampak secara struktur kalimat. Hal ini terlihat dari penghilangan tanda baca dari kalimat. Merujuk pada Tabel II, tanda koma dihilangkan sebagai konsekuensi dari proses ini. Melihat dari fungsinya di dalam kalimat, salah satu fungsi koma adalah mengindikasikan pemisahan antara anak kalimat dan induk kalimat [15]. Konsekuensinya adalah tidak adanya tanda yang memisahkan antara anak kalimat dan induk kalimat dalam struktur sintaksis kalimat. Selain itu, pemisahan antara anak kalimat dan induk kalimat yang dinyatakan oleh tanda koma juga menandakan aspek gramatika dari suatu kalimat. Jika terdapat tanda koma yang memisahkan antara anak kalimat dan induk kalimat, hal ini menyatakan bahwa kalimat tersebut mengikuti kaidah gramatika. Namun, jika anak kalimat diikuti induk kalimat tanpa adanya tanda koma, maka hal ini menunjukkan adanya ketidakpatuhan secara gramatika.

3) Filtering

Set data yang telah melalui proses *data cleaning*, selanjutnya masuk ke proses *filtering*. Hal yang dilakukan peneliti di proses ini adalah menginput fungsi *stopwords* dalam pemrograman. *Stopwords* adalah kata yang mempunyai nilai informasi yang minim di dalam kalimat [16]. Tujuan dari penggunaan *stopwords* adalah membuang kata-kata yang minim informasi, seperti artikel *a*, *an*, *the* dan lain-lain yang mungkin ada di dalam set data [5], [17].

Berdasarkan contoh yang disampaikan oleh [5], [17], fungsi *stopwords* adalah membuang kata-kata yang tergolong sebagai *function words* dalam perspektif linguistik. *Function words* adalah kata-kata yang berfungsi untuk menerangkan fungsi gramatikalnya di dalam kalimat [18]. Mengacu pada contoh artikel dalam bahasa Inggris yang telah disinggung, secara linguistis artikel berfungsi untuk menerangkan atau memberikan informasi yang berkorelasi dengan nomina, seperti kuantitas yang direpresentasikan oleh nomina atau spesifik atau tidaknya informasi yang terkandung di dalam nomina tersebut. Hal ini sejalan dengan pendapat [19] yang menyatakan demikian. Berikut tampilan data sebelum dan sesudah mengalami proses *filtering*.

TABEL IIIII
FILTERING

Pra *Filtering*	**Pasca *Filterng***
dari kasus omicron ri impor dari negaranegara ini	kasus omicron ri impor negaranegara

Mengacu pada data Tabel III, dua preposisi *dari* dihilangkan. Alasan yang melatarbelakangi dua preposisi *dari* dihilangkan adalah keduanya dideteksi sebagai *stopwords* sehingga mesin menghilangkan dua kata tersebut. Hal ini secara eksplisit sejalan dengan pendapat [5] dan [17] yang menjelaskan tentang proses *filtering*. Hal yang dipahami dari proses ini adalah filtering tidak hanya menyaring kata-kata gramatikal, tetapi juga menunjukkan implikasi berupa pengutamaan *content words* dalam data. *Content words* merupakan kata-kata yang terklasifikasi sebagai nomina, verba, adjektiva, adverbia, dan preposisi dalam perspektif linguistik [19]. Menyaring kata-kata yang ada di data sehingga *content words* tetap ada dan *function words* hilang menyebabkan absennya komponen gramatikal dari data. Implikasi dari hal ini adalah keberterimaan kalimat dalam sudut pandang linguistik. Keberterimaan dapat dilihat dari dua sisi, yaitu sisi gramatika dan semantis [20]. Merujuk pada Tabel III, sisi gramatika pada proses ini dapat dilihat dari preposisi *dari* yang hilang. Penghilangan preposisi ini mengakibatkan adanya kekosongan unsur gramatika yang semula ada menjadi tidak ada. Masih berhubungan dangan hal ini, sisi semantis dapat pula dilihat dari kata yang sama, yaitu preposisi *dari*. Penghilangan preposisi *dari* yang pertama menyebabkan hilangnya makna frasa preposisi dan penghilangan *dari* yang kedua menyebabkan hilangnya makna yang menunjukkan direksi dalam kalimat.

4) Stemming

Tahapan terakhir dari *text preprocessing* adalah *stemming*. *Stemming* adalah proses morfologis berupa pengonversian kata derivasi ke bentuk awalnya [3]. Kata-kata yang tergolong bentukan derivasi, seperti kata-kata afiksasi dipisahkan dari imbuhannya sehingga bentuk dasar dari kata tersebut dapat diperoleh. Sebagai contoh, kata dimakan dan berlari dipisahkan dari prefiksnya dan diubah ke bentuk dasarnya, yaitu makan dan lari sebagai konsekuensi dari proses stemming.

Sehubungan dengan penelitian ini yang mengangkat persoalan judul berita berbahasa Indonesia, maka *stemming* dilakukan berdasarkan sistem bahasa Indonesia. Untuk mengubah kata derivasi ke bahasa Indonesia, penulis menggunakan *Sastrawi Stemmer* pada proses ini. *Sastrawi Stemmer* adalah *library* yang menyediakan fitur berupa *stemming* bahasa Indonesia sehingga para peneliti dapat mengonversi kata derivasi ke bentuk awalnya. Perhatikan tabel 4 berikut yang menyajikan cuplikan hasil *stemming*.

TABEL IVV
STEMMING

Pra *Stemming*	**Pasca *Stemming***
ogah divaksin covid pria ini terciduk mau tipu nakes pakai tangan palsu	ogah vaksin covid pria ciduk mau tipu nakes pakai tangan palsu

Merujuk pada tabel, proses *stemming* mengubah morfologis verba, yaitu menghilangkan imbuhan yang melekat padanya. Pada Tabel IV, verba pasif *divaksin* dipisahkan dari imbuhannya sehingga bentuk dasar *vaksin* dapat dicapai pada proses ini. Penghapusan imbuhan dan pengembalian verba ke

bentuk dasar mencerminkan bahwa verba adalah *content words*. Hal ini senada dengan pendapat [21] yang menyatakan bahwa *content words* adalah kata-kata yang mengalami proses derivasi secara morfologis. Proses pengimbuhan (afiksasi) merupakan salah satu proses derivasi yang dapat terjadi di kata-kata pada suatu bahasa tertentu. Selain itu, proses *stemming* menimbulkan implikasi pada ranah semantisnya. Pada kasus ini, implikasi yang ditimbulkan adalah penghilangan makna pasif yang semula dimiliki verba hilang akibat proses ini. Depasifasi verba ini tidak hanya menunjukkan implikasi penghilangan makna pasif, melainkan juga memberikan pengaruh pada relasi makna derivasional. Relasi makna derivasional adalah relasi makna yang memiliki peran dalam proses pembentukan kata [22]. Berdasarkan tabel, relasi antara imbuhan yang melekat dan bentuk dasar adalah pasifasi yang berakibat terbentuknya makna pasif. Namun, proses *stemming* mengembalikan verba ke bentuk dasar yang berakibat pada ketiadaan relasi makna pasif yang semula ada.

B. *Klasifikasi Sentimen*

Selepas proses *text preprocessing* selesai dijalankan, set data masuk ke tahapan klasifikasi sentimen. Untuk menjalankan tahapan klasifikasi sentimen, set data menjalani beberapa proses, yaitu *training and testing data*, *Term Frequency-Inverse Document Frequency* (TF-IDF), *Naïve Bayes Classifier*, dan *Confusion Matrix*.

Pertama, set data menjalani tahapan *training and testing data*. Pada tahapan ini, data yang telah diproses di tahapan sebelumnya dibagi ke dalam dua kelompok, yaitu data yang dilatih (*data training*) dan data yang diuji (*data testing*). Teknik *testing* dibagi, yaitu berdasarkan persentase dan *K-fold cross validation*.

Pada *testing* yang menggunakan teknik K-Fold, hasil akurasi diperoleh sebesar 91.5%. Tingkat Akurasi ini diperoleh dengan melakukan 10 kali *fold* pada dataset. Pada teknik kedua, Set data dengan jumlah 399 *headline* berita dibagi dengan proporsi 80% pada *data training* dan 20% pada *data testing*. Spesifiknya, 319 data dimasukkan sebagai data yang dilatih dan 80 data lainnya dimasukkan sebagai data yang diuji. Lebih rinci, 319 data yang dilatih terdiri dari 118 data dengan label netral, 87 data dengan label positif, dan 114 data berlabel negatif dan 80 data yang diuji terdiri dari 29 data berlabel netral, 27 data berlabel positif, dan 24 data berlabel negatif. Implikasi dari hal ini adalah model yang dikembangkan pada penelitian ini tergolong *supervised learning*.

Kedua, peneliti mengekstraksi vektor TF-IDF ke dalam proses komputasi. [13] menjelaskan bahwa vektor TF-IDF adalah metode yang melakukan pembobotan terhadap kata yang kerap muncul di dokumen tertentu dan bukan di banyak dokumen dari sebuah korpus. TF-IDF sendiri terdiri dari dua komponen, yaitu TF dan IDF. *Term Frequency* (TF) adalah teknik yang mencoba menemukan frekuensi relatif sebuah kata di dalam dokumen tertentu, sedangkan *Inverse Document Frequency* (IDF) adalah teknik yang memastikan kata yang kerap digunakan, seperti kata-kata yang tergolong sebagai komponen gramatikal harus diberikan bobot yang lebih rendah dibandingkan dengan kata lain yang jarang digunakan [16]. Sederhananya, mesin memberikan pembobotan terhadap set data yang telah dibagi sebelumnya, yaitu pembobotan terhadap data yang dilatih dan pembobotan terhadap data yang diuji.

Ketiga, penginputan algoritma *Naïve Bayes Classifier* ke dalam proses komputasi. Algoritma *naïve bayes classifier* adalah algoritma atau model yang berusaha mencari kata kunci dalam suatu set dokumen yang merupakan variabel target dari suatu prediksi [17]. Algoritma yang diimplementasikan pada tahapan ini adalah MultinomialNB yang merupakan salah satu jenis dari algoritma *naïve bayes classifier*. Pengimplementasian algoritma ini didasarkan pada pendapat [13] yang menyatakan bahwa MultinomialNB adalah salah satu algoritma dari *naïve bayes classifier* dan algoritma ini kerap digunakan untuk klasifikasi teks data. Melalui pengimplementasian algoritma ini, mesin dapat melalukan pembelajaran dan pengujian terhadap set data yang telah diberikan. Hasil yang didapatkan dari eksekusi algoritma ini adalah didapatnya tingkat akurasi sebar 72.5%. Persentase ini menunjukkan seberapa akurat mesin memprediksi set data yang telah dilabeli, dilatih, dan diuji. Tingkat akurasi dapat digambarkan lebih rinci melalui tabel berikut yang dinyatakan dalam desimal.

TABEL V
AKURASI K-FOLD

Hasil	Rata-rata (desimal)
Training Accurary	0.915
Training Precision Weighted	0.918
Training Recall Weighted	0.915
Training F1 *Weighted*	0.915
Validation Accuracy	0.566
Validation Precision Weighted	0.608
Validation Recall Weighted	0.566
Validation F1 *Weighted*	0.548

TABEL VI
AKURASI BERDASARKAN PROPORSI

	Precision	***Recall***	***F1-Score***	***Support***
Netral	0.63	0.83	0.72	29
Positif	0.85	0.63	0.72	27
Negatif	0.77	0.71	0.74	24
Accuracy			0.73	80
Macro Average	0.75	0.72	0.73	80
Weighted Average	0.75	0.72	0.73	80

Tingkat akurasi pada Tabel VI adalah tingkat akurasi yang diperoleh melalui pelaksanaan semua tahapan preprocessing data, yaitu *case folding*, *data cleaning*, *filtering*, dan *stemming* dengan melakukan teknik kedua. Pada penelitian ini, peneliti bereksperimen dengan menghilangkan proses tertentu untuk

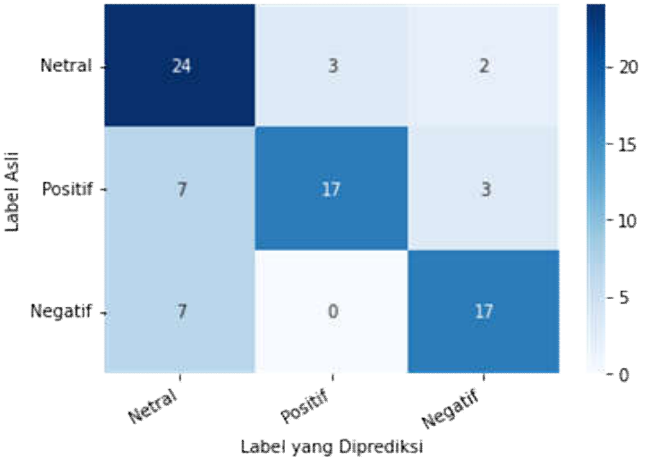
menguji tingkat presisi model pembelajaran. Perhatikan Tabel VI berikut.

TABEL VII
UJI COBA TINGKAT AKURASI

No.	Uji Coba Proses	Tingkat Akurasi
1.	Melaksanakan semua tahapan *preprocessing*	0.725
2.	Tanpa melakukan *case folding*	0.713
3.	Tanpa melakukan *data cleaning*	0.713
4.	Tanpa melakukan *filtering*	0.725
5.	Tanpa melakukan stem*ing*	0.662
6.	Tanpa melakukan *preprocessing data*	0.700

Berdasarkan Tabel VII, terlihat bahwa uji coba dengan menerapkan teknik kedua pada penghilangan proses *case folding* dan *data cleaning* tidak memberikan dampak yang signifikan terhadap tingkat akurasi. Hal ini dilatarbelakangi oleh proses yang dijalankan di kedua tahapan tersebut, yaitu mengubah huruf dan menghilangkan karakter yang tidak relevan. Bahkan, uji coba berupa penghilangan filtering tidak mengubah tingkat akurasi sama sekali. Hal ini mengimplikasikan bahwa kata-kata gramatikal tidak memengaruhi tingkat akurasi. Hal ini juga menunjukkan pertentangan dengan pendapat [5] dan [17] yang menyatakan bahwa *filtering* bertujuan untuk menghilangkan kata-kata gramatikal yang minim informasi. Berbeda dengan tiga proses tersebut, proses *stemming* jika dihilangkan, maka memengaruhi tingkat akurasi mesin dalam memprediksi label. Uji coba penghilangan *stemming* dalam *preprocessing data* memperlihatkan bahwa mesin membutuhkan proses ini dalam memprediksi label karena proses ini mengembalikan kata-kata yang mengalami proses derivasi ke bentuk dasarmya. Hal ini dibuktikan dengan menurunnya tingkat akurasi jika proses *stemming* dihilangkan. Selain itu, pengujian dengan tanpa melakukan semua preprocessing data menunjukkan adanya penurunan tingkat akurasi walaupun tidak signifikan. Mesin tetap dapat memprediksi label walaupun set data tidak menjalani *preprocessing data* sama sekali.

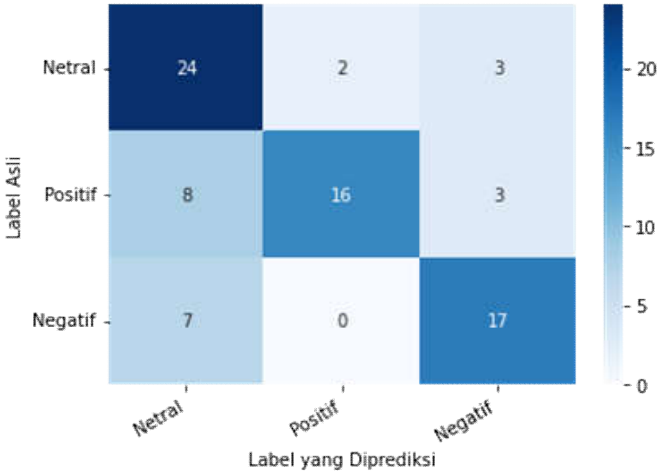
Dan keempat adalah visualisasi melalui *confusion matrix*. *Confusion matrix* adalah matriks yang memperlihatkan hasil evaluasi klasifikasi dan mengkalkulasi jumlah prediksi yang benar oleh mesin [5], [13]. Berdasarkan definisi ini, *confusion matrix* menyediakan fitur berupa visualisasi hasil model pembelajaran yang telah dikerjakan oleh peneliti pada penelitian ini. Adapun hasil model pembelajaran pada penelitian ini dapat dilihat melalui gambar matriks berikut.



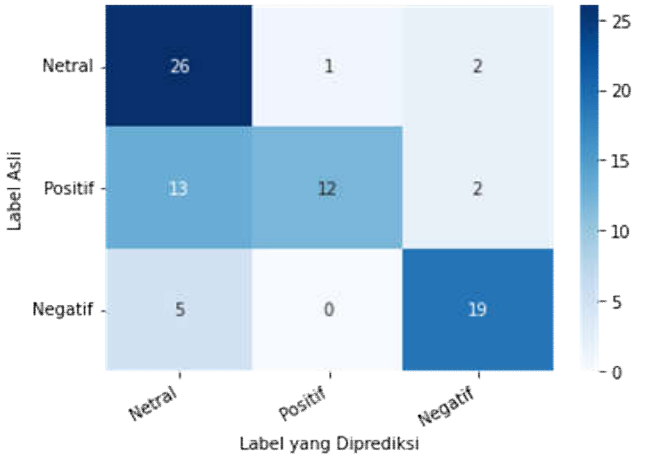
Gbr. 2 *Confusion Matrix* Penelitian

Mengacu pada Gambar 2, hal yang harus diperhatikan adalah label asli pada sumbu vertikal, label yang diprediksi pada sumbu horizontal, dan garis diagonal yang terbentuk dari kiri atas ke kanan bawah. Pada sumbu vertikal, keterangan yang tertera adalah label asli atau dalam hal ini adalah set label yang digunakan oleh peneliti. Pada sumbu horizontal, keterangan yang tertera adalah label yang diprediksi oleh mesin model pembelajaran dan berkenaan. Dan, garis diagonal adalah kesamaan antara hasil yang diekspektasi dan hasil prediksi oleh mesin. Sebagai contoh, perhatikan angka 24 di posisi kiri atas. Angka tersebut menandakan bahwa *data testing* yang telah dilabeli oleh peneliti sebagai sentimen netral diidentifikasi oleh mesin sebagai data dengan sentimen netral. Di posisi kanannya, 3 data sentimen netral diprediksi oleh mesin sebagai data positif dan 2 data lainnya diprediksi sebagai data dengan sentiment negatif. Hal ini mengindikasikan kesesuaian prediksi dengan ekspektasi dan pada saat yang bersamaan juga menunjukkan kekeliruan hasil prediksi mesin.

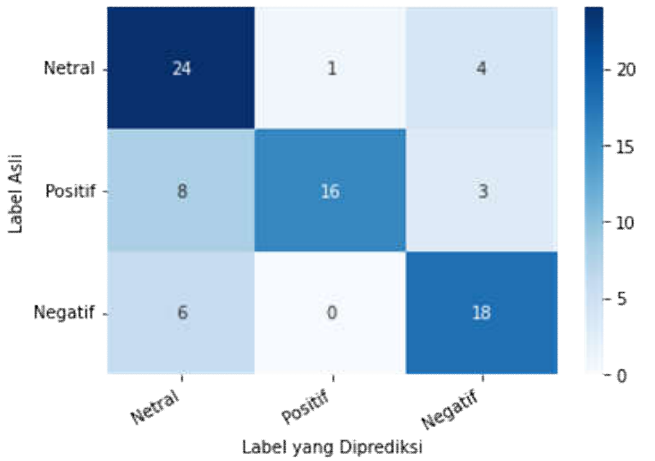
Pada label positif, mesin berhasil memprediksi 17 data dengan sentimen positif dan keliru memprediksi 10 data sentimen positif yang diidentifikasi sebagai 7 data netral dan 3 data negatif. Sama halnya dengan label positif, mesin berhasil memprediksi data negatif dengan besaran data 17 data. Namun, mesin juga keliru dalam mendeteksi 7 data lainnya dan diidentifikasi sebagai data netral, serta tidak ada data yang diidentifikasi sebagai data positif. Hal ini memperlihatkan bahwa kekeliruan mesin dalam mendeteksi sentimen negatif hanya terjadi pada label netral saja.



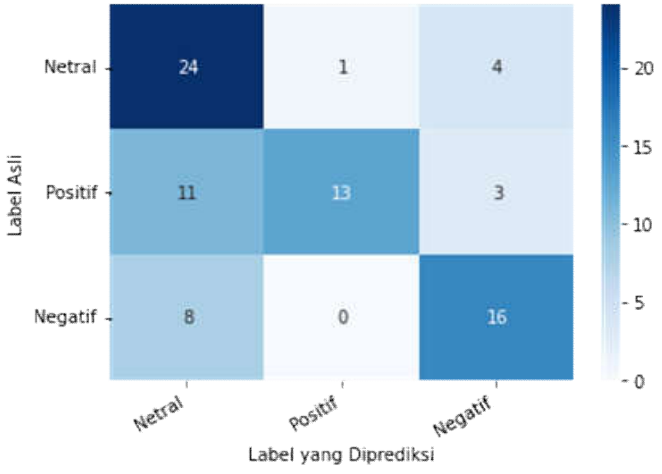
Gbr. 3 *Confusion Matrix* tanpa *Case Folding*



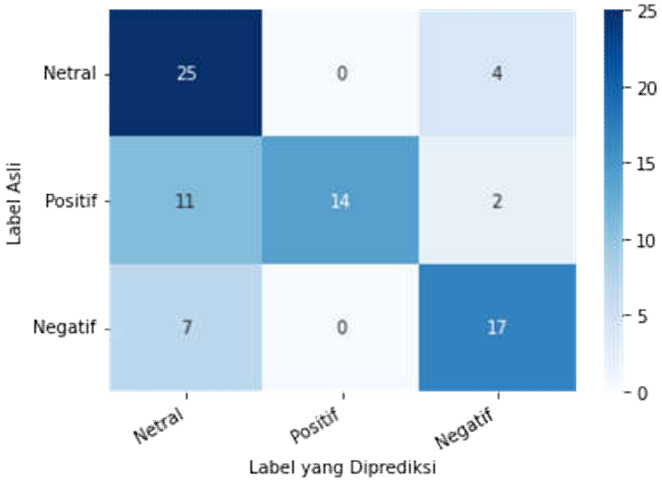
Gbr. 4 Confusion Matrix tanpa Data Cleaning



Gbr. 5 Confusion Matrix tanpa Filtering



Gbr. 6 *Confusion Matrix* tanpa *Stemming*



Gbr. 7 Confusion Matrix tanpa Preprocessing Data

Merujuk pada Gambar 3-7, terdapat diferensiasi hasil prediksi label sebagai konsekuensi dari penghilangan salah satu atau keseluruhan proses di *preprocessing data*. Namun, peneliti menemukan kesamaan yang dapat dipahami, yaitu label netral adalah label dengan hasil prediksi benar tertinggi di setiap matrsiknya. Matriks ini menunjukkan bahwa mesin sudah cukup memadai dalam memprediksi label netral di dalam data. Setelah label netral, disusul prediksi label negatif dan positif.

V. SIMPULAN

Berdasarkan pemaparan hasil dan pembahasan di bagian sebelumnya, maka simpulan dari penelitian ini adalah ditemukan 399 judul berita sepannjang tahun 2021 di media daring DetikHealth. 399 judul berita tersebut terbagi ke dalam tiga proporsi, yaitu judul berita berlabel sentimen netral berjumlah 147 data, label positif 114 data, dan label negatif berjumlah 138 data. Setelah melalui proses *text preprocessing* dan implementasi algoritma *naïve bayes classifier* pada proses klasifikasi sentimen, persentase akurasi yang diperoleh melalui model pembelajaran ini sebesar 72.5%. Tingkat akurasi diperoleh melalui pelatihan dengan memasukkan 319

data sebagai *data training* dan pengujian dengan memasukkan 80 data sebagai *data testing*. Lalu, pengujian berupa penghilangan salah satu atau keseluruhan proses dari preprocessing data dapat memberikan dampak, baik yang tidak signifikan maupun yang relatif signifikan atau bahkan tidak memengaruhi tingkat akurasi sama sekali. Penuruan tingkat akurasi yang relatif signifikan terlihat melalui pengujian tanpa melibatkan *stemming* dalam *preprocessing data*. Implikasinya adalah *stemming* merupakan proses yang relatif berdampak pada *preprocessing data* dan pengukuran tingkat akurasi mesin. Sebab, data yang dikembalikan ke bentuk awal menghasilkan pemahaman linguistik yang berbeda dengan data tanpa dikembalikan ke bentul awal. Selain itu, sentimen netral adalah label dengan prediksi benar tertinggi. Hal ini berimplikasi pada proses pembelajaran yang sudah cukup baik dalam memprediksi data dengan label netra. Namun, kekeliruan masih dapat ditemukan di model pembelajaran ini. Kekeliruan terlihat dari visualisasi yang diperlihatkan pada *confusion matrix*. Kekeliruan paling tampak dalam memprediksi label positif. Melalui penelitian ini terlihat bahwa mesin masih mengalami kesulitan dalam memprediksi label positif walaupun hasil prediksi label positif sudah cukup baik.

Mengingat masih adanya kekeliruan dalam proses prediksi label sentimen, penulis menyarankan pada penelitian selanjutnya dapat memperbanyak kuantitas data, baik total data maupun data di tiap sentimennya. Kuantitas data yang mumpuni dan seimbang di setiap labelnya diharapkan dapat memberikan dampak positif bagi performa model dalam memprediksi sentimen. Tidak hanya itu, peneliti juga menyarankan untuk melakukan riset lebih lanjut mengenai pangaruh dan dampak *preprocessing data* terhadap tingkat akurasi mesin dalam memprediksi label.

REFERENSI

[1] Kemendikbud, "KBBI V." 2016.
[2] I. A. Bolshakov and A. Gelbukh, *Computational Linguistics: Models, Resources, Applications*, 1st Ed. Mexico: Instituto Politecnico Nacional, 2004.
[3] M. Z. Kurdi, *Natural Language Processing and Computational Linguistics 1*, 1st Ed. London: ISTE, 2016.
[4] J. Turow, *Media Today: An Introduction to Mass Communication*, 3rd Ed. New York: Routledge, 2009.
[5] A. Kedia and M. Rasu, *Hands-On Python Natural Language Processing*. Birmingham: Packt, 2020.
[6] B. Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics*. Birmingham: Packt, 2018.
[7] Pusara Digital Tenaga Kesehatan, "Tenaga Kesehatan Indonesia Gugur Melawan COVID-19," *26 Mei*, 2022. https://nakes.laporcovid19.org/statistik
[8] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 1st Ed. New York: Cambridge University Press, 2015.
[9] S. Khairunnisa, Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 2, pp. 406–414, 2021, doi: 10.30865/mib.v5i2.2835.
[10] M. I. Aditama, R. I. Pratama, K. H. U. Wiwaha, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19," *JIEET (Journal Inf. Eng. Educ. Technol.*, vol. 4, no. 2, pp. 90–92, 2020, doi: 10.26740/jieet.v4n2.p90-92.
[11] F. Sidik, I. Suhada, A. H. Anwar, and F. N. Hasan, "Analisis Sentimen Terhadap Pembelajaran Daring dengan Algoritma Naïve Bayes Classifier," *J. Linguist. Komputasional*, vol. 5, no. 1, pp. 34–43, 2022, doi: 10.26418/jlk.v5i1.79.
[12] W. Daelemans and A. Van Den Bosch, "Memory-Based Learning," in *The Handbook of Computational Linguistics and Natural Language Processing*, 1st Ed., A. Clark, C. Fox, and S. Lappin, Eds. West Sussex: Wiley-Blackwell, 2010, pp. 154–179.
[13] A. C. Muller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. California: O'Reilly Media, Inc, 2017.
[14] H. Kridalaksana, *Kamus Linguistik*, 2nd Ed. Jakarta: PT Gramedia Pustaka Utama, 2008.
[15] A. Prihantini, *Master Bahasa Indonesia*, 1st Ed. Yogyakarta: Penerbit B First, 2015.
[16] K. Bhavsar, N. Kumar, and P. Dangeti, *Natural Language Processing with Python Cookbook*, 1st Ed. Birmingham: Packt, 2017.
[17] H. Lane, C. Howard, and H. M. Hapke, *Natural Language Processing in Action*. New York: Manning Publications, 2019.
[18] A. Akmajian, R. A. Demers, A. K. Farmer, and R. M. Harnish, *Linguistics: An Introduction to Language and Communication*, 6th Ed. Massachusetts: The MIT Press, 2010.
[19] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, *Linguistics: An Introduction*, 1st Ed. Cambridge: Cambridge University Press, 2009.
[20] A. Chaer, *Sintaksis Bahasa Indonesia: Pendekatan Proses*. Jakarta: Rineka Cipta, 2020.
[21] V. A. Fromkin, *Linguistics: An Introduction to Linguistic Theory*, 1st ed. Massachusetts: Blackwell Publishing, 2000.
[22] T. F. Djajasudarma, *Semantik 2: Relasi Makna Paradigmatik, Sintagmatik, dan Derivasional*. Bandung: Refika Aditama, 2016.