PENERAPAN METODE *NAÏVE BAYES CLASSIFIER* DAN *LEXICON BASED* UNTUK ANALISIS SENTIMEN *CYBERBULLYING* PADA BPJS

Madonna Al Khadafi, Kurnia Paranita Kartika, Filda Febrinita
Program Studi Teknik Informatika S1, Fakultas Teknologi Informasi
Universitas Islam Balitar, Jalan Majapahit No. 2-4, Kota Blitar, Indonesia
madonnakhadafi53@gmail.com

ABSTRAK

Presiden Joko Widodo menerbitkan Instruksi Presiden Republik Indonesia Nomor 1 Tahun 2022 tentang Optimalisasi Pelaksanaan Program Jaminan Kesehatan Nasional. Inpres tersebut mengatur syarat mengurus sejumlah layanan publik seperti jual beli tanah, membuat SIM, SKCK, haji dan umrah harus terdaftar sebagai peserta BPJS Kesehatan. Peraturan tersebut dimulai per 1 Maret 2022. Tetapi, yang menjadi persoalan adalah ketika berpendapat tidak berlandaskan etika, sehingga muncul adanya cemooh atau menyudutkan pihak yang bersangkutan, maka bisa mengakibatkan adanya tindak *cyberbullying*. Untuk itu, perlu dilakukan menganalisis sentimen pada komentar Twitter untuk mengklasifikasikan *tweet* yang mengandung *cyberbullying* atau *non cyberbullying*. Metode yang digunakan dalam penelitian ini adalah deskriptif kualitatif dengan algoritma pengolahan data *cyberbullying* menggunakan *Naive Bayes Classifier* dan *Lexicon Based*. Proses dimulai dari pengambilan data *tweet, pre-processing,* pembobotan TF-IDF, performa *Naive Bayes Classifier*. Kemudian dilakukan proses klasifikasi menggunakan metode *Lexicon Based* dengan hasil keluaran sistem berupa identifkasi apakah *tweet* termasuk *cyberbullying* atau *non cyberbullying*. Pada penelitian ini, didapatkan hasil performansi dari *Naive Bayes Classifier* menghasilkan akurasi 80%, sedangkan *Lexicon Based* menghasilkan akurasi 22%. Setelah membandingkan kedua metode, maka kesimpulan yang dapat diambil adalah *Naive Bayes Classifier* lebih baik dan lebih akurat daripada *Lexicon Based*.

Kata kunci : *Naive Bayes Classifier*, *Lexicon Based*, Analisis Sentimen, *Cyberbullying*, BPJS Kesehatan

1. PENDAHULUAN

Presiden Joko Widodo, telah menerbitkan Instruksi Presiden Republik Indonesia (Inpres) Nomor 1 Tahun 2022 tentang Optimalisasi Pelaksanaan Program Jaminan Kesehatan Nasional (JKN) (Liputan6, 2022). Inpres tersebut antara lain mengatur syarat mengurus sejumlah layanan publik seperti jual beli tanah, membuat SIM, STNK, SKCK, haji dan umrah yang harus terdaftar sebagai peserta BPJS Kesehatan. Tentunya peraturan itu menunai polemik lantaran mewajibkan kartu peserta BPJS Kesehatan sebagai syarat jual beli tanah. Peraturan tersebut dimulai pada tanggal 1 Maret 2022, Kementerian Agraria dan Tata Ruang atau Badan Pertanahan Nasional (ATR atau BPN) telah merilis aturan baru, yakni jual beli tanah wajib menyertakan kartu BPJS Kesehatan.

Cyberbullying (perundung dunia maya) ialah *bullying* atau perundungan dengan menggunakan teknologi digital. Dalam hukum Indonesia, *cyberbullying* sudah memiliki kebijakan yang diatur dalam peraturan perundang-undangan yakni, Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik (UU *ITE*) dan memiliki sanksi bagi pelaku yang melakukan *cyberbullying* [1]. Dalam kasus yang sudah dijelaskan diatas, bahwa pelaku *cyberbullying* atau komunikator bisa berinteraksi melalui pesan personal, komentar, bahkan status. Disini peniliti lebih memilih media sosial Twitter dibandingkan dengan media sosial lainnya, karena permasalahan yang ada pada Twitter lebih akurat dan mengerucut. Terlebih lagi Twitter merupakan salah satu media sosial dan layanan yang bekerja secara *real-time*, memungkinkan pengguna untuk mengekspresikan opini atau perasaan pengguna mengenai banyaknya isu serta permasalahan dalam bentuk pesan.

Twitter adalah situs berita dan jejaring sosial *online* tempat orang berkomunikasi dalam pesan singkat yang disebut *tweet*. *Tweet*ing memposting pesan singkat untuk siapa saja yang mengikuti anda di Twitter, dengan harapan kata-kata anda bermanfaat dan menarik bagi seseorang di audiens [2]. Twitter sendiri memiliki beberapa fitur, seperti *tweet* atau kicauan, *following*, *followers*, *biography*, *profile*. Pada Twitter juga terdapat fitur top trending yang digunakan untuk mempermudah penggunanya untuk melihat *tweet* yang paling popouler dan paling sering di *tweet* oleh para pengguna Twitter lainnya [3].

Naïve Bayes Classifier adalah metode klasifikasi dalam penambangan teks yang digunakan dalam analisis sentimen. Metode ini berpotensi baik dalam klasifikasi dalam hal presisi dan komputasi data [4]

Lexicon Based adalah suatu pendekatan yang meliputi frase, bentuk ekspresi, atau konten yang berupa teks yang umumnya terdapat pada obrolan, dialog, post, review, dan lainnya. *Lexicon Based* merupakan pendekatan yang menggunakan suatu kamus sentimen berisi kata positif dan kata negatif yang dicocokkan dengan kata pada kalimat untuk diketahui tingkat polaritasnya.

Salah satu cara untuk melakukan analisa pada media sosial dapat dilakukan dengan menggunakan metode dalam data *mining*, yaitu *Naive Bayes Classifier* dan *Lexicon Based*. Keterkaitan antara Twitter dengan metode *Naive Bayes Classifier* adalah untuk klasifikasi, sedangkan dengan metode *Lexicon Based* hanya digunakan untuk pembanding saja, karena *Lexicon Based* didasarkan oleh orientasi kontekstual pada jumlah orientasi sentimen untuk setiap kata ataupun kalimat. Jika dilihat dari kompleksitas, metode *Naive Bayes Classifier* ini cenderung lebih sederhana dan konvensional daripada metode-metode yang lainnya.

Diharapkan penelitian ini, dapat mencapai tujuan penelitian dalam pengklasifikasian *tweets* positif dan negatif tentang jual beli tanah harus menyertakan surat BPJS Kesehatan itu termasuk ke dalam *cyberbullying* atau tidak dengan membandingkan metode *Naive Bayes Classifier* dan *Lexicon Based*. Manfaat yang dapat diambil dari penelitian ini diharapkan dapat memberikan informasi yang berguna bagi pihak instansi BPJS sebagai bahan evaluasi program.

2. TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen merupakan salah satu bidang dari *Natural Languange Processing* (NLP) yang membangun sistem untuk mengenali dan mengekstraksi opini ke dalam bentuk teks. Menurut [5] analisis sentimen atau *opinion mining* adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, sikap dan emosi orang-orang terhadap suatu objek seperti produk, layanan, organisasi, individu, isu, peristiwa, topik dan lainnya.

2.2. Twitter

Twitter merupakan salah satu media sosial dan layanan *microblogging* yang kerjanya secara *real time*, serta memungkinkan bahwa pengguna bisa mengekspresikan opini atau perasaan pengguna mengenai isu yang ada pada saat itu. Twitter sendiri adalah salah satu media sosial yang memiliki pengguna kurang lebih perharinya mencapai 126 juta. Selain itu, Twitter memiliki Twitter Trending Topic yang merupakan daftar topik terpopuler dan di*update* setiap waktu [6].

2.3. *Naïve Bayes Classifier*

Naive Bayes Calssifier adalah metode klasifikasi dalam penambangan teks yang digunakan dalam analisis sentimen. Metode ini berpotensi baik dalam klasifikasi dalam hal presisi dan komputasi data [4] *Naive Bayes Calssifier* menghitung probabilitas dengan ketentuan bahwa class keputusan adalah benar karena vektor informasi objek, mengasumsikan bahwa atribut objek ini independent [7]. Berikut ini rumus *Naïve Bayes Classifier*:

$$\boldsymbol{P(C|X) = \frac{P(X|C)P(c)}{P(x)}} \quad (1)$$

Keterangan:

X : vecktor input

c : sebuah *class* spesifik

P(C|X) : probabilitas *class* berdasar vektor input yang diketahui (*posteriori probability*)

P(c) : probabilitas *class* yang dicari (*prior probability*) dari keseluruhan data

P(x|c) : probabilitas tiap input berdasarkan kondisi pada *class*

P(x) : probabilitas suatu input dari keseluruhan data

2.4. *Lexicon Based*

Lexicon Based merupakan suatu pendekatan yang meliputi frase, bentuk ekspresi, atau konten yang berupa teks yang umumnya terdapat pada obrolan, dialog, post, review, dan lainnya. *Lexicon Based* merupakan pendekatan yang menggunakan suatu kamus sentimen berisi kata positif dan kata negatif yang dicocokkan dengan kata pada kalimat untuk diketahui tingkat polaritasnya. Klasifikasi sentimen dengan *Lexical Based* adalah klasifikasi berdasarkan kata positif dan kata negatif yang ada pada *tweets* yang telah dibersihkan.

2.5. *Text Mining*

Text mining memiliki definisi menambang data yang berupa teks di mana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen. *Text mining* dapat memberikan solusi dari permasalahan seperti pemprosesan, pengorganisasian atau pengelompokkan dalam jumlah besar [8]. Dalam *text mining*, terdapat tahap *pre-processing*. *Pre-processing* merupakan suatu proses pengumpulan data mentah untuk diolah menjadi data yang bermanfaat [9]. Tahapan dalam *pre-processing* ada 7 yaitu :

a. *Labelling*

Labelling dilakukan dengan memberikan label pada setiap komentar atau *tweets* setiap pengguna aplikasi Twitter. Untuk pelabelan dibagi menjadi dua bagian yakni, positif dan negatif. Proses pelabelannya yaitu dari kalimat atau kata-kata yang dimasukkan.

b. *Case Folding*

Case folding yaitu merubah semua karakter huruf pada sebuah kalimat menjadi huruf kecil dan menghilangkan karakter yang dianggap tidak valid seperti angka, tanda baca, dan *Uniform Resources Locator* (URL) [10].

c. *Cleaning*

Proses *cleaning* merupakan proses penghilangan komponen tertentu yang terdapat dalam data yakni seperti *Uniform Resource Locator* (URL), *username*, *hashtag*.

d. *Tokenizing*

Tokenizing yaitu memotong sebuah kalimat berdasarkan tiap kata yang menyusunnya [10].

e. *Stopword Removal*

Stopword removal merupakan tahap dimana kata-kata yang tidak penting atau tidak menggambarkan isi dari sebuah komentar akan dihilangkan atau dipotong.

f. *Stemming*

Stemming yaitu merubah berbagai kata berimbuhan menjadi kata dasarnya, tahap ini pada umumnya dilakukan untuk teks dengan bahasa Inggris, karena teks dengan bahasa Inggris memiliki struktur imbuhan yang tetap [10].

g. *Weighting*

Algoritma yang digunakan setelah proses tahapan *text mining* yaitu *Term-Frequency Inverse Document Frequency* (TF-IDF). Suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen [11]. Berikut ini adalah rumus dari TF-IDF.

$$IDFt = \log\left(\frac{D}{df}\right) \quad (2)$$

$$Wd.t = tfd.t * IDFt \quad (3)$$

Keterangan (2):

IDF = inversed document frequency
t = kata ke-*t* dari kata kunci
D = total dokumen
df = banyak dokumen yang mengandung kata yang dicari

Keterangan (3):

Pada persamaan (2) digunakan untuk mencari nilai IDF
d = dokumen ke-*d*
t = kata ke-*t* dari kata kunci
W = bobot dokumen ke-*d* terhadap kata ke-*t*
tf = banyak kata yang dicari pada sebuah dokumen

2.6. *Crawling*

Crawling adalah semacam pengambilan data dari media sosial kemudian di kumpulkan menjadi satu untuk di evakuasi dan di bentuk agar menjadi sebuah penelitian. *Crawling* data merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu *database*. Pengumpulan data dari penelitian ini yaitu data yang diunduh dari *server* Twitter berupa *user* dan *tweet* beserta atribut-atributnya [12].

2.7. *Confusion Matrix*

Confusion matrix adalah tabel dengan empat kombinasi berbeda dari nilai prediksi dan nilai aktual. *Confusion matrix* adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [13]. Pada kelas 1 merupakan positif dan kelas 0 merupakan negatif.

Tabel 1. *Confusion Matrix*

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:
TP (*True Positive*) = jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
TN (*True Negative*) = jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
FP (*False Positive*) = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
FN (*False Negative*) = jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0

Rumus *confusion matrix* untuk menghitung *accurazy*, *precision*, *recall* dan F1-*score* sebagai berikut.

$$accuracy = \frac{TP+TN}{Total} \quad (4)$$

$$precision = \frac{TP}{TP+FP} \quad (5)$$

$$recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ score = (2 * \frac{recall*precision}{recall+precision}) \quad (7)$$

2.8. *Cyberbullying*

Media sosial tentu saja memiliki dampak yang postif dan negatif, begitu juga dengan pengguna media sosial yang semakin hari semakin tinggi juga diikuti dengan lahirnya kebebasan berpendapat. *Cyberbullyig* dikenal sebagai bentuk ancaman atau serangan yang dilakukan seseorang terhadap orang lain yang disampaikan melalui pesan elektronik lewat media [9]. *Cyberbullying* dapat menyebabkan korban yang di intimidasi mengalami gangguan secara emosional. Pelaku *cyberbullying* menganggap bahwa melakukan perundungan dalam dunia maya lebih mudah dilakukan daripada kekerasan konvensional, karena pelaku tidak perlu bertatap muka secara langsung dengan korban.

2.9. BPJS

BPJS adalah singkatan dari Badan Penyelenggara Jaminan Sosial, yakni lembaga khusus yang bertugas untuk menyelenggarakan jaminan kesehatan dan ketenagakerjaan bagi masyarakat, PNS, serta pegawai swasta. BPJS dibagi menjadi 2 bagian yaitu ada BPJS Kesehatan dan BPJS Ketenagakerjaan. BPJS Kesehatan adalah asuransi di bidan kesehatan, atau biasa disebut Jaminan Kesehatan Nasional (JKN). JKN memberikan pelayanan kesehatan secara komprehensif melalui rujukan berjenjang tergantung pada indikasi medis pasien. Sedangkan BPJS Ketenagakerjaan adalah Jaminan Hari Tua (JHT) yang pembayarannya ditanggung oleh pengusaha dan pekerja.

2.10. Kajian Penelitian

Pada sebuah karya ilmiah dibutuhkan kajian penelitian sebagai tolak ukur penelitian selanjutnya, kajian pustaka merupakan daftar referensi mulai dari jurnal, buku, artikel dan karya ilmiah yang lainnya. Berikut ini merupakan beberapa penelitian yang bisa menjadi sumber referensi untuk penulis, antara lain.

Penelitian pertama ialah penelitian yang dikemukakan oleh [14] dengan judul "Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode *Lexicon-Based*". Penelitian ini bertujuan untuk memberi tahu kepada masyarakat khususnya pengguna internet, bahwasannya situs pembajakan itu nyata dan benar adanya. Penelitian ini lebih mengerucut tentang pembajakan artikel penelitian yaitu Sci-Hub, yang bilamana akses untuk referensi tidak bisa diperoleh karena hambatan *paywall*. Data yang digunakan adalah *tweets* yang berjumlah 33.755 dengan *tweets* yang berbahasa Inggris sebanyak 28.193 dan *tweets* yang berbahasa Indonesia sebanyak 7.555, pengklasifikasian ini berdasarkan dengan *rule* dalam *dictionary* InSet dengan hasil kualifikasi sentimen positif 59.51%, sentimen negatif 26.42%, dan sentimen netral 14.07%.

Penelitian kedua ialah penelitian yang dikemukakan oleh [4] dengan judul "Analisis Sentimen Pembelajaran Daring pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode *Naive Bayes*". Penelitian ini bertujuan untuk menganalisis opini publik terhadap pembelajaran daring pada masa pandemi COVID-19 di Indonesia pada awal November 2020. Namun, dengan adanya pembelajaran daring ini yang awal mulanya sebagai kontroversi karena singkatnya proses adaptasi. Perubahan mendadak dari pembelajaran tatap muka ke pembelajaran daring pada skala besar menyebabkan berbagai respon di kalangan masyarakat. Dengan adanya kasus ini peneliti ingin melakukan penelitian pembelajaran daring pada sebuah aplikasi Twitter menggunakan metode NBC. Dan temuan peneliti menunjukkan bahwa pembelajaran daring memiliki 30% sentimen positif, 69% sentimen negatif, dan 1% netral pada periode tersebut.

Penelitian ketiga ialah penelitian yang dikemukakan oleh [7] dengan judul "Analisis Sentimen *Cyberbullying* di Jejaring Sosial Twitter Dengan Algoritma *Naïve Bayes*". Penelitian ini bertujuan untuk mengklasifikasi kata-kata yang mengandung unsur *cyberbullying* atau tidak pada *tweet* atau cuitan yang ada pada Twitter. Sumber data yang digunakan ialah dari akun yang sering membuat *tweet* mengenai politik menggunakan Twitter API (*Application Programming Interface*). Jika pada data latih *tweet* yang tidak mengandung unsur *cyberbullying* maka akan diberikan label positif, sedangkan pada data latih *tweet* yang mengandung unsur *cyberbullying* maka diberi label negatif. Setelah meneliti menggunakan metode NBC, dapat disimpulkan bahwa hasil pengujian dengan data uji *real time*, peneliti mendapatkan nilai akurasi sebesar 76%.

Penelitian keempat ialah penelitian yang dikemukakan oleh [15] dengan judul "Sentiment Analysis of Full Day Scholl Policy Comment Using *Naïve Bayes Classifier* Algorithm". Penelitian ini bertujuan untuk melihat pendapat atau objek oleh seseorang, baik yang cenderung berpandangan negatif maupun positif. Tujuan lain dari penelitian ini adalah untuk megetahui kinerja dari algoritma NBC. Untuk pelabelan data latih, peneliti menggunakan metode *Lexicon*. Nilai akurasi pada algoritma NBC dengan pemilihan fitur trigram, model pelatihan 300 data telah mencapai nilai 80%. Simulasi ini telah membuktikan bahwa semakin besar data pelatihan dan pemilihan fitur maka akan mempengaruhi hasil akurasinya.

Penelitian kelima ialah penelitian yang dikemukakan oleh [16] dengan judul "Perbandingan Metode *Naïve Bayes Classifier* Dan *Holistic Lexicon Based* Dalam Analisis Sentimen Mahasiswa". Penelitian ini bertujuan untuk menentukan kelas dari data opini mahasiswa yang terdapat dalam *form* angket mahasiswa berupa kelas positif, negatif, dan netral. Disini peneliti menggunakan dua metode yakni, *Naïve Bayes Classifier* Dan *Holistic Lexicon Based*. Dikarenakan jika memilih perbandingan maka akan menemukan metode yang terbaik untuk memecahkan sebuah kasus. Disini peneliti menggunakan metode NBC dengan menggunakan data latih untuk menentukan kelasnya, sedangkan metode HLB menggunakan kamus kata sifat untuk penentuan kelasnya. Dapat disimpulkan bahwa metode NBC memiliki nilai *precission* dan tingkat *accurazy* yang lebih baik dibandingkan dengan metode HLB.

Penelitian keenam ialah penelitian yang dikemukakan oleh [9] "Identifikasi *Cyberbullying* pada Komentar Instagram menggunakan Metode *Lexicon-Based* dan *Naïve Bayes Classifier*". Penelitian ini bertujuan untuk mengklasifikasikan komentar yang mengandung *cyberbullying* atau *non cyberbullying* pada media sosial Instagram, terkait pemilihan Presiden Indonesia tahun 2019. Proses sistem ini dimulai dari *text preprocessing* dengan tahapan *cleaning*, *casefolding*, dan *stemming*, kemudian dilakukan proses klasifikasi menggunakan metode LB dan NBC. Dapat disimpulkan bahwa hasil performansi dari metode LB menghasilkan akurasi sebesar 58%, sedangkan metode NBC menghasilkan akurasi sebesar 97%.

Penelitian ketujuh ialah penelitian yang dikemukakan oleh [6] dengan judul "Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon Based* Dan *Support Vector Machine*". Penelitian ini bertujuan untuk mengetahui ulasan komentar pengguna Go-pay di media sosial Twitter dengan pelabelan sentimen menggunakan LB bahwa pengguna banyak berkomentar positif dengan jumlah

923 ulasan, sedangkan untuk negatif berjumlah 287 ulasan. Dari penelitian ini, bahwa klasifikasi menggunakan metode SVM dapat membandingkan kernel dengan cukup baik, untuk kernel linear mendapatkan akurasi sebesar 89.17%, sedangkan kernel polynominal mendapatkan akurasi sebesar 84.38%. dapat disimpulkan bahwa sesuai dengan hipotesa sebelumnya, presentase pengguna Go-pay lebih banyak yang berkomentar positif dibandingkan dengan komentar negatif, yaitu untuk komentar positif mendapatkan hasil 17% dari kernel linear SVM.

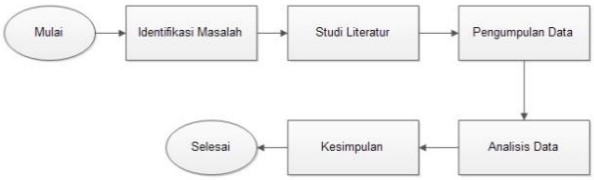
Penelitian kedelapan ialah penelitian yang dikemukakan oleh [17] dengan judul "Implementasi *Lexicon Based* Dan *Naive Bayes* Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019". Penelitian ini bertujuan untuk mengklasifikasi opini-opini masyarakat itu termasuk ke golongan yang positif, negatif, maupun netral. Salah satu sarana yang peneliti ambil untuk masyarakat gunakan dalam mekekspresikan pilihan politiknya adalah melalui media sosial Twitter. Peneliti memilih tahapan analisis sentimen yang terdiri antara lain, akuisisi data, *pre-processing*, klasifikasi data, evaluasi data dan visualisasi data. *Pre-processing* dilakukan dengan *case folding*, normalisasi data, *filtering*, ubah kata baku, *stopword* dan *stemming*. Kemudian hasil akhir dari analisis ini akan dihitung nilai akurasinya menggunakan *confusion matrix* dan visualisasi menggunakan *web server*. Dapat disimpulkan bahwa hasil dari penelitian ini adalah tingkat keakurasian antara sentimen prediksi dan sentimen aktual dengan metode LB sebesar 64.49% pada data uji, sedangkan pada data latih sebesar 94.2% serta dengan menggunakan Labelisasi dan NBC sebesar 86.53% pada data uji, sedangkan pada data latih sebesar 94.08%.

Penelitian kesembilan ialah penelitian yang dikemukakan [18] dengan judul "Sentiment Analysis of Student's Comment Using Lexicon Based Approach". Pada penelitian ini bertujuan untuk mengidentifikasi sikap positif atau negatif dari segi umpan balik siswa dengan mengukur kualitas pengajaran. Untuk menganalisis umpan balik teks siswa secara otomatiis, peneliti menggunakan pendekatan berbasis *Lexicon* untuk memprediksi tingkat pengajaran. Basis data yang digunakan pada kata-kata sentimen berbahasa Inggris dibuat sebagai sumber leksikal untuk mendapatkan polaritas kata. Peneliti dapat menentukan hasil pendapat dari guru, untuk menggambarkan tingkat pendapat positif atau negatif. Sistem telah menunjukkan bahwa hasil dari pendapat guru yang direpresentsikan sebagai positif kuat, positif sedang, positif lemah, sangat negatif, negatif lemah, atau netral.

Penelitian kesepuluh ialah penelitian yang dikemukakan oleh [19] dengan judul "Prediksi *Rating* Film Menggunakan Metode *Naïve Bayes*". Penelitian ini bertujuan untuk menganalisis agar dapat mengetahui minatnya seorang penikmat film yaitu dengan membuat penilaian-penilaian yang nantinya dapat digunakan untuk mengetahui *rating* sesuai film menggunakan metode NBC yang dimana metode ini dapat melakukan pendekatan berdasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan resiko yang ditimbulkan dalam keputusan-keputusan tersebut. Dapat disimpulkan bahwa hasil dari penelitian ini menunjukkan hasil prediksi *rating* film menggunakan metode NBC memiliki *accurazy* sebesar 55.80%, *precission* sebesar 32.41%, dan *recall* sebesar 46.70%.

Berdasarkan hasil dari beberapa penelitian yang sudah dijadikan rujukan di atas, maka dapat diambil kesimpulan bahwa metode *Naive Bayes Classifier* dan *Lexicon-Based* mempunyai hasil akhir yang berbeda. Untuk hasil metode *Naive Bayes Classifier* lebih akurat daripada metode *Lexicon-Based*. Perbedaan penelitian ini dengan penelitian yang dilakukan oleh [9] dengan menggunakan objek komentar Instagram sedangkan pada penelitian ini menggunakan objek komentar atau *tweets* pengguna media sosial Twitter, karena data dari Twitter lebih lengkap dan mengerucut dengan menggunakan fitur *hashtag* dan *top tranding*. Kelebihan selanjutnya terletak pada tahap *pre-processing*, pada penelitian terdahulu rata-rata hanya menggunakan *case folding*, *cleaning*, dan *stemming* saja. Sedangkan pada penelitian ini, menggunakan tahap *pre-processing* seperti *case folding*, *cleaning*, *tokenizing*, *stopwrod removal*, *stemming*, serta menggunakan pembobotan TF-IDF.

3. METODE PENELITIAN



Gambar 1. Metode Penelitian

3.1. Identifikasi Masalah

Tahap awal dari penelitian ialah identifkasi masalah. Permasalahan yang terjadi ada pada media sosial Twitter, yakni aturan baru yang dibuat oleh Presiden Joko Widodo terkait pembelian tanah harus menyertakan surat BPJS Kesehatan. Dengan adanya aturan baru yang telah beredar tidak sedikit kemungkinan menimbulkan pro dan kontra antara lain timbulnya *cyberbullying* atau perundungan dan ujaran kebencian terhadap orang yang dituju.

3.2. Studi Literatur

Pada studi literatur, peneliti telah mencoba memahami permasalahan tersebut dengan memanfaatkan sejumlah jurnal-jurnal dan buku. Hingga saat ini sudah melakukan studi literatur berupa *cyberbullying* dan Twitter.

3.3. Pengumpulan Data

Proses pengumpulan data (*harvesting*) *tweet* dilakukan dengan memanfaatkan *Twitter Streaming APIs*. Pencarian dan pengumpulan opini masyarakat di Twitter berdasarkan dengan kata kunci yang telah ditetapkan sebelumnya dalam kurun waktu tiga bulan. Data yang didapat dari *harvesting* disimpan ke dalam *database*.

3.4. Analisis Data

Analisis data merupakan suatu proses mengolah data menjadi informasi baru. Disini penulis mempunyai 3 tahap yakni, tahap *Pre-Processing*, pengujian dan kamus *cyberbullying* dan *non cyberbullying*. Untuk tahap-tahap *Pre-Processing* ada *labelling*, *case folding*, *cleaning*, *tokenizing*, *stopword removal*, dan *stemming*. Pengujiannya penulis menggunakan metode *Confusion Matrix*. Berikut ini merupakan contoh kamus untuk menentukan kalimat atau *tweet* tersebut termasuk ke dalam *cyberbullying* atau *non cyberbullying*.

Tabel 2. Contoh Kamus *Cyberbullying* dan *Non Cyberbullying*

No.	*Cyberbullying*	*Non Cyberbullying*
1	pea	aneh
2	goblok	iuran
3	tolol	syarat
4	ruwet	akibat
…	…	…
26	kere	duit
27	tai	rakyat
28	anjing	lucu
29	penjilat	logika
30	gila	untung

Hasil dari ulasan positif dan negatif diambil ulasan yang negatif, kemudian diklasifikasi lagi ke ulasan *cyberbullying* atau *non cyberbullying*.

a. *Cyberbullying* : terdapat berbagai *tweet* negatif yang ditunjukkan kepada suatu objek dengan tujuan mendeskriminasi seseorang, mengucilkan orang lain, serta mengungkapkan rasa kebencian kepada orang lain dengan menggunakan kata kasar.
b. *Non Cyberbullying* : terdapat pada *tweet* yang tidak ada kata kasar

4. HASIL DAN PEMBAHASAN

4.1. Pengambilan Data *Tweet*

Dengan adanya kegiatan penelitian ini, maka data yang digunakan ialah data sekunder berupa *tweets* yang diperoleh melalui proses *crawling* dengan menggunakan bahasa pemrograman *Python* serta menggunakan *pandas* dan *tweepy*. Pada proses *crawling* data terdapat beberapa tahapan yang harus dilakukan. Tahapan pertama yakni melakukan *crawling* data yaitu proses installasi *pandas* dan *tweepy*. Setelah melakukan installasi *pandas* dan *tweepy*, tahap selanjutnya adalah menentukan jumlah data yang akan diambil. Selanjutnya data yang diperoleh disimpan dalam bentuk *Comma Separated Values* (CSV).

4.2. *Pre-Processing* Data

Setelah data yang diperoleh melalui *crawling* data di *Google Colab*, maka selanjutnya dilakukanlah *Pre-Processing*. Untuk tahap *Pre-Processing* ini digunakan agar pengolahan data mentah menjadi data yang siap digunakan. Berikut ini merupakan tahapan-tahapan yang peneliti lakukan pada *pre-processing* data.

a. *Labelling*

Tabel 3. *Labelling* pada data *Tweets*

Tweet	*Labelling*
Yg riel radikul itu; BBM naik, Tarik Toll Naik, Minyak Goreng Langka, Gula dan Terigu naik, Listrik Naik, Daging juga naik, Semua hrs ikut BPJS, spy bisa membuat SIM, jual beli tanah.. Puncaknya Pemilu Ingin Ditunda…	Positif
Niat banget ngumpulin duit rakyat. Padahal gak semua orang mampu bayar iuran tiap bulan, tapi ngurus SIM, SKCK sampe jual beli tanah wajib punya BPJS Kesehatan. Ada yang ngejar setoran?	Negatif

b. *Case Folding* dan *Cleaning*

Tabel 4. *Case Folding* pada data *Tweets*

Tweet	*Case Folding dan Cleaning*
Yg riel radikul itu; BBM naik, Tarik Toll Naik, Minyak Goreng Langka, Gula dan Terigu naik, Listrik Naik, Daging juga naik, Semua hrs ikut BPJS, spy bisa membuat SIM, jual beli tanah. Puncaknya Pemilu Ingin Ditunda…	yg riel radikul itu; bbm naik, tarik toll naik, minyak goreng langka, gula dan terigu naik, listrik naik, daging juga naik, semua hrs ikut bpjs, spy bisa membuat sim, jual beli tanah puncaknya pemilu ingin ditunda
Niat banget ngumpulin duit rakyat. Padahal gak semua orang mampu bayar iuran tiap bulan, tapi ngurus SIM, SKCK sampe jual beli tanah wajib punya BPJS Kesehatan. Ada yang ngejar setoran?	niat banget ngumpulin duit rakyat. padahal gak semua orang mampu bayar iuran tiap bulan, tapi ngurus sim, skck sampe jual beli tanah wajib punya bpjs kesehatan. ada yang ngejar setoran

c. *Tokenizing*

Tabel 5. *Tokenizing* pada data *Tweets*

Case Folding dan Cleaning	*Tokenizing*
yg riel radikul itu; bbm naik, tarik toll naik, minyak goreng langka, gula dan terigu naik, listrik naik, daging juga naik, semua hrs ikut bpjs, spy bisa membuat sim, jual beli tanah puncaknya pemilu ingin ditunda	yg,riel,radikul,itu,bbm,naik, tarik,toll,naik, minyak,goreng,langka,gula, dan,terigu,naik, listrik,naik,daging,juga,naik ,semua,hrs,ikut,bpjs,spy,bis a,membuat,sim,jual,beli,tan ah,puncaknya, pemilu,ingin,ditunda
niat banget ngumpulin duit rakyat. padahal gak semua	niat,banget,ngumpulin,duit,r akyat,padahal,gak,semua,or

orang mampu bayar iuran tiap bulan, tapi ngurus sim, skck sampe jual beli tanah wajib punya bpjs kesehatan. ada yang ngejar setoran	ang,mampu,bayar,iuran,tiap ,bulan,tapi,ngurus,sim,skck, sampe,jual,beli,tanah,wajib, punya,bpjs,kesehatan,ada,ya ng,ngejar,setoran

d. *Stopword Removal*

Tabel 6. *Stopword Removal* pada data *Tweets*

Tokenizing	***Stopword Removal***
yg, riel,radikul,itu,bbm,naik,tarik,toll ,naik,minyak,goreng,langka,gula, dan,terigu,naik.listrik,naik,daging ,juga,naik,semua,hrs,ikut,bpjs,sp y,bisa,membuat,sim,jual,beli,tana h,puncaknya,pemilu,ingin,ditund a	yg,riel,radikul,itu,bbm, tarik,toll,minyak,goren g,langka,gula,terigu,lis trik,daging,hrs,bpjs,sp y,sim,jual,beli,tanah,p uncaknya,pemilu,ditun da
niat,banget,ngumpulin,duit,rakya t,padahal,gak,semua,orang,mamp u,bayar,iuran,tiap,bulan,tapi,ngur us,sim,skck,sampe,jual,beli,tanah ,wajib,punya,bpjs,kesehatan,ada, yang,ngejar,setoran	niat,banget,ngumpulin, duit,rakyat,gak,orang,b ayar,iuran,ngurus,sim,s kck,sampe,jual,beli,tan ah,wajib,bpjs,kesehata n,ngejar,setoran

e. *Stemming*

Tabel 7. *Stemming* pada data *Tweets*

Stopword Removal	***Stemming***
yg,riel,radikul,itu;,bbm,tarik ,toll,minyak,goreng,langka, gula,terigu,listrik,daging,hrs ,bpjs,spy,sim,jual,beli,tanah, puncaknya,pemilu,ditunda	yg riel radikul itu bbm tarik toll minyak goreng langka gula terigu listrik daging hrs bpjs spy sim jual beli tanah puncak milu tunda
niat,banget,ngumpulin,duit,r akyat,gak,orang,bayar,iuran, ngurus,sim,skck,sampe,jual, beli,tanah,wajib,bpjs,keseha tan,ngejar,setoran	niat banget ngumpulin duit rakyat gak orang bayar iur ngurus sim skck sampe jual beli tanah wajib bpjs sehat ngejar setor

4.3. Pembobotan TF-IDF

Setelah melewati tahap *pre-processing* diatas, maka proses selanjutntya adalah pemberian nilai atau bobot pada setiap *term*. Pada setiap pembobotan nilai dari *term* tersebut akan dijadikan inputan pada saat melakukan proses klasifikasi data. Berikut ini merupakan sebagian hasil dari pembobotannya.

```
  (0, 59)    0.3495033932767893
  (1, 24)    0.364641828888114
  (2, 24)    0.30550565550020903
  :    :
 (296, 8)    0.28049609834739037
 (297, 44)   0.22201665299353462
 (298, 69)   0.22893831564732442
```

Pada hasil pembobotan diatas, maka dapat ditarik kesimpulan jika suatu kata atau *term* terdapat dalam suatu dokumen sebanyak 5 kali maka diperoleh bobot 1 (satu). Tetapi jika *term* tidak terdapat dalam dokumen tersebut, maka bobotnya adalah 0 (nol).

4.4. Performa Algoritma *Naïve Bayes Classifier* dan *Lexicon Based*

Naive Bayes memang cocok untuk klasifikasi *biner* dan *multiclass*. Metode ini menerapkan teknik *supervised* klasifikasi objek di masa depan dengan menetapkan label kelas. Berikut ini merupakan hasil dari performa NBC.

```
Accuracy    :  0.8
Precission  :  0.7857142857142857
Recall      :  0.55
f1_Score    :  0.6470588235294117
```

Berdasarkan analisa diatas, dapat disimpulkan bahwa *tweets* negatif lebih banyak daripada *tweets* positif nya, dengan tingkat ke *accuracy*an sebanyak 80%.

Berikut ini merupakan hasil *tweets* yang mengandung unsur *cyberbullying* dan *non cyberbullying*.

Tabel 8. Contoh *Tweet Cyberbullying* dan *Non Cyberbullying*

No.	***Tweet***	***Labelling***
1	niat banget ngumpulin duit rakyat gak orang bayar iur ngurus sim skck sampe jual beli tanah wajib bpjs sehat ngejar setor	*non cyberbullying*
2	wajib kartu bpjs sehat jual beli tanah ngurus sim sampe haji trus ntar tau2 aja iur bulanan nya apalll banget main nya	*non cyberbullying*
3	atur perintah rakyat bingung kemarin puksin skrg bpjs semua layan dri jual beli tanah umroh stnk sim wajib hrus bpjs mau peras rakyat dgn atur ini another	*non cyberbullying*
…	…	**…**
194	tidak logis lah pea penting bayar pajak apa tau kalau bpjs ladang manipulsi rumah sakit	*cyberbullying*
195	goblok contoh bpjs buat syarat jual beli tanah apa relevansi sama dengan bpjs buat masuk masjid ada gimana caranya negara narik duit dari rakyat selain dari pajak lewat bpjs tolol	*cyberbullying*
196	aturan apa ini anjing	*cyberbullying*

4.5. Pembahasan

Pada pembahasan ini, penulis ingin menyampaikan bahwa pengambilan data dari media sosial Twitter kemudian dikumpulkan menjadi satu untuk di evakuasi dan di bentuk agar menjadi sebuah penelitian dengan cara *crawling* data. Setelah proses

crawling, maka data yang di dapat sebanyak 512 *tweet* dengan menggunakan *Google Colab*. Tahap selanjutnya adalah menghapus atau menghilangkan *tweet* yang bersfiat *double* atau tidak membahas tentang jual beli tanah, pembuatan SIM, SKCK, haji dan umrah. Yang sebelumnya data *tweet* sebanyak 512 menjadi 298 data *tweet* saja. Setelah proses pengambilan data selesai, maka *file* data *tweet* yang sudah diperoleh akan tersimpan menjadi *file Comma Separated Values* (CSV).

Selanjutnya masuk pada tahap *pre-processing*. Pada tahap ini tidak ada kendala dan *program* bisa memanggil *file* dengan format nama tweetbpjs (2). Lanjut ke tahap *labelling* atau pelabelan positif dan negatif. Setelah tahap pelabelan selesai, maka lanjut ke tahap *case folding, cleaning, tokenizing, stopword removal*, dan *stemming*. Jika sudah melalui tahap *pre-processing* maka, langkah selanjutnya adalah tahap pembobotan menggunakan TF-IDF. Setelah pembobotan selesai, maka lanjut ke tahap yang terkakhir yaitu, klasifikasi *Naive Bayes Classifier* dengan dua *class* positif dan negatif menggunakan perhitungan *Confusion Matrix* dengan menghasilkan nilai *accracy* 0.8, *precission* 0.78, *recall* 0.55 dan *F1_Score* 0.64. Akurasi diperoleh dengan membandingkan jumlah data hasil klasifikasi (prediksi) yang sesuai dengan jumlah keseluruhan data. Semakin tinggi nilai akurasi yang diperoleh, maka hasil klasifikasi semakin baik. Oleh karena itu penelitian ini juga menghitung nilai *recall* dan presisi. *Recall* diperoleh dengan membandingkan jumlah data hasil klasifikasi yang relevan. Sedangkan f1-*Score* digunakan untuk mengetahui keseimbangan antara presisi dan *recall* yang didapat dari perhitungan tersebut.

Setelah proses pelabelan data *tweet* negatif yang berjumlah 186 *tweet* menjadi 36 *tweet* untuk *cyberbullying* dan 160 *tweet* untuk yang *non cyberbullying*. Maka, *cyberbullying* mendapatkan nilai 0,22 sedangkan *non cyberbullying* mendapatkan nilai 0,77. Disini dapat disimpulkan bahwa *tweet* yang mengandung unsur *cyberbullying* memiliki nilai lebih rendah daripada *non cyberbullying*. Disini penulis mendapatkan *feedback* dari institusi BPJS Kesehatan terkait laporan yang peneliti tulis yaitu dengan banyaknya *tweets* negatif yang ada di Twitter maka pihak BPJS Kesehatan akan melakukan evaluasi lagi dan memperbaiki kualitas pelayanan pada pengguna BPJS.

Setelah semua proses dilalui maka kesimpulan yang didapatkan dengan perbandingan dua metode yakni, *Naive Bayes Classifier* menghasilkan akurasi sebesar 0,8 sedangkan *Lexicon Based* menghasilkan akurasi sebesar 0,22.

5. KESIMPULAN DAN SARAN

Kesimpulan yang didapat bahwa penerapan metode *Naive Bayes Classifier* dan *Lexicon Based* digunakan untuk menganalisis sentimen pada *tweet* atau komentar yang membahas terkait aturan baru BPJS. Untuk pengujian menggunakan metode *Confusion Matrix* dan hasil pada klasifikasi memperlihatkan bahwa metode *Naive Bayes Classifier* dan *Lexicon Based* memberikan nilai presentase *accuracy 80%*, *precission* 80%, *recall* 90% dan *F1_Score* 86%. Sedangkan pada *Lexicon Based* memberikan nilai presentase akurasi sebanyak 22%. Berdasarkan dari pengkajian hasil penelitian di lapangan maka penulis bermaksud memberikan saran yang mudah-mudahan dapat bermanfaat antara lain, untuk menambah data latih dan data uji lebih banyak lagi, penambahan kata pada kamus sentimen sesuai konteks masalah yang akan diidentifikasi sehingga akan meningkatkan perolehan akurasi, dan peneliti lebih lanjut diharapkan menerapkan algoritma yang berbeda untuk mendapatkan hasil yang lebih akurat dalam analisis sentimen.

DAFTAR PUSTAKA

[1] Al Zimantyo. (2021). Hukum pada Cyber Bullying https://www.kompasiana.com/alzimantyo2045053/60d098d46ae34e34c92fd6f2/hukum-pada-cyberbullying

[2] Destitus, C., Wella, W., & Suryani S. (2020). Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia https://ejournals.umn.ac.id/index.php/SI/article/view/1740

[3] Basri, H. (2017). Peranmedia Sosial Twitter Dalam Interaksi Sosial Pelajar Sekolah Menengah Pertama Di Kota Pekanbaru (studi kasus pelajar SMPN 1 kota Pekanbaru). *Strategi Bertahan Hidup Petani Penggarap Di Jorong Sarilamak Nagari Sarilamak Kecamatan Harau Kabupaten Lima PuluhKota*,*4*(1),1–13. https://media.neliti.com/media/publications/183768-ID-partisipasi-masyarakat-dalam-pelaksanaan.pdf

[4] Samsir, S., Ambyar, A., & Verawardina U. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes http://stmik-budidarma.ac.id/ejurnal/index.php/mib/article/view/2580

[5] Lin, Y., Wang, X., & Zhou, A. (2016). Opinion spam detection. *Opinion Analysis forOnlineReviews*,*May*,79–94. https://doi.org/10.1142/9789813100459_0007

[6] Mahendrajaya, R., Buntoro, G. A., & Setyawan, M. B. (2019). Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine. *Komputek*, *3*(2), 52. https://doi.org/10.24269/jkt.v3i2.270

[7] Maulana, F. A., Ernawati, I., Labu, P., & Selatan, J. (2020). Analisa sentimen cyberbullying di jejaring sosial twitter dengan algoritma naïve bayes. *Seminar Nasional*

Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA, 529–538.https://conference.upnvj.ac.id/index.php/senamika/article/view/619

[8] Nurzahputra, A., & Muslim, M. A. (2016). Analisis sentimen pada opini mahasiswa menggunakan languange processing https://ilkom.unnes.ac.id/snik/prosiding/2016/18.%20SNIK_366_Analisis%20Sentimen.pdf

[9] Syarif, R. D., Herdiani, A., & Astuti, W. (2019). … Cyberbullying Pada Komentar Instagram Menggunakan Metode Lexicon-based Dan Naïve Bayes Classifier (studi Kasus: Pemilihan Presiden Indonesia Tahun ….*eProceedings…*,*6*(2),8838–8851. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/download/9876/9735

[10] Indraloka, D. S., & Santosa, B. (2017). Penerapan Text Mining untuk Melakukan Clustering Data Tweet Shopee Indonesia. *Jurnal Sains dan Seni ITS*, *6*(2), 6–11. https://doi.org/10.12962/j23373520.v6i2.24419

[11] Yunarfi, G. R., & Simdy, R. (2021). Implementasi Text Mining untuk mengetahui Kata Abreviasi dalam Percakapan Media Sosial http://journal.uvers2.ac.id/index.php/jodens/article/view/43

[12] Sembodo, J. E., & Setiawan, E. B. (2016). Data Crawling Otomatis pada Twitter https://www.researchgate.net/profile/Erwin-Setiawan-3/publication/308815966_Data_Crawling_Otomatis_pada_Twitter/links/5bb5832392851ca9ed37a37e/Data-Crawling-Otomatis-pada-Twitter.pdf

[13] Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

[14] Undap, M. G., Rantung, V. P., & Rompas, P. T. D. (2021). Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based. *Jointer – Journal of Informatics Engineering*, *2*(2), 39–46

[15] Al Fath, M. K., Arini, A., & Hakiem, N. (2020). Sentiment Analysis Of Full Day School Policy Comment Using Naïve Bayes Classifier Algorithm. *SinkrOn*, *5*(1), 107–114. https://doi.org/10.33395/sinkron.v5i1.10564

[16] Suprianto, S. (2019). Perbandingan Metode Naïve Bayes Classifier Dan Holistic Lexicon Based Dalam Analisis Sentimen Angket Mahasiswa. *JSI: Jurnal Sistem Informasi (E-Journal)*, *11*(2), 1763–1771. https://doi.org/10.36706/jsi.v11i2.9140

[17] Aulia, G. N., & Patriya, E. (2019). Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019. *JurnalIlmiahInformatikaKomputer*,*24*(2),140–153.

[18] Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 149–154. https://doi.org/10.1109/ICIS.2017.7959985

[19] Pratiwi, R. W., Yusuf, D., & Nugroho, S. (2016). Prediksi Rating Film Menggunakan Metode Naïve Bayes. *Jurnal Teknik Elektro*, *8*(2), 60-63. https://journal.unnes.ac.id/nju/index.php/jte/article/view/7764