e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

ANALISIS SENTIMEN TERHADAP VAKSIN NUSANTARA PADA MEDIA SOSIAL *YOUTUBE* MENGGUNAKAN METODE NAÏVE BAYES DAN SELEKSI FITUR PARTICLE SWARM OPTIMIZATION

Taufik Adi Prasetyo[1], Didit Widiyanto[2], Desta Sandya Prasvita[3]
Prodi S1 Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta 12450
taufikadipras@gmail.com, didit.widiyanto@upnvj.ac.id, desta.sandya@upnvj.ac.id

Abstrak. *Youtube* adalah salah satu dari sekian banyak media sosial yang dapat digunakan untuk memberikan komentar yang dapat diakses masyarakat Indonesia, Penelitian kali ini, menggunakan salah satu video di *youtube* yang berjudul "Peneliti Utama Jawab Kontroversi Vaksin Nusantara - ROSI (1)" dan di unggah oleh akun KOMPASTV. Komentar yang diambil menggunakan komentar pertama yang diketik bukan merupakan balasan dari suatu komentar. Setelah itu dilakukan pelabelan untuk data komentar tersebut, dan dilakukan praproses teks, kemudian kata-kata tersebut diberikan bobot dengan menggunakan pembobotan TF-IDF (*Term Frequency – Inverse Document Frequency*). Setelah itu dilakukan seleksi fitur menggunakan PSO (*Particle Swarm Optimization*) hingga 1000 iterasi dan didapatkan fitur yang digunakan sejumlah 933 fitur, lalu untuk pembuatan model klasifikasi, dilakukan sampling menggunakan SMOTE, yang awalnya 645 data negatif, dan 354 positif, menjadi berimbang 645 negatif, dan 645 positif, didapatkan hasil klasifikasi menggunakan *Naïve Bayes* dengan akurasi 82,5%, nilai presisi 78,7%, dan nilai *recall* 89,1%.

Kata Kunci: *Youtube*, Analisis Sentimen, *Naïve Bayes*, *Particle Swarm Optimization*

1 Pendahuluan

Virus corona merupakan wabah penyakit yang dapat menular dari satu individu ke individu lainnya, penyakit ini sangat berbahaya dikarenakan mengincar bagian vital dan menyebabkan gangguan, salah satunya gangguan pernapasan. Penyakit ini diidentifikasi pertama kali di Wuhan, China, pada Desember 2019. Sejak saat itu Indonesia sudah waspada akan datangnya virus tersebut dan menyiapkan beberapa langkah untuk menghentikan penyebaran corona virus. Penanganan awal yang dilakukan pemerintah Indonesia pada saat itu adalah memberi instruksi kepada kedutaan Indonesia yang ada di China untuk memperhatikan WNI yang ada di Wuhan. Memasuki tahun 2020, dampak corona virus di Indonesia terus mengalami kenaikan dari waktu ke waktu, hingga memakan banyak korban jiwa. Oleh karena itu, pemerintah Indonesia merencanakan program vaksinasi untuk menghadapi Coronavirus, diawali dengan mempersiapkan 1,2 juta dosis vaksin buatan Sinovac Biotech.

Namun, karena vaksin ini merupakan vaksin jenis baru, dan cenderung belum memiliki umur yang lama dibanding vaksin milik lembaga lainnya, maka penulis merasa perlu menganalisa bagaimana respon atau tanggapan orang-orang terhadap Vaksin Nusantara, baik yang sudah mendapatkannya, maupun yang belum, respon atau tanggapan tersebut didapatkan dari media sosial *youtube*. diperlukan analisis sentimen untuk mengklasifikasikan respon tersebut, karena ada banyak respon atau tanggapan yang diberikan oleh masyarakat.

Pada penelitian sebelumnya, yang sudah ada dengan topik analisis sentimen yang ditulis oleh (Nurjanah, Perdana, & Fauzi, 2017) dengan metode KNN (*K Nearest Neighbor*) untuk tayangan televisi menghasilkan akurasi sebesar 82,5%, penelitian analisis sentimen lainnya oleh (Apriani & Gustian, 2019) dengan *naïve bayes* untuk ulasan produk di aplikasi tokopedia berhasil menghasilkan akurasi sebesar 97,13%, penelitian (Sari & Hayuningtyas, 2019) yaitu analisis sentimen berbasiskan web dengan metode *naïve bayes* memiliki akurasi sebesar 70%, dan penelitian analisis sentimen oleh (Que, Iriani, & Purnomo, 2020) dengan menggunakan *Support Vector Machine* dan *Particle Swarm Optimization* (PSO) menghasilkan akurasi sebesar 96,04%.

Berdasarkan penelitian yang telah dilakukan sebelumnya, diperlukan *text mining* analisis sentimen untuk bisa mengolah respon masyarakat menjadi sentimen positif, atau negatif. Untuk melakukan analisis sentimen, diperlukan juga metode klasifikasi yang menggunakan salah satu algoritma *machine learning* yaitu *naïve Bayes*

(NB), dimana pada penelitian sebelumnya dengan *naïve bayes* mendapatkan akurasi 90%, serta pada penelitian sebelumnya yang dengan penambahan algoritma *Particle Swarm Optimization* (PSO) menghasilkan akurasi 96,04%, maka hal ini dirasa penulis dapat membantu membangun pembuatan model klasifikasi penelitian ini, kemudian akan dilakukan percobaan penelitian dengan data tidak seimbang (*imbalanced data*), serta metode sampling seperti *random undersampling* dan SMOTE (*Synthetic Minority Oversampling Technique*) yang merupakan metode untuk menangani ketidakseimbangan kelas (Siringoringo, 2018), sehingga data yang didapat dapat menjadi bahan evaluasi bagi pemerintah Indonesia untuk menghadirkan vaksin yang sesuai dengan masyarakat Indonesia, baik itu dari segi produk maupun layanan.

2 Tinjauan Pustaka

2.1 *Youtube*

Youtube merupakan salah satu situs media sosial yang paling sering digunakan, yang dibuat pada tahun 2005 dan sekarang memiliki lebih dari satu miliar pengguna, memungkinkan ratusan juta jam total waktu menonton video setiap hari. Media sosial seperti *Youtube* memiliki potensi besar untuk memberikan kemudahan akses informasi[1].

2.2 PSO (*Particle Swarm Optimization*) dan PySwarms

Particle Swarm Optimization (PSO) adalah teknik pencarian heuristik yang secara iteratif meningkatkan serangkaian solusi kandidat fitur yang diberikan ukuran nilai yang objektif, Implementasi dari PSO dapat ditemukan di beberapa algoritma *python* evolusioner.
PySwarms adalah library untuk *Particle Swarm Optimization* (PSO) yang menyediakan satu set kelas primitif yang berguna untuk menyelesaikan optimasi berkelanjutan dan masalah kombinatorial. [2].
Pada penelitian ini, PSO digunakan untuk melakukan seleksi fitur apa saja yang akan digunakan dalam penelitian ini didefinisikan pada persamaan (1) dan (2) :

$$Vi(t) = Vi(t-1) + c1r1[Xpbesti - Xi(t)] + c2r2[Xgbest - Xi(t)]\ ) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (1)$$

$$Xi(t) = Xi(t-1) + Vi(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots... (2)$$

dengan $Vi(t)$ kecepatan partikel i saat iterasi t, dan $Xi(t)$ posisi partikel i saat iterasi t, dan $c1$, $c2$ merupakan Learning rates kemampuan individual dan pengaruhnya, dan $r1$ dan $r2$: Bilangan uniformal distribusi interval 0 dan 1, dan $Xpbesti$ adalah posisi terbaik partikel i, lalu $Xgbes$t adalah Posisi terbaik global

2.3 SMOTE *(Synthetic Minority Oversampling Technique)*

SMOTE kepanjangan dari *Synthetic Minority Oversampling Technique,* merupakan teknik oversampling dengan cara menyeimbangkan data,dengan mengambil k data dari k-NN dari setiap data komentar dari kelas yang minor [3].

Tahapan dari algoritma SMOTE dapat dijabarkan seperti berikut [4] :

1. Pilih sampel random pada data, contoh D1.
2. Menerapkan K-NN pada D1.
3. Mengambil tetangga D1 yang memenuhi kriteria dari K-NN (*random*), misal D2.
4. Membuat data baru D1' dengan persamaan (2).
5. Lalu proses diatas dilakukan berulang kali hingga didapatkan jumlah data kelas minoritas = kelas mayoritas.

Persamaan untuk membuat data baru menggunakan persamaan (2) sebagai berikut :

$$D1' = D1 + rand(0,1)*(D2-D1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

dengan *D1'* merupakan Data baru dari D1, lalu *D1 adalah* Data yang akan di duplikasi, dan *D2* adalah Data tetangga dari data D1

2.4 *Naïve Bayes*

Algoritma *naïve bayes* adalah algoritma dengan performa yang sangat baik untuk data yang tidak memiliki hubungan antar atribut, walaupun ketidakterikatan antar data biasanya sulit ditemukan pada data di kehidupan nyata, dengan sifatnya yang tidak terhubung (terkait) dengan atribut lain, maka algoritma *naïve bayes* merupakan algoritma yang paling efisien untuk melakukan proses klasifikasi [5].

Berikut persamaan (3) dari teorema bayes :

$$P(C|X) = P(X|C).P(C) \,/\, P(X) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

dengan X adalah data yang belum diketahui kelasnya, dan C adalah data hipotesis yang termasuk ke suatu kelas P(C|X) adalah probabilitas hipotesis C berdasarkan X, dan P(X|C) adalah probabilitas hipotesis X berdasarkan C P(C) adalah probabilitas hipotesis C (prior probability), dan P(X) adalah probabilitas X

Dalam melakukan klasifikasi untuk data penelitian, akan dipakai persamaan (6) berikut [6] :

$$CMAP = argmax\ p\ P(p) \prod i\ P(wi \mid p) \ldots \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

dengan $CMAP$, adalah kelas dengan probabilitas terbesar, dan $argmax$ p adalah nilai maksimum untuk kelas p, dan P(p) adalah probabilitas tampilnya dokumen untuk kelas p, dan P(Wi | p) adalah probabilitas tampilnya kata Wi untuk kelas p.

2.5 *Term Frequency – Inverse Document Frequency (TF – IDF)*

Pembobotan *Term Frequency – Inverse Document Frequency (TF – IDF)* dilakukan untuk merubah data dari hasil *preprocessing* menjadi data numerikal statistik, untuk memutuskan hubungan kata (*term*) terhadap dokumen dengan mendapatkan bobot yang digunakan pada proses klasifikasi [7].

Persamaan *Term Frequency* seperti berikut :

$$TF(word) = f_{word,d} / \sum word,d \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Keterangan :

$fword,d$ = frekuensi kata ($word$) di dokumen

$\sum word,d$ = total semua data yang ada di dokumen

Setelah itu menghitung *Inverse Document Frequency* seperti berikut :

$$IDF(word) = log\ \frac{N}{dfword} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

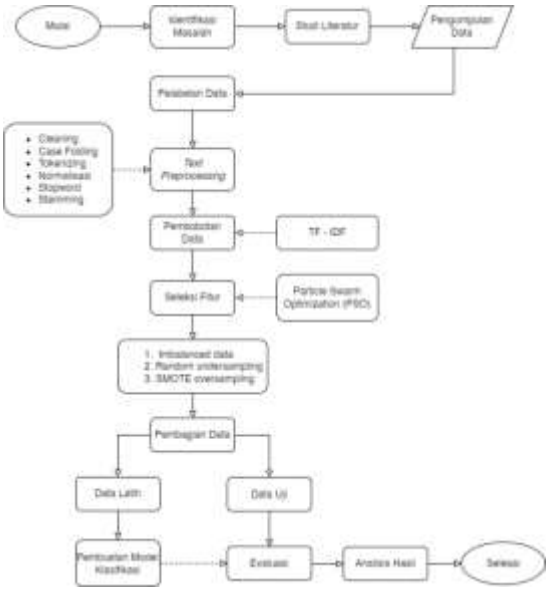
Keterangan :

N = total dokumen

df_{word} = total dokumen dimana *term* t muncul di dalamnya

3 Metodelogi Penelitian

Dalam melakukan klasifikasi pada penelitian kali ini hingga mencapai tujuan penelitian, Flowchart untuk analisis sentimen yang dilakukan pada penelitian ini memiliki alur dan tahapan seperti pada Gambar 1 sebagai berikut :



Ganmbar 1. Alur Penelitian

Studi literatur. Setelah mengidentifikasi topik yang akan dibahas, selanjutnya mencari referensi dan bacaan sebagai sumber pustaka sebagai dasar acuan ilmu pengetahuan dengan metode yang dilakukan, yaitu membaca literatur seperti buku berbentuk *e-book* maupun jurnal yang terkait dengan penelitian *text mining,* proses klasifikasi menggunakan algoritma *naïve bayes*, atau analisis sentimen yang menggunakan proses klasifikasi lain, serta seleksi fitur *particle swarm optimization*, dan SMOTE *oversampling.* Setelah melakukan proses tersebut kemudian menuliskan sumber tersebut ke dalam daftar pustaka.

Identifikasi Masalah. Tahap pertama yang dilakukan adalah identfikasi masalah, proses identifikasi merupakan bagian yang menjelaskan masalah lebih mendalam, tentang isu yang akan dibahas, seperti pada penelitian ini, yaitu permasalahan pengklasifikasikan komentar video *youtube* terhadap Vaksin Nusantara menggunakan algoritma klasifikasi *naïve bayes* dan *particle swarm optimization*.

Pengumpulan Data. Kemudian dilakukan pengumpulan data yang akan digunakan untuk penelitian kali ini, data yang digunakan diambil dari *youtube* dengan menggunakan *Google Apps Script* dan dihubungkan dengan video yang akan diambil komentarnya, setelah itu data di *crawling* dengan menggunakan fungsi *JavaScript*, data yang diperoleh adalah komentar yang terdapat pada video "Peneliti Utama Jawab Kontroversi Vaksin Nusantara - ROSI (1)"

Pelabelan Data. Langkah awal adalah melakukan pelabelan pada suatu komentar, Standar penentuan untuk kelas positif adalah komentar yang memiliki sentimen mendukung vaksin nusantara serta pihak terkait, dan untuk kelas negatif diberikan kepada komentar yang mengandung kalimat tidak mendukung dan bernada kebencian terhadap vaksin nusantara serta pihak yang terkait, bahkan untuk kerumitan yang ada pada data, contohnya ulasan dengan tata Bahasa yang kurang baik atau ambigu.

Text Preprocessing. Selanjutnya melakukan *text preprocessing* pada data yang diteliti untuk membersihkan data awal tahapan untuk membersihkan data secara umum dengan melakukan *Cleaning, Case Folding, Tokenization, Normalization, Stopword Removal,* dan *Stemming*.

Pembobotan Kata. Setelah melakukan praproses pada data yang akan diteliti, kemudian melakukan pembobotan untuk seluruh kata, dimana nantinya untuk setiap kata memiliki nilai, dan nilai tersebut akan digunakan pada proses klasifikasi. Pembobotan yang dilakukan menggunakan TF-IDF *(Term Frequency — Inverse Document Frequency)* yang dilakukan sesuai dengan persamaan, untuk mencari nilai yang mempresentasikan data latih.

PSO (*Particle Swarm Optimization*). *Particle Swarm Optimization* (PSO) adalah algoritma seleksi fitur yang terinspirasi dari perilaku kawanan burung yang secara berkawanan mencari makanan. penggunaan *Particle Swarm Optimization* bisa diibaratkan menggunakan perumpaan fitur sebagai burung yang sedang berusaha untuk mencari makanan, kemudian secara berulang-ulang mencari dan mengikuti burung yang paling dekat dari target makanan, karena itu pada penelitian ini akan digunakan *stop criteria* pada jumlah iterasi tertentu supaya semakin terpilih fitur fitur yang bernilai paling optimal.

SMOTE (*Synthetic Minority Oversampling Technique*) Pada penelitian ini, terdapat ketidakseimbangan data (*imbalanced data*), maka digunakan algoritma *oversampling* SMOTE untuk menanganinya, SMOTE kepanjangan dari *Synthetic Minority Oversampling Technique,* merupakan teknik *oversampling* dengan cara menyeimbangkan data, dengan mengambil k data dari k-NN dari setiap data komentar dari kelas yang minor [3].

Klasifikasi. Kemudian melakukan klasifikasi menggunakan algoritma *naïve bayes* data komentar *Youtube* dengan melakukan pembobotan pada kata dan membagi data latih dan data uji dengan konfigurasi pembagian secara seimbang sebesar 80% untuk data latih, dan 20% untuk data uji. Setelah itu membentuk model klasifikasi untuk menguji data yang akan diuji menggunakan persamaan *naïve bayes.*

Evaluasi. Langkah berikutnya adalah melakukan peoses evaluasi untuk melihat bagaimana keakuratan dan akurasi klasifikasi yang telah dijalankan, Pengujian yang digunakan pada penelitian ini yaitu menggunakan metode *confusion matrix* untuk melihat keakuratan dengan cara melihat 4 variabel representasi yang terdiri dari TP (data positif yang benar diprediksi sebagai positif), TN (data negatif yang benar diprediksi sebagai negatif), FP (data negatif yang salah diprediksi sebagai positif), FN (data positif yang salah diprediksi sebagai negatif). untuk melihat tingkat akurasi, presisi, dan *recall*.

4 Hasil dan Pembahasan

Data yang digunakan pada penelitian ini, merupakan data yang diambil dari komentar pada sebuah video pada aplikasi media sosial *Youtube*, dengan rincian video tersebut adalah video berjudul "Peneliti Utama Jawab Kontroversi Vaksin Nusantara – ROSI (1)" dengan akun pengunnggah video yaitu KOMPASTV. Proses pengumpulan data menggunakan *Integrated Development Environment (IDE) Google Apps Script,* yang merupakan platform *JavaScript* berbasiskan internet serta mengotomatisasikan beberapa layanan milik *Google.* Data yang diambil merupakan komentar yang ada pada rentang waktu April 2021 hingga Februari 2022, didapatkan sebanyak 3652 data, kemudian dilakukan seleksi secara manual untuk menghapus komentar yang bersifat netral ataupun tidak berhubungan dengan vaksin nusantara, maka didapatkan sisa 1008 data, dan disimpan ke dalam *Google Sheet* ke dalam file ekstensi berbentuk xlsx. Setelah data berhasil diambil kemudian melakukan proses labeling atau memberi label pada data, pada tahap ini pelabelan dilakukan secara manual dan dikerjakan oleh 3 annotator, berdasarkan opini pribadi, pengetahuan yang dimiliki dari masing masing annotator. Label yang dipakai pada penelitian kali ini menggunakan dua buah kelas label, yaitu kelas positif, dan kelas negatif, label kelas netral tidak digunakan karena tidak sesuai dengan tujuan mencari sentimen baik dan buruk dari topik "Vaksin Nusantara".

Cleaning. Tahap awal yang dilakukan yaitu *Cleaning,* yaitu membersihkan data dari noise yang terdiri dari, menghapus URL atau *link* yang tersedia, lalu nama atau *username*, simbol-simbol, *hashtag*, tanda baca, dan karakter *unicode*. Tahapan ini bertujuan untuk memudahkan proses berikutnya mendapatkan makna yang penting dari data komentar, berikut juga dilakukan pengecekan untuk data yang bernilai *null*, dan data yang memiliki duplikasi, dan setelah dilakukan pemeriksaan, maka didapatkan 0 data yang bernilai *null*, dan 9 data yang memiliki duplikasi, maka 9 data tersebut kemudian dihilangkan untuk membersihkan data, sehingga tersisa 999 data yang diolah**.**

Tabel 1 Data Sampel Cleaning Data

Data Asli Komentar Youtube	Hasil Cleaning Data
Ini bahaya bawa2 TNI dalam vaksin nusantara.... Ini harus di konfirmasi Pak Kasad apa benar organisasi TNI terlibat di vaksin nusantara	Ini bahaya bawa TNI dalam vaksin nusantara ini harus di konfirmasi Pak Kasad apa benar organisasi TNI terlibat di vaksin nusantara

Case folding. Case folding dilakukan untuk membuat semua huruf yang terkandung dalam data komentar menjadi seragam atau sama menjadi huruf yang kecil (*lowercase*), sehingga beberapa kata yang mengandung huruf kapital (*uppercase*) diubah seluruhnya untuk meminimalkan terjadinya *case sensitive.*

Tabel 2. Hasil Case folding

Hasil Cleaning Data	Hasil Case Folding Data
Ini bahaya bawa TNI dalam vaksin nusantara ini harus di konfirmasi Pak Kasad apa benar organisasi TNI terlibat di vaksin nusantara	ini bahaya bawa tni dalam vaksin nusantara ini harus di konfirmasi pak kasad apa benar organisasi tni terlibat di vaksin nusantara

Tokenization. Setelah tahap *case folding* dilakukan, kemudian melanjutkan ke proses tokenisasi, pada tahap ini dilakukan pemisahan pada teks menjadi sebuah kata dengan dibagi dari spasi.

Tabel 3. Hasil Tokenization

Hasil Case Folding Data	Hasil *Tokenization*
ini bahaya bawa tni dalam vaksin nusantara ini harus di konfirmasi pak kasad apa benar organisasi tni terlibat di vaksin nusantara	['ini', 'bahaya', 'bawa', 'tni', 'dalam', 'vaksin', 'nusantara', 'ini', 'harus', 'di', 'konfirmasi', 'pak', 'kasad', 'apa', 'benar', 'organisasi', 'tni', 'terlibat', 'di', 'vaksin', 'nusantara']

Normalization. Selanjutnya setelah melakukan *normalization*, dilakukan pemeriksaan seluruh kata kata yang terdapat pada komentar youtube untuk melihat apakah sudah sesuai dengan ejaan yang ada di KBBI (Kamus Besar Bahasa Indonesia), pada proses ini pemeriksaan dilakukan secara manual untuk melihat kata kata apa saja yang perlu diperbaiki ejaannya, karena setiap komentar diketik oleh masing masing orang, maka kemungkinan terdapat banyak variasi kata tergantung dari cara masing masing si penulis komentar.

Tabel 4. Hasil Noramlization

Hasil *Tokenization*	Hasil *Normalization*
['ini', 'bahaya', 'bawa', 'tni', 'dalam', 'vaksin', 'nusantara', 'ini', 'harus', 'di', 'konfirmasi', 'pak', 'kasad', 'apa', 'benar', 'organisasi', 'tni', 'terlibat', 'di', 'vaksin', 'nusantara']	['ini', 'bahaya', 'bawa', 'tni', 'dalam', 'vaksin', 'nusantara', 'ini', 'harus', 'di', 'konfirmasi', 'pak', 'kasad', 'apa', 'benar', 'organisasi', 'tni', 'terlibat', 'di', 'vaksin', 'nusantara']

Stopword Removal. Pada tahap *stopword removal*, kata dari data komentar yang tidak penting dihapus supaya tidak mengganggu proses klasifikasi, *Stopword* yang digunakan adalah *Stopword* sastrawi yang berbahasa Indonesia, dan juga menggunakan *stopwords* sendiri untuk memperlengkap kata kata yang tidak diinginkan

Tabel 5. Hasil Stopword Removal

Hasil *Normalization*	Hasil *Stopword* Removal
['ini', 'bahaya', 'bawa', 'tni', 'dalam', 'vaksin', 'nusantara', 'ini', 'harus', 'di', 'konfirmasi', 'pak', 'kasad', 'apa', 'benar', 'organisasi', 'tni', 'terlibat', 'di', 'vaksin', 'nusantara']	bahaya bawa tni vaksin nusantara konfirmasi kasad organisasi tni terlibat vaksin nusantara

Stemming. Selanjutnya, setelah dilakukan *stopword removal*, maka dilakukan proses *stemming*, yaitu memeriksa seluruh kata dan memproses dengan merubah kata kata yang memiliki imbuhan menjadi kata dasarnya saja.

Tabel 6. Hasil Stemming

Hasil *Stopword* Removal	**Hasil *Stemming***
bahaya bawa tni vaksin nusantara konfirmasi kasad organisasi tni terlibat vaksin nusantara	bahaya bawa tni vaksin nusantara konfirmasi kasad organisasi tni libat vaksin nusantara

Pembobotan Kata TF-IDF. Tahap berikutnya dari data komentar yang telah dibersihkan, data tersebut, yang berjumlah 999 komentar merupakan data yang bersifat kualitatif, sehingga perlu diubah terlebih dahulu menjadi data kuantitatif, dimana dilakukan vektorisasi atau pembobotan agar dapat diolah oleh model *machine learning,* metode pembobotan yang digunakan pada kali ini adalah menggunakan TF-IDF (*Term Frequency – Inverse Document Frequency).* Proses pertama pada tahap ini adalah menghitung *term frequency*, kemudian menghitung *inverse document frequency.* Dan yang terakhir menghitungnya menggunakan rumus TF-IDF.

Seleksi Fitur PSO (*Particle Swarm Optimization*).
Metode PSO (*Particle Swarm Optimization*) yang digunakan berasal dari library *pyswarm,* selanjutnya menggunakan fungsi optimizer untuk memilih fitur dengan nilai terbaik yang akan digunakan pada permodelan klasifikasi, pada percobaan ini dicoba menggunakan beberapa iterasi karena sifat PSO yang mempelajari berulang ulang maka dipilih beberapa kali jumlah iterasi, seperti 10 iterasi, 100 iterasi, 500 iterasi, dan 1000 iterasi, yang ditunjukan hasilnya pada Tabel 7. kemudian dipilih iterasi dengan fitur yang paling optimal.

Tabel 7. Hasil seleksi fitur *chi square*

Iterasi PSO	Fitur
10	1026
100	954
500	962
1000	933

SMOTE (*Synthetic Minority Oversampling Technique*). Metode SMOTE (Synthetic Minority Oversampling Technique) digunakan dengan library *imblearn,* awalnya menentukan nilai **k** neighbour yaitu sebesar 3 untuk kelas minor, setelah itu menghitung dengan ***l*** = N/100 dimana N merupakan jumlah total data kelas minor yang berjumlah 354. Dan didapatkan nilai ***l*** menggunakan rumus tersebut sejumlah 4 yang merupakan pembulatan dari 3.54. Yang mana menandakan 4 data yang berasal dari 3 neighbour akan membentuk data-data acak yang sebanding dengan positif minor. Karena ***l*** > **k** maka akan menggunakan neighbour yang sama dengan terus berulang hingga data berimbang.

Tabel 8. Data Sampel Dokumen Baru

	dukung	vaksin	nusantara	moga	pakai	masyarakat	khusus	indonesia	Label
D_1	0	0.352	0.352	0	0	0	0	0	Negatif
D_3	0.477	0.176	0.176	0.477	0.477	0.477	0.477	0.477	Positif
D1’	0.0954	0.316	0.316	0.0954	0.0954	0.0954	0.0954	0.0954	Negatif

Contoh data baru yang ditambahkan dapat dilihat pada Tabel 8, Setelah melakukan *oversampling*, data positif dan negatif sekarang memiliki jumlah yang seimbang, dengan jumlah 645 data kelas positif, dan 645 data kelas negatif

4.1 Klasifikasi Naïve Bayes

Tahapan terakhir setelah kita mendapatkan bobot dari setiap kata, dan melakukan sampling dengan SMOTE, berikutnya memprediksi label yang akan diberikan kepada data uji menggunakan metode klasifikasi. Setelah menggunakan SMOTE maka data yang digunakan sudah terbagi secara merata menjadi 645 kelas positif dan 645 kelas negatif, dan kemudian akan terbagi menjadi 80:20.

Proses Latih. Untuk menggambarkan proses pelatihan pada pembuatan model klasifikasi ini, maka akan digunakan sampel data latih yang akan diolah dan mensimulasikan proses latih, Dari data latih sampel pada tabel, maka sudah dilakukan pembobotan pada setiap kata-kata dari hasil pembobotan TF-IDF

Tabel 9. Dala Latih Sampel

Data Komentar	
D1	bahaya bawa tni vaksin nusantara konfirmasi kasad organisasi tni libat vaksin nusantara
D2	mohon info mana beli ya percaya dokter terawan serta jajar percaya bpom
D3	dukung vaksin nusantara moga pakai masyarakat khusus indonesia

Pada penelitian ini, jika ada kata yang tidak terdapat pada salah satu kelas maka secara otomatis akan memiliki nilai 0, maka digunakan *Laplace Smoothing* pada setiap persamaan untuk setiap kelas untuk menghindari hal tersebut, dengan rumus sebagai berikut :

$$P(W_i|p) = \frac{count(Wij,p) + 1}{|p| + |V| * 1}$$

Keterangan : (jumlah nilai kata) V adalah 28, dan (jumlah nilai) TF di kelas positif adalah 20, dan (jumlah nilai) TF di kelas negative adalah 12

Selanjutnya menghitung probabilitas untuk masing-masing kelas menggunakan rumus diatas :

$$P(Positif) = 2 / 3 = 0{,}67$$
$$P(Negatif) = 1/3 = 0{,}33$$

Selanjutnya menghitung probabilitas untuk masing-masing kelas menggunakan rumus diatas, hingga didapat hasil probabilitas pada tabel 10 berikut :

Tabel 9. Probabilitas Data Latih Sampel

Term	D	P Positif	P negatif
bahaya	0,477	0,0208	0,0369
bawa	0,477	0,0208	0,0369
tni	0,477	0,0208	0,0369
vaksin	0,352	0,0245	0,0338
nusantara	0,352	0,0245	0,0338
konfirmasi	0,477	0,0208	0,0369
kasad	0,477	0,0208	0,0369
organisasi	0,477	0,0208	0,0369
tni	0,477	0,0208	0,0369
libat	0,477	0,0208	0,0369

Proses Uji. Selanjutnya, mengambil data sampel untuk dilakukan penghitungan proses uji menggunakan model pelatihan yang sudah dibuat

Tabel 10. Dala Uji Sampel

Data Sampel Uji
pribumi dukung vaksin nusantara

Tabel diatas merupakan data sampel yang akan digunakan menjadi data uji yang akan dicari probabilitasnya dengan algoritma *naïve bayes* menggunakan nilai kemungkinan (*probability*) dari data latih yang sudah diolah sebelumnya dan didapatkan nilai bobotnya, tetapi setelah itu pada data sampel uji disertakan juga penghitungan nilai bobot (TF-IDF), Untuk menenetukan dimana kelas yang memiliki probabilitas paling besar terhadap data sampel uji maka digunakan rumus perhitungan persamaan yang tepat untuk menentukan kelas dari kata yang berada pada data latih dan uji.

$P(Positif|data\ uji) = P(Positif) * P(\text{pribumi}|Positif) * P(\text{dukung}|Positif) * P(\text{vaksin}|Positif) * P(\text{nusantara}|Positif)\ P(Positif|data\ uji) = 0{,}67 * 0{,}0208 * 0{,}0307 * 0{,}0245 * 0{,}0245 =$ **0,0000002787535**

$P(\text{Negatif}|data\ uji) = P(\text{Negatif}) * P(\text{pribumi}|\ \text{Negatif}) * P(\text{dukung}|\ \text{Negatif}) * P(\text{vaksin}|\ \text{Negatif}) * P(\text{nusantara}|\ \text{Negatif})\ P(\text{Negatif}\ |data\ uji) = 0{,}33 * 0{,}0250 * 0{,}0250 * 0{,}0294 * 0{,}0294 =$ **0,000000195941695**

Dari hasil nilai probabilitas yang didapat, dapat dilihat bahwa nilai probabilitas dari kelas positif lebih besar dibanding probabilitas yang didapat dari kelas negatif, karena Probabilitas Positif > Probabilitas Negatif, maka data uji ditetapkan sebagai sebuah kelas positif.

4.3 Evaluasi

Selanjutnya, karena model *machine learning* menggunakan *multinomial naïve bayes* sudah terbentuk, maka dapat dilakukan pemeriksaan untuk melihat performa klasifikasi, dimana pada penelitian kali ini menggunakan *confusion matrix*.

Tabel 11. Tabel Evaluasi *Confusion Matrix*

	Prediksi	
	Negatif	Positif
Aktual Negatif	(TN) 98	(FP) 31
Aktual Positif	(FN) 14	(TP) 115

Analisis Perbandingan. Pada penelitian ini, sebelum menggunakan *naïve bayes* + PSO dengan *oversampling* SMOTE, turut dilakukan percobaan dengan menggunakan *Imbalanced Data*, dan juga *Undersampling*, data percobaan yang dilakukan dapat dilihat pada tabel berikut

Tabel 12. Tabel Analisis Perbandingan

	Percobaan 1 NB+PSO *Imbalanced Data*	**Percobaan 2** NB+PSO *Random Undersampling*	**Percobaan 3** NB+PSO SMOTE
Akurasi	79%	77,4%	82,5%
Presisi	75,4%	70,5%	78,7%
Recall	100%	94,3%	89,1%
Akurasi Positif	100%	94,3%	89,1%
Akurasi Negatif	40%	60,5%	76%

Dapat dilihat dari Tabel 12, setelah dilakukan 3 percobaan penelitian, maka dapat dilihat bahwa sebelum dilakukan sampling dimana data komentarnya tidak berimbang (*imbalanced data*) mendapatkan hasil akurasi sebesar 79%, presisi 75,4%, dan *recall* 100%. Hasil akurasi positif dan akurasi negatif memiliki perbedaan yang sangat jauh dengan 100% untuk akurasi positif, dan 40% untuk akurasi negatif, lalu ketika dilakukan percobaan ke 2 dengan metode *undersampling* didapat akurasi sebesar 77,4%, lalu presisi 70,5%, dan *recall* 94,3%, dan akurasi positif dan negatif masing-masing sebear 94,3% dan 60,5%, dan setelah menggunakan metode *oversampling* SMOTE maka didapatkan akurasi sebaesar 82,5%, presisi sebesar 78,7%, dan *recall* sebesar 89,1%, lalu dengan nilai akurasi positif dan negatifnya masing-masing sebememiliki nilai yang paling berimbang

5 Kesimpulan

Berdasarkan hasil yang didapat pada analisis dan pembahasan yang sudah dibahas di bagian sebelumnya bisa diambil kesimpulan sebagai berikut :

1. Pada pembangunan model klasifikasi sentimen terhadap Vaksin Nusantara ini, proses yang dilakukan adalah mengambil data komentar dari video "Peneliti Utama Jawab Kontroversi Vaksin Nusantara - ROSI (1)" yang ada di media sosial *youtube*, kemudian melakukan labeling dengan tiga orang *annotator*, kemudian melakukan praproses pada data dengan *cleaning*, *case folding*, *tokenization*, *normalization*, *stopword removal*, dan *stemming*, kemudian melakukan pembobotan data menggunakan TF-IDF, lalu menerapkan seleksi fitur PSO (*Particle Swarm Optimization*), dan setelah itu melakukan sampling dengan 3 percobaan, dengan *imbalanced data*, *random undersampling*, dan SMOTE *oversampling*, lalu data dibagi menjadi data latih dan data uji yang kemudian mengklasifikasikan kelas menggunakan algoritma *naïve bayes*, dimana jika peluang sebuah data atau dokumen memiliki peluang kelas positif lebih besar daripada kelas negatif, maka dokumen tersebut diklasifikasikan ke dalam kelas positif, ataupun berlaku juga sebaliknya.
2. Performa algoritma *naïve bayes* dan PSO (*Particle Swarm Optimization*) dengan iterasi ke 1000 yang mendapat fitur optimal sebanyak 933 fitur mendapatkan hasil terbaik akurasi sebesar 82,5%, nilai presisi 78,7%, dan nilai *recall* 89,1% yang didapat dengan menambahkan SMOTE *oversampling*.
3. Metode sampling yang diuji disini menyatakan informasi bahwa metode *oversampling* menggunakan SMOTE menghasilkan akurasi yang lebih baik dibandingkan menggunakan *random undersampling*, maupun dibandingkan *imbalanced data*, dimana jika dengan *undersampling* menghasilkan nilai akurasi 77,4%, `nilai presisi sebesar 70,5%, dan nilai *recall* sebesar 94,3%, sementara ketika menggunakan SMOTE didapatkan nilai akurasi 82,5%, nilai presisi 78,7%, dan nilai r*ecall* 89,1%, akurasi sebesar 82,5%, nilai presisi 78,7%, dan nilai *recall* 89,1% yang didapat dengan menambahkan SMOTE *oversampling*.

6 Daftar Pustaka

[1] Drozd, B., Couvillon, E., & Suarez, A. (2018). Medical YouTube Videos and Methods of Evaluation: Literature Review. *JMIR Med Educ 2018;4(1):e3*.

[2] Miranda, L. J. (2018). PySwarms: a research toolkit for Particle Swarm. *Journal of Open Source Software*, 433.

[3] W. Satriaji, & R Kusumaningrum. (2018). Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis. *2nd International Conference on Informatics and Computational Sciences (ICICoS)*.

[4] Siringoringo, R. (2018). KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR. *Jurnal ISD Vol.3*, 44-49.

[5] Huda, A. S., Awangga, R. M., & Fathonah, R. N. (2020). *Prediksi Penerimaan Pegawai Baru Dengan Metode Naïve Bayes*.

[6] Afrianti, E., Fathoni, & Heroza, R. I. (2020). Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Keterangan Laporan dan Durasi Recovery Time Laporan Gangguan Listrik PT. PLN (Persero) WS2JB Area Palembang. *JSI : Jurnal Sistem Informasi (E-Journal)*, 1955-1961.

[7] Deolika, A., Kusrini, & Luthfi, E. T. (2019). ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING. *(Jurnal Teknologi Informasi) Vol.3*, *No.2*, 179-184.