DECODE: Jurnal Pendidikan Teknologi Informasi
ISSN Cetak : 2775-2984 | ISSN Online : 2775-1813
Vol. 3, No. 2, September 2023, Hal: 161-171, Doi: http://dx.doi.org/10.51454/decode.v3i2.131
http://journal.umkendari.ac.id/index.php/decode

TEKNIK PREPROCESSING PADA TEXT MINING MENGGUNAKAN DATA TWEET "MENTAL HEALTH"

Dianda Rifaldi[1)*], Abdul Fadlil[1)], Herman[1)]
[1] Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: 2207048010@webmail.uad.ac.id

Abstrak

Indonesia merupakan salah satu negara di kategorikan pengguna media sosial twitter terbanyak yaitu mencapai 18,45 pada periode januari tahun 2022 juta pengguna sehingga data pada twitter dapat digunakan dalam melakukan bebagai penelitian. Data penelitian ini menggunakan data media sosial twitter yang diambil dengan metode *crawling* dan mendapatkan data sebanyak 9739 yang diambil dari tanggal 19 oktober 2022 sampai 4 desember 2022 dengan menggunakan *keyword* "*mental health*". Data hasil *crawling* masih berbentuk mentah dan tidak terstruktur, sehingga perlu dilakukan *preprocessing* agar data dapat di proses ke tahap selanjutnya dan menghasilkan data yang dapat diolah menggunakan *tools* pengolah data. Tujuan penelitian ini adalah melakukan *preprocessing* pada data yang sudah diperoleh melalui twitter. Pengolahan data menggunakan model *machine learning* diperlukan tahap persiapan data yaitu dengan melakukan *preprocessing* agar data yang digunakan dapat diolah dengan baik. hasil penelitian ini adalah data yang melewati tahap *preprocessing* telah berbentuk kata dasar dan siap diolah untuk melakukan penelitian terkait *mental health*. Beberapa tahapan yang dilakukan pada *preprocessing* yaitu perubahan bentuk kata dasar, menghapus kata yang tidak penting, menghapus imbuhan, dan konjungsi dari dokumen tweet. Selanjutnya data yang telah melewati tahap *preprocessing* siap untuk dilakukan pembuatan model analisis sentimen yang berguna dalam pengambilan keputusan terhadap permasalahan tersebut.

Kata kunci: mental health; preprocessing; twitter.

PREPROCESSING TECHNIQUES IN TEXT MINING: "MENTAL HEALTH" TWEET DATA

Abstract

Indonesia is one of the countries categorized as the most Twitter social media users, reaching 18.45 million users in the January 2022 period so that data on Twitter can be used in conducting various researches. This research data uses Twitter social media data taken by crawling method and obtained 9739 data taken from October 19, 2022 to December 4, 2022 using the keyword "mental health". Crawled data is still raw and unstructured, so preprocessing is needed so that the data can be processed to the next stage and produce data that can be processed using data processing tools. The purpose of this study is to preprocess the data that has been obtained through twitter. Data processing using machine learning models requires a data preparation stage, namely by preprocessing so that the data used can be processed properly. The result of this study is that data that passes the preprocessing stage has been in the form of basic words and is ready to be processed to conduct research related to mental health. Some of the stages carried out in preprocessing are changing the form of basic words, removing unimportant words, removing affixes, and conjunctions from the tweet document. Furthermore, data that has passed the preprocessing stage is ready to make sentiment analysis models that are useful in making decisions on these problems.

Keywords*: mental health; preprocessing; twitter.*

Submitted: 6 Februari 2023	Reviewed: 7 Februari 2023	Accepted: 6 April 2023	Published: 30 April 2023

Penerbit: Pendidikan Teknologi Informasi FKIP Universitas Muhammadiyah Kendari

PENDAHULUAN

Statistik pengguna twitter di Indonesia mencapai 18.45 juta pada periode Januari 2022 yang dikelompokkan berdasarkan jenis kelamin, laki-laki sebesar 68,5% dan perempuan sebesar 31.5%. Pengguna twitter dalam kategori usia, usia 25-34 tahun sebesar 26.6%, usia 18-24 tahun sebesar 25.2%, usia 45 tahun keatas sebesar 12% dan usia 13-17 tahun sebesar 78% (Admin, 2022)**.**Masalah dalam kesehatan mental merupakan masih menjadi masalah yang besar terutama di Indonesia. Pada tahun 2018 Riset Kesehatan Dasar mengatakan "lebih dari 19 juta penduduk berusia dari 15 tahun mengalami gangguan mental emosional dan lebih dari 12 juta penduduk berusia lebih dari 15 tahun mengalami depresi. Organisasi Kesehatan Dunia (WHO) menyatakan lebih dari 264 juta jiwa dari berbagai kalangan usia menderita depresi, penyakit ini tidak memandang jenis kelamin, status sosial, dan usia (Mulyani et al., 2022). Dalam kasus ini jika masyarakat minim pengetahuan terkait masalah kesehatan mental, maka memungkinkan untuk menambahkan jumlah penderita masalah kesehatan mental.

Analisis sentimen bertujuan untuk mengetahui arah polaritas dari emosi positif, negatif atau netral dari data dokumen, kalimat, paragraf (Zhao, 2015) (Firdaus et al., 2022) (Meetei et al., 2021) (Merinda Lestandy et al., 2021) (Xu et al., 2022) (Leelawat et al., 2022). Proses analisis biasa dilakukan dengan mengumpulkan pendapat dalam bentuk *Text* dari berbagai sumber data dari internet dan berbagai platform media sosial. Analisis dilakukan secara otomatis untuk mengenali apakah segmen *Text* bersifat emosional dan menentukan polaritasnya. Analisis sentimen telah digunakan untuk mendapatkan persentase polaritas emosi (positif atau negatif) terhadap ulasan tokopedia di Google Playstore dengan data yang digunakan sebanyak 992 komentar dengan penggunaan algoritma *Naïve Bayes* sebesar 75.30% dan tingkat akurasi algoritma K-Nearest Neighbor 86.09% (Firdaus et al., 2022).

Kemudian terhadap opini masyarakat terhadap Kominfo dalam pemblokiran situs non-PSE dengan jumlah data 1234 tweet yang telah di *Preprocessing* cenderung negatif sebesar 82.82% positif 10.53% dan netral 6.65% (Rahmawati & Sukmasetya, 2022). Terhadap vaksin *Covid*-19 dengan jumlah data 4.078 data tweet, terdapat 2.525 sentimen positif (43,0%), 771 sentimen negatif (16,4%), dan 1.912 sentimen netral (40,6%). Hasil dari 80% (3766) data latih dan 20% (942) data uji diperoleh skor akurasi sebesar 73,6%. Dari penelitian ini dapat disimpulkan bahwa kecenderungan masyarakat Indonesia pada saat pengambilan data sampling lebih menerima (respon positif) terhadap kebijakan pemerintah terkait program vaksinasi Covid-19 (Syah & Witanti, 2022).

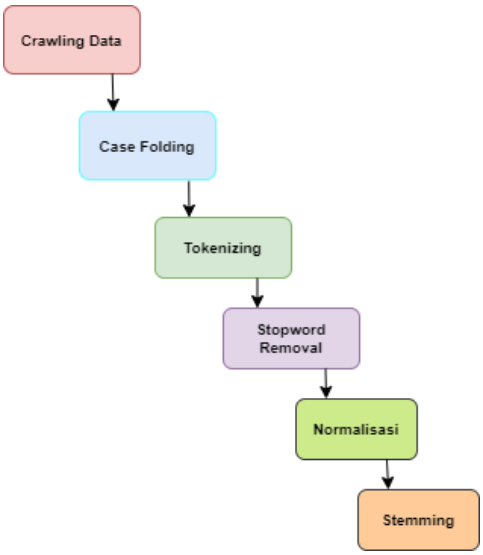
Pada penelitian ini data yang digunakan adalah data tweet pada media sosial Twitter yang diambil dengan menggunakan teknik *Crawling*, data tersebut berhasil diambil namun masih dalam bentuk *Text* utuh dari tweet pengguna. *Keyword* data yang digunakan pada penelitian ini yaitu "*Mental Health*" dengan jumlah data 9739 yang diambil dari 19 Oktober 2022 sampai dengan 04 Desember 2022. yang akan berfokus pada tahap *Preprocessing*. Penelitian membahas tahap *Preprocessing* data tweet sehingga siap digunakan untuk melakukan analisis sentimen, penelitian ini menerapkan teknik *Preprocessing* diantaranya *case folding, tokenizing, stop word removal*, normalisasi dan *stemming* yang nantinya hasil dari *Preprocessing* tersebut dapat digunakan dalam pembuatan model *Machine Learning* Analisis Sentimen. *Preprocessing* merupakan proses pembersihan dan distandarisasi data, karena data mentah yang dihasilkan pengguna biasanya tidak terstruktur dan tidak dapat dianalisis untuk melakukan analisis sentimen (Meetei et al., 2021) (Kurniawan et al., 2020) (Leurs et al., 2022). Teknik *Preprocessing* sering digunakan dalam *Natural Language Processing* untuk menyiapkan *Text* untuk klasifikasi (Sohrabi & Hemmatian, 2019).

Sebelumnya sudah dilakukan penelitian dalam teknik *Preprocessing* pada data sebelum ke tahap pemodelan dalam *Machine Learning* yaitu untuk Analisis Sentimen, penelitian (Vijayarani, Ms. J. Ilamathi, 2015) melakukan *Preprocessing* untuk *Text Mining* untuk

memecahkan berbagai jenis masalah penelitian dalam terkait penambangan data seperti penambangan *Text*, penambangan web, penambangan gambar, penambangan pola sekuensial penambangan spasial, penambangan medis, penambangan multimedia, penambangan struktur dan penambahan grafik. Pada teknik *Preprocessing* dalam penelitian ini menggunakan beberapa tahap diantaranya *Preprocessing, Extraction, Stopword Removal, Stemming, TF/IDF*. (Duong & Nguyen-Thi, 2021) melakukan tahap *Preprocessing* menormalkan data seperti penggantian ikon emoji, penghapusan karakter memanjang, penanganan negasi, penanganan intensifikasi. Kemudian (Sohrabi & Hemmatian, 2019) dengan tujuan untuk mencapai *Text* standar yang dapat diterima dengan menggunakan Metode *Word2vec* yang cepat dan akurat dalam mengubah susunan kata menjadi *Vector Numeric* yang menggunakan tahap *Preprocessing* dengan beberapa tahap yaitu *Conversion of Uppercase Characters to Lowercases, Removing the Accent, Removing Extra Blank Spaces, Removing hyphens, Removing Punction Marks, Removing Watermarks, Removing Watermarks, Removing Stopwords, Tokenizing*. Setelah itu ke tahap mengubah kata ke *numeric* dengan klasifikasi *Supervised Learning* diantaranya *Decision Tree, neural network, Support Vector Machine, Random Forest*, dan melakukan *Evaluation* diantaranya *precision, recall, accuracy*. Tujuan dari penelitian ini adalah untuk melakukan *Preprocessing* pada data yang sudah di *Crawling* sebelumnya dengan *keyword* "*Mental Health*" di media sosial Twitter dengan berfokus pada metode *Case Folding, Tokenizing, Stopword Removal,* Normalisasi, dan *Stemming*. Sebelum masuk kedalam pembuatan model *Machine Learning* diperlukan tahap awal yaitu melakukan *Preprocessing*, data yang sudah melalui proses *Preprocessing* ini nantinya akan dapat digunakan dalam membentuk model *Machine Learning* seperti Analisis Sentimen.

METODE

Tahap *Preprocessing* memiliki peran yang sangat penting dalam teknik dan aplikasi *Text Mining*, ini merupakan langkah pertama yang harus dilakukan dalam proses *Text Mining* (Vijayarani, Ms. J. Ilamathi, 2015). Penelitian ini menggunakan 5 tahapan *Preprocessing* yaitu *Case Folding, Tokenizing, Stopword Removal*, Normalisasi, dan *Stemming*. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahap Preprocessing

Berikut uraian tahapan penelitian yang dilakukan untuk mendapatkan hasil meliputi: (1) *Data Crawling*, merupakan tahap mengambil data pada media sosial salah satunya twitter (Rahmawati & Sukmasetya, 2022) (Nurkholis et al., 2022) (Yudhana et al., 2019). Untuk melakukan *Crawling Data* sebelumnya harus menentukan *keyword* untuk menentukan *topic* tweet apa yang akan diambil dan menentukan tenggat waktu dari tweet yang dilakukan oleh pengguna; (2) *Case Folding* merupakan proses mengubah huruf kapital pada kalimat tweet menjadi huruf kecil (Syafaat Amardita & Dwifebri Purbolaksono, 2022) (Muzaki & Witanti, 2021) (Jannah & Prasetyo, 2022). Untuk *Case Folding* ini tahap yang digunakan mengubah seluruh *Text* tweet yang sudah di *Crawling* sebelumnya menjadi huruf kecil atau *lowercase*. Karena kondisi tweet yang diambil beragam bentuk strutkur *Text* nya dan tidak beraturan; (3) *Tokenizing* merupakan pemisahan kata pada dokumen sehingga membentuk serangkaian token (Syafaat Amardita & Dwifebri Purbolaksono, 2022) (Duong & Nguyen-Thi, 2021) (Ulfah et al., 2022). *Tokenizing* ini adalah proses memisahkan kata perkata dalam setiap kalimat kemudian mendapatkan hasil akhir berupa serangkaian token pada kalimat yang dokumen tweet; (4) *Stopword Removal* merupakan proses menghilangkan kata yang dianggap tidak memiliki makna pada dokumen (Syafaat Amardita & Dwifebri Purbolaksono, 2022) (Ulfah et al., 2022) (Kadhim et al., 2015). Kata-kata yang tidak penting atau dianggap tidak memiliki makna dalam dokumen tweet penelitian ini akan dihilangkan agar dapat mempermudah pemrosesan; (5) Normalisasi, penelitian ini memperbaiki huruf yang ganda pada suatu kalimat serta memperbaiki kalimat yang salah ketik. Tujuannya agar kata bisa dikenali dan dideteksi makna dari kata-kata tersebut. Perbandingan antara tweet yang sudah dilakukan normalisasi dan sebelum dilakukan normalisasi (Rahmawati & Sukmasetya, 2022) (Duong & Nguyen-Thi, 2021) (Saputra, 2019); dan (6) *Stemming*, merupakan proses untuk mendapatkan kata dasar dari kata turunan dengan menghilangkan imbuhan sufiks, infiks, dan konfiks (Syafaat Amardita & Dwifebri Purbolaksono, 2022) (Kadhim et al., 2015). Dengan menggunakan *library sastrawi* ini berfungsi untuk mengubah kata dasar.

HASIL DAN PEMBAHASAN

Data awal yang masih dalam bentuk tidak terstruktur serta masih terdapat kombinasi dari simbol-simbol pada *Text* tidak akan dapat digunakan untuk proses pembentukan model *Machine Learning*. Penelitian ini menggunakan data tweet dari media sosial Twitter dengan *keyword "Mental Health"*. *Keyword yang digunakan hanya 1 yaitu Mental Health* untuk membatasi banyaknya data yang *akan diproses pada penelitian ini. diharapkan penelitian selanjutnya dapat menggunakan model yang dibangun ini dengan keyword yang lain yang bersesuaian*. Data diambil dari tweet pengguna terhitung sejak 19 Oktober 2022 sampai dengan 04 Desember 2022 dengan jumlah yang dihasilkan 9739 *record* data yang memiliki empat atribut diantaranya *Datetime*, Tweet, *Text*, dan Label. Tahap *Preprocessing* tidak menggunakan seluruh atribut tapi hanya menggunakan atribut *Text* yang berisi tweet pengguna Twitter yang akan di Analisis Sentimen. Potongan data dapat dilihat pada Tabel 1.

Tabel 1. Detail Dataset Mental Health

No	Atribut	Tipe Data
1	Datetime	Object
2	Tweet ID	Float
3	Text	Object
4	Label	int

Data yang sudah berhasil di *Crawling* merupakan data mentah yang belum dapat dilakukan proses Analisis Sentimen, oleh karena itu di perlukan tahap *preprocessing* untuk memperoleh data yang terstruktur sehingga dapat dilakukan pembuatan model *Machine Learning*. Penelitian yang dilakukan berhasil memperoleh data yang siap digunakan pada proses pembentukan model *Machine Learning* untuk melakukan Analisis Sentimen. Tahapan awal akan mengubah bentuk *Text* menjadi *lowercase*, melakukan penghapusan *tab, new line, back slice, emoticon, mention, url, hashtag*. Hasil *dataset* dapat dilihat pada Tabel 2.

Tabel 2. Dataset *Mental Health*

Datetime	Text	Label
2022-12-04 18:14:19	@uumiftah Huhu thank you so much for sharing smangat kita smua yg sdg berjuang dg kesehatan mental smga diberi kekuatan! Izin titip konseling gratis	1
2022-12-26 19:47:09	Mental udah hancur, masa iya kesehatan gak dijaga?	-1
2022-12-04 14:53:03	Hahaha aku lemah dalam mempertahakan hubungan, dan lebih mementingkan kesehatan mental. Skip	1
….	….	….
2022-11-27 00:23:13	"Terlalu banyak konsumsi medsos, konsumsi konten-konten viral, asupan menfess toxic, radikal, bikin kesehatan mental menjadi turun drastis	-1
2022-12-04 11:50:29	Bulan pertama berjuang sana sini sendirian, akhirnya nyerah karna bener2 makan kesehatan mental. Bulan kedua mulai terbuka sama temen2, yang Puji Tuhan diahadiahi orang2 yg super baikkkk ðŸ¥²	1

Dataset pada tabel 2 merupakan hasil dari *Crawling* data yang mempunyai 4 atribut yaitu *Datetime, Text, Username* dan *Label*. Data diatas melewati tahap yang dinamakan sebagai *Crawling* yang bertujuan untuk mengambil data dari *platform* media sosial terutama Twitter yang digunakan pada penelitian ini. *Keyword* yang digunakan pada saat pengambilan data di Twitter ini menggunakan "*Mental Health*".

Case Folding, Tahap ini akan mengubah bentuk Text yang awalnya bercampur dengan huruf besar menjadi *Lowercase* dari hasil dataset yang sudah didapatkan melalui tahap *Crawling*. Berikut Gambar 2 hasil penerapan tahap *Case Folding* dalam *Preprocessing* pada data yang sudah didapatkan.

```
Case Folding Result :

                   Datetime      Tweet ID  \
0  2022-12-04 18:14:19+00:00  1.599470e+18
1  2022-12-04 15:04:33+00:00  1.599420e+18
2  2022-12-04 14:53:03+00:00  1.599420e+18
3  2022-12-04 14:30:25+00:00  1.599410e+18
4  2022-12-04 14:09:10+00:00  1.599410e+18

                                        Text        Username  Label
0  @uurmiftah huhu thank you so much for sharing ...  Orbitingtoyeom      1
1  i hope he would take it easy if he starts draw...      rietveldkz      1
2  hahaha aku lemah dalam mempertahankan hubungan...         mhmdbrq      1
3  @nekareanemesis @representatif @rainraingoaw @...     sastrawanto      1
4  ya mungkin emg bnyk yg seolah "dibuat-buat" ta...    kadalkejepit      1
```

Gambar 2. Tahap *Case Folding*

Gambar 2 *merupakan* hasil dari dataset yang sebelumnya sudah dilakukan *Case Folding*, atribut Text tersebut merupakan data tweet dari setiap akun pengguna yang hasil akhirnya berbentuk *type Text lowercase.*

Tokenizing, Setelah melakukan tahap *Case Folding*, proses akan melanjutkan ke tahap *Tokenizing* untuk melakukan pemisahan kata per kata pada dokumen *Text* yang dihasilkan pada saat *Crawling*. Sehingga setelah *Tokenizing* sudah dilakukan maka didapatkan hasil token-token dari kalimat tweet. Berikut Gambar 3 hasil penerapan tahap *Tokenizing* dalam *Preprocessing* pada data yang sudah didapatkan.

```
Tokenizing Result :

0    [huhu, thank, you, so, much, for, sharing, sma...
1    [hope, he, would, take, it, easy, if, he, star...
2    [hahaha, aku, lemah, dalam, mempertahankan, hu...
3    [mereka, lupa, yg, sedang, dilawan, itu, gt, p...
4    [ya, mungkin, emg, bnyk, yg, seolah, dibuatbua...
Name: tweet_tokens, dtype: object
```

Gambar 3. Tahap *Tokenizing*

Stopword Removal, Setelah tahap *Tokenizing* selesai akan melanjutkan tahap *Stopword Removal* yang berfungsi untuk melakukan pemilihan kata-kata yang tidak penting yang ada dalam dokumen, Berikut Gambar 4 hasil penerapan tahap *Stopword Removal* dalam *Preprocessing* pada data yang sudah didapatkan.

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\asus\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

0    [huhu, thank, you, so, much, for, sharing, sma...
1    [hope, he, would, take, it, easy, if, he, star...
2    [hahaha, lemah, mempertahankan, hubungan, meme...
3    [lupa, dilawan, gt, populasi, manusia, bumi, d...
4    [emg, bnyk, dibuatbuat, butuh, program, kyk, g...
Name: tweet_tokens_WSW, dtype: object
```

Gambar 4. Tahap *Stopword Removal*

Gambar 4 merupakan hasil akhir dari pemilihan kata-kata penting yang ada di dokumen, dari data-data tweet yang sudah didapatkan sebelumnya kemudian dipilih kembali kata-kata yang tidak penting kemudian menghapus nya. Jika semuanya sudah, maka jumlah kalimat di setiap dokumen otomatis berkurang jumlahnya karena sudah dilakukan tahap penghapusan kata yang tidak penting. Untuk tahap *Stopword Removal* ini dilakukan dengan cara manual dengan menginputkan bentuk kata singkat yang tidak penting dalam sebuah dokumen tersebut diantaranya "yg", "dg", "rt", "dgn", "ny", "d", 'klo', 'kalo', 'amp', 'biar', 'bikin', 'bilang', 'gak', 'ga', 'krn', 'nya', 'nih', 'sih', 'si', 'tau', 'tdk', 'tuh', 'utk', 'ya', 'jd', 'jgn', 'sdh', 'aja', 'n', 't', 'nyg', 'hehe', 'pen', 'u', 'nan', 'loh', 'rt','&amp', 'yah'.

Normalisasi, Setelah tahap *Stopword Removal* dilakukan selanjutnya akan ke Normalisasi yang berguna untuk untuk memperbaiki huruf yang ganda pada suatu kalimat serta memperbaiki kalimat yang salah ketik. Tujuannya agar kata bisa dikenali dan dideteksi makna dari kata kata tersebut Berikut Gambar 5 hasil penerapan tahap Normalisasi dalam *Preprocessing* pada data yang sudah didapatkan.

```
Out[6]: 0    [huhu, thank, you, so, much, for, sharing, sma...
        1    [hope, he, would, take, it, easy, if, he, star...
        2    [hahaha, lemah, mempertahankan, hubungan, meme...
        3    [lupa, dilawan, gt, populasi, manusia, bumi, d...
        4    [emg, bnyk, dibuatbuat, butuh, program, kyk, g...
        5    [kewarasan, otak, kesehatan, mental, utama, mo...
        6    [jaga, kesehatan, mental, fisik, ngelakuin, ba...
        7    [berjuang, nyerah, karna, bener, makan, keseha...
        8    [kesehatan, mental, diatas, nyari, kerjaan, di...
        9    [gangguan, mental, sadar, dunia, sosmed, buruk...
        Name: tweet_normalized, dtype: object
```

Gambar 5. Tahap Normalisasi

Stemming, merupakan tahap akhir dalam proses *Preprocessing* untuk mendapatkan kata dasar dari kata turunan dengan menghilangkan imbuhan sufiks, infiks, dan konfiks dengan menggunakan *library sastrawi*. *Library* Sastrawi ini merupakan *library* yang di sediakan oleh python untuk dapat membantu mengurangi kata-kata *infleksi* dalam bahasa Indonesia menjadi ke bentuk kata dasarnya. Berikut Gambar 6 hasil penerapan tahap *Stemming* dalam *preprocessing* pada data yang sudah didapatkan.

Gambar 6. Tahap *Stemming*

Gambar 6 merupakan hasil dari proses *Stemming* yang berfungsi untuk mencari kata dasar dari setiap dokumen tweet yang sudah di *Crawling* sebelumnya, kata dasar tersebut diambil secara keseluruhan data yang ada dari atribut *Text*.

Setelah melalui semua tahap *Preprocessing* dari data yang di *Crawling* terdapat perbedaan jumlah kata yang ada pada dokumen tweet tersebut. Untuk melihat jumlah kata pada dokumen tweet dapat dilihat pada Tabel 3.

Tabel 3. Kata Setelah Tahap *Preprocessing*

No	Text	Jumlah Kata				
		Case Folding	*Tokenizing*	*Stopword removal*	*Normalisasi*	*Stemming*
1	Huhu thank you so much for sharing smangat kita smua yg sdg berjuang dg kesehatan mental smga diberi kekuatan! Izin titip konsleing https://t.co/KKhOd7KZh	23	23	19	19	19
2	I hope he would take it easy if he starts drawing again. Vagabond sendiri dulu indefinite hiatus karena kesehatan mangkanya terganggu/ I heard if affect his mental health	26	26	23	23	23

3	Hahaha aku lemah dalam mempertahankan hubungan dan lebih mementingkan kesehatan mental. Skip	12	12	8	8	8
4	@represetatif @rainraingoaw @BayuSmdro mereka lupa yg sedang dilawan itu >95% populasi manusia di bumi… dan dikatain ‘phobic’ pula tanpa alasan. Main di kebingungan linguistic, makanya pada obsesi tuh ama “pronoun” dan jerit2 identitas demi kesehatan mental dll dst… hash.. jadi curcol dah. Wes ah bobo	43	43	27	27	27
5	Ya mungkin emng bnyk yg seoalh “dibuat-buat” tapi banyak juga yg butuh program kyk gini, artinya udh banyak aware sm kesehatan mental. Ntar ujung2nya “jaman kami diginiin biasa aja, blaba” yg bagus, tapi skrng jaman udh berubah dan sbnrnya bnyk yg butuh ini jg dari jaman ke jaman	48	48	28	28	28
….	….	….	….			
100	Live instagram #YEScurhat “APAKAH KAMU PEOPLE PLEASER?” bersama dr.Ekanita M.S.,Sp.KJ 😍 https://t.co/m0HOLABtMN #mental #mentalhealth #mentalhealth tips #kesehatanmental	10	10	7	7	7

Tabel 3 menunjukan hasil dari perbandingan jumlah kalimat sebelum dan sesudah dilakukan *Preprocessing*, yang mana terdapat perubahan jumlah kata dan jumlah ukuran pada file yang mana ukuran file sebelum dilakukan *Preprocessing* sebesar 2,118 KB dan setelah dilakukan tahap *Preprocessing* sebesar 8,182 KB. Fungsi dari membandingkan disini adalah untuk melihat perbedaan jumlah dari kata yang sebelumnya belum dilakukan *Preprocessing* dan setelah dilakukannya *Preprocessing.*

Kemudian, disini mencoba untuk membandingkan hasil dari labeling otomatis dengan *textlob* dan manual. Hasil manual dalam pelabelan pada data tweet dari pengguna Twitter di dapatkan berjumlah 100% positif dan 0 negatif dengan menggunakan *textlob* dan 75% positif dan 25% negatif dengan menggunakan manual, untuk keterangan dari labeling dibawah adalah angka 1 merupakan positif sedangkan 0 negatif. Detail dapat dilihat pada Tabel 4.

Tabel 4. Perbandingan Labeling

No	Text	Label	
		Textlob	Manual
1	Huhu thank you so much for sharing smangat kita smua yg sdg berjuang dg kesehatan mental smga diberi kekuatan! Izin titip konsleing https://t.co/KKhOd7KZh	1	1
2	I hope he would take it easy if he starts drawing again. Vagabond sendiri dulu indefinite hiatus karena kesehatan mangkanya terganggu/ I heard if affect his mental health	1	1
3	Hahaha aku lemah dalam mempertahankan hubungan dan lebih mementingkan kesehatan mental. Skip	1	1
4	@represetatif @rainraingoaw @BayuSmdro mereka lupa yg sedang dilawan itu >95% populasi manusia di bumi… dan dikatain 'phobic' pula tanpa alasan. Main di kebingungan linguistic, makanya pada obsesi tuh ama "pronoun" dan jerit2 identitas demi kesehatan mental dll dst… hash.. jadi curcol dah. Wes ah bobo	1	0
5	Ya mungkin emng bnyk yg seoalh "dibuat-buat" tapi banyak juga yg butuh program kyk gini, artinya udh banyak aware sm kesehatan mental. Ntar ujung2nya "jaman kami diginiin biasa aja, blaba" yg bagus, tapi skrng jaman udh berubah dan sbnrnya bnyk yg butuh ini jg dari jaman ke jaman	1	1
….	….	….	….
100	Live instagram #YEScurhat "APAKAH KAMU PEOPLE PLEASER?" bersama dr.Ekanita M.S.,Sp.KJ😍 https://t.co/m0HOLABtMN #mental #mentalhealth #mentalhealth tips #kesehatanmental	1	1

Tabel 4 menunjukan hasil pelabelan yang dilakukan perlabelan secara manual dan menggunakan *library textlob,* hasil akurasi yang didapatkan pada perbandingan perlabelan dengan perhitungan 75/100 * 100% dan 25/100 * 100% yaitu dengan hasil yang akhir 75% benar dan 25% salah.

KESIMPULAN DAN SARAN

Penelitian berhasil melakukan *Preprocessing* untuk data tweet yang di *Crawling*, data ini sudah dapat digunakan untuk ke tahap selanjutnya hingga mencapai tahap akhir untuk melakukan Analisis Sentimen. Dari 4 atribut yang sudah didapatkan dan 9739 *record* data yang dihasilkan untuk *Preprocessing* hanya menggunakan atribut *Text*. tahapan yang sudah dilalui dalam *Preprocessing* ini adalah *Case Folding, Tokenizing, Stopword Removal*, Normalisasi, dan *Stemming*. Hasil penelitian adalah data yang sudah melewati *Preprocessing* sudah dapat langsung digunakan untuk melakukan Analisis Sentimen terhadap *Mental Health*. Dari penelitian ini mendapatkan hasil dalam melakukan *Preprocessing* dalam membuang kata-kata yang tidak penting, simbol-simbol dan memisahkan kata per kata yang ada di dalam dokumen *text* menjadi token.

Saran untuk penelitian selanjutnya dapat ditambahkan pada tahap pemilihan kata-kata tidak penting yang ada di dokumen dan melanjutkan pembuatan model *Machine Learning* untuk analisis sentimen pada studi kasus *Mental Health*.

DAFTAR PUSTAKA

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), 1-16.

El Firdaus, M. F., Nurfaizah, N., & Sarmini, S. (2022). Analisis Sentimen Tokopedia Pada Ulasan di Google Playstore Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor. *JURIKOM (Jurnal Riset Komputer)*, *9*(5), 1329-1336. http://dx.doi.org/10.30865/jurikom.v9i5.4774

Jannah, Y. A. N., & Prasetyo, R. B. (2022). Analisis Sentimen dan Emosi Publik pada Awal Pandemi COVID-19 Berdasarkan Data Twitter dengan Pendekatan Berbasis Leksikon. *Seminar Nasional Official Statistics*, *2022*(1), 597-608.

Jianqiang, Z. (2015, December). Pre-processing boosting Twitter sentiment analysis?. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 748-753.

Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2015). Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering. *Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014*, 69-73. https://doi.org/10.1109/ICAIET.2014.21

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Kurniawan, S., Gata, W., Puspitawati, D. A., Parthama, I. K. S., Setiawan, H., & Hartini, S. (2020). Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis. *IOP Conference Series: Materials Science and Engineering*, *835*(1), 1-8. https://doi.org/10.1088/1757-899X/835/1/012057

Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., Yudha, A. K., & Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon*, *8*(10), e10894. https://doi.org/10.1016/j.heliyon.2022.e10894

Leurs, W. L. M., Lammers, L. A. S., Compagner, W. N., Groeneveld, M., Korsten, E. H. H. M., & van der Linden, C. M. J. (2022). Text mining in nursing notes for text characteristics associated with in-hospital falls in older adults: A case-control pilot study. *Aging and Health Research*, *2*(2), 100078. https://doi.org/10.1016/j.ahr.2022.100078

Meetei, L. S., Singh, T. D., Borgohain, S. K., & Bandyopadhyay, S. (2021). Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, *55*(4), 947-969. https://doi.org/10.1007/s10579-021-09541-9

Merinda Lestandy, Abdurrahim Abdurrahim, & Lailis Syafa'ah. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 802-808. https://doi.org/10.29207/resti.v5i4.3308

Mulyani, S., & Novita, R. (2022). Implementation Of The Naive Bayes Classifier Algorithm For Classification Of Community Sentiment About Depression On Youtube. *Jurnal Teknik Informatika (Jutif)*, *3*(5), 1355-1361.

https://doi.org/10.20884/1.jutif.2022.3.5.374

Muzaki, A., & Witanti, A. (2021). Sentiment Analysis of the Community in the Twitter To the 2020 Election in Pandemic Covid-19 By Method Naive Bayes Classifier. *Jurnal Teknik Informatika (Jutif)*, *2*(2), 101-107. https://doi.org/10.20884/1.jutif.2021.2.2.51

Nurkholis, A., Alita, D., & Munandar, A. (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(2), 227-233. https://doi.org/10.29207/resti.v6i2.3906

Rahmawati, C., & Sukmasetya, P. (2022). *Sentimen Analisis Opini Masyarakat Terhadap Kebijakan Kominfo atas Pemblokiran Situs non-PSE pada Media Sosial Twitter*. *9*(5), 1393-1400. https://doi.org/10.30865/jurikom.v9i5.4950

Saputra, N. (2019). Sentiment Analisys With Lexicon Preprocessing. *Dinamika Informatika*, *7*(1), 45-57.

Sohrabi, M. K., & Hemmatian, F. (2019). An Efficient Preprocessing Method For Supervised Sentiment Analysis By Converting Sentences To Numerical Vectors: A Twitter Case Study. *Multimedia Tools and Applications*, *78*(17), 24863-24882. https://doi.org/10.1007/s11042-019-7586-4

Syafaat Amardita, R., & Dwifebri Purbolaksono, M. (2022). Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN. *Jurnal Riset Komputer)*, *9*(1), 2407-389. https://doi.org/10.30865/jurikom.v9i1.3793

Syah, H., & Witanti, A. (2022). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM). *Jurnal Sistem Informasi Dan Informatika (Simika)*, *5*(1), 59-67. https://doi.org/10.47080/simika.v5i1.1411

Ulfah, A. N., Anam, M. K., Sidratul Munti, N. Y., Yaakub, S., & Firdaus, M. B. (2022). Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19. *JUITA : Jurnal Informatika*, *10*(2), 209. https://doi.org/10.30595/juita.v10i2.12308

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*(April), 100073. https://doi.org/10.1016/j.dajour.2022.100073

Yudhana, A., Fadlil, A., & Rosidin, M. (2019). Indonesian Words Error Detection System Using Nazief Adriani Stemmer Algorithm. *International Journal of Advanced Computer Science and Applications*, *10*(12), 219-225. https://doi.org/10.14569/ijacsa.2019.0101231

How to cite:

Rifaldi, D., Fadlil, A., & Herman, H. (2023). Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet "Mental Health". *DECODE: Jurnal Pendidikan Teknologi Informasi, 3(2),* 161-171. http://dx.doi.org/10.51454/decode.v3i2.131