KRESNA: Jurnal Riset dan Pengabdian Masyarakat
Volume 1, Nomor 1, November, 2021, Hal: 34-43 ISSN: 2809-6509 (Online)

Implementasi Algoritma *Naïve Bayes Clasifier* untuk Mengelompokkan Naskah Berita Pendidikan dan berita Covid-19

Implementation of the Naïve Bayes Classifier to Classify Education News and Covid-19 News

Muhammad Azhar Mujahid[1], Windarto[2*], Mohammad Syafrullah[3]

[1,2,3] Fakultas Teknologi Informasi, Universitas Budi Luhur
E-mail: [1]1711503290@student.budiluhur.ac.id, [2*]windarto@budiluhur.ac.id, [3]mohammad.syafrullah@budiluhur.ac.id
(* corresponding author)

Abstract

As the day progressed, many information distribution institutions that initially conveyed news through printed media or electronic media, such as newspapers and television, turned into digital media such as news portals that use the internet. Generally, the news portal categorized news into several categories, such as education news, health news, and other categories. Yet, some people still do it manually when grouping that news into proper categories. For instance, they invite experts to assist news labeling into such categories. Compared to the manual method, news grouping can be classified using an algorithm that can automatically classify it into its exact categories based on its content, either from its title or its scripts. Naïve Bayes is one of the algorithms for grouping news based on its text. Before classifying based on its text, it is necessary to collect the news data from both of its news scripts along with its title. This study uses data taken from several online media sources through Google Alerts. It generates a search engine based on the selected criteria. Based on the testing results, using 295 news scripts, 236 training data, and 59 testing data, based on the testing results, the study is obtained 74.58% of accuracy.

Keywords : *naïve bayes, classifying, grouping, news, texts.*

Abstrak

Seiring dengan perkembangan jaman, banyak lembaga penyaluran informasi yang pada awalnya menyampaikan berita melalui media cetak atau media elektronik, seperti koran dan televisi, beralih ke media digital berupa portal berita digital yang menggunakan jaringan internet. Pada umumnya berita yang disampaikan dalam portal berita tersebut terdiri dari beberapa kategori, seperti berita tentang pendidikan, berita kesehatan kesehatan, maupun berita dengan kategori lainnya. Namun, dalam membagi berita ke dalam kategori kategori tersebut, masih ada yang melakukan secara manual yakni dengan mengumpulkan beberapa narasumber untuk menyepakati sebuah berita masuk ke kategori yang mana. Dibandingkan dengan menggunakan cara manual, pengelompokan berita dapat dilakukan secara otomatis dengan menggunakan sebuah algoritma yang dapat mengelompokkan berita berdasarkan teks berita tersebut, baik dari teks judul ataupun teks isi berita. Naïve Bayes adalah salah satu algoritma yang dapat diimplementasikan untuk mengelompokkan berita berdasarkan teksnya. Sebelum melakukan pengklasifikasian berita berdasarkan teks, perlu dilakukan pengumpulan data berupa naskah berita beserta dengan judulnya. Dalam penelitian ini, data yang digunakan diambil dari beberapa sumber media online melalui layanan Google Alerts yang menghasilkan mesin telusur berdasarkan kriteria yang dipilih. Berdasarkan hasil pengujian dengan menggunakan data sebanyak 295 buah naskah berita, 236 buah data training, dan 59 buah data testing, didapatkan hasil akurasi sebesar 74.58%.

Kata kunci : naïve bayes, klasifikasi, pengelompokan, berita, teks

1. PENDAHULUAN

Berita adalah informasi berdasarkan fakta atau laporan mengenai suatu kejadian yang sedang atau telah terjadi dan dipublikasikan melalui media cetak, siaran, internet maupun dari mulut ke mulut [1]. Dengan adanya berita, masyarakat menjadi lebih tahu mengenai kejadian terkini. Di masa pandemic covid-19 ini, berita vaksin covid-19 dan berita tentang pembelajaran dalam jaringan merupakan salah satu berita yang menjadi topik utama berita saat ini. Dalam mengelompokkan berita ke dalam kategori berdasarkan teks judul berita ataupun isi dari sebuah naskah berita, saat ini masih ada yang melakukan secara manual yakni dengan mengumpulkan beberapa narasumber untuk menyepakati sebuah berita masuk ke kategori yang mana.

Dibandingkan dengan menggunakan cara manual, pengelompokan berita dapat dilakukan secara otomatis dengan menggunakan sebuah algoritma yang dapat mengelompokkan berita berdasarkan teks berita tersebut, baik dari teks judul ataupun teks isi berita. Naïve Bayes adalah salah satu algoritma yang dapat diimplementasikan untuk mengelompokkan berita berdasarkan teksnya. Sebelum melakukan pengklasifikasian berita berdasarkan teks, perlu dilakukan pengumpulan data berupa naskah berita beserta dengan judulnya. Dalam penelitian ini, data diambil dari beberapa sumber media online melalui layanan Google yang yang menghasilkan mesin telusur berdasarkan kriteria yang dipilih, yaitu Google Alerts.

Dalam mengelompokkan sebuah teks, terdapat dua pekerjaan utama, pertama adalah pembangunan model sebagai sebuah *prototype* untuk disimpan sebagai memori dan yang kedua adalah penggunaan model untuk melakukan sebuah pengenalan klasifikasi serta prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang mudah disimpan (*testing data*).

Algoritma pengelompokan yang digunakan pada penelitian ini adalah Naïve Bayes Classifier. Menurut Friedman, klasifikasi atau pengelompokan adalah model probabilistik yang sederhana untuk mengelompokkan data ke dalam kelas yang spesifik berdasarkan fitur data yang berbeda [1]. Sebelum berita dapat dikelompokkan, perlu dilakukan tahapan *preprocessing*. *Preprocessing* itu sendiri terdiri dari beberapa tahapan, yaitu: *case folding*, *tokenizing*, *filtering* dan *stemming*. Proses *stemming* merupakan proses yang terpenting karena pada prose *stemming* terjadi penghilangan imbuhan pada kata di dalam sebuah kalimat pada berita sehingga menghasilkan kata dasar.

Berdasarkan latar belakang tersebut, dapat dirumuskan pertanyaan riset dalam penelitian ini yaitu seberapa besar tingkat ketepatan metode Naïve Bayes Classifier dalam mengelompokkan berita online ke dalam kategori "Kesehatan" dan "Pendidikan" dari Google Alerts pada periode waktu tertentu. Penelitian ini dibatasi oleh beberapa hal diantaranya adalah kategori yang digunakan dalam pengelompokan naskah berita yaitu kesehatan dan pendidikan, pengambilan data maksimal dalam sekali pengambilan data dari Google Alerts berjumlah 20 data naskah berita, berita yang digunakan sebagai data sampling dan data training diambil pada kurun 03 Desember 2020 hingga 10 Desember 2020. Tujuan akhir dari penelitian ini adalah mengetahui hasil ketepatan klasifikasi naskah berita dengan menggunakan metode Naïve Bayes Classifier dalam dua kategori yang telah disebutkan sebelumnya.

Berikut ini adalah beberapa penelitian yang telah dilakukan sebelumnya yang berkaitan dengan pengelompokan berita yang digunakan sebagai referensi dalam penelitian ini:
Penelitian yang dilakukan oleh [2] dengan judul "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting" menjabarkan bahwa penelitian yang dilakukan merupakan bagian dari *text mining* untuk klasifikasi konten berita yang telah memiliki label berdasarkan katagori berita pada situs detik.com. Proses yang dilakukan adalah melakukan permodelan dan pengolahan data, mulai proses *pre-processing*, proses seleksi fitur *information gain*, dan penerapan model algoritma *Naive Bayes Classifier* dengan *Bayesian Boosting*. Hasil yang diperoleh atas model tersebut mendapatkan nilai evaluasi terhadap

akurasi, *recall*, dan ketepatan sebesar 73.2%. Sedangkan dengan model yang lebih ringkas yaitu model algoritma *Naive Bayes Classifier*, dengan *Bayesian Boosting* mendapatkan nilai evaluasi yang sama besar yaitu 73.2%. Penilaian atas hasil evaluasi model yang telah terlaksanakan berkesimpulan bahwa penerapan seleksi fitur *Information Gain* tidak berpengaruh besar atas kenaikan hasil performa terhadap kondisi label *Polynomial*.

Penelitian yang dilakukan oleh [3] dengan judul "Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia". Teknologi informasi khususnya internet sangat mendukung terjadinya pertukaran informasi dengan sangat cepat. Kondisi tersebut memunculkan masalah untuk mengakses informasi yang diinginkan secara akurat dan cepat. Untuk mengatasi masalah tersebut, salah satu Teknik yang dapat digunakan adalah dengan mengklasifikasikan teks dokumen tersebut sesuai dengan karakteristik, fitur, maupun kelasnya berdasarkan aturan baku bahasa yang akan diolah. Dalam penelitian ini Bahasa Indonesia adalah bahasa yang digunakan sebagai sumber acuan. Jenis penelitian ini termasuk kepada penelitian terapan (*Applied Research*). Objek dalam penelitian ini adalah dokumen Teks Berbahasa Indonesia. *Vector Space Model* (VSM) adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu *query*. Pada model ini, *query* dan dokumen dianggap sebagai vektor-vektor pada ruang n-dimensi, dimana n adalah jumlah dari seluruh *term* yang ada di dalam daftar. Tujuan dari penelitian ini menganalisis efektifitas model sistem klasifikasi/kategorisasi dokumen dalam penerapan *Vector Space Model* berdasarkan pembobotan termdokumen dan query, juga menerapkan metode stemming Bahasa Indonesia dengan algoritma nazief adriani, menghasilkan nilai similarity dengan fungsi *cosine* yang berpengaruh pada pemeringkatan hasil kategorisasi dokumen yang relevan.

Dalam penelitian yang dilakukan oleh [4] dengan judul "Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram" menjabarkan bahwa perkembangan berita digital telah tumbuh sangat cepat. Saat ini diperkirakan 80% berita digital dalam bentuk tidak terstruktur. Tingginya volume dokumenteks ini dipicu oleh aktivitas dari berbagai sumber berita. Kebutuhan analisis *text mining* sangat diperlukan dalam menangani teks yang tidak terstruktur tersebut. Untuk mengklasifikasikan berita, banyak peneliti yang berusaha untuk melakukan klasifikasi terhadap berita ini secara otomatis, salah satunya adalah dengan menggunakan klasifikasi naïve bayes. Pada penelitian ini selain menggunakan naïve bayes, peneliti juga akan menggunakan fitur N-Gram. Diharapkan dengan penambahan metode ini, dapat meningkatkan tingkat akurasi dari klasifikasi naïve bayes.

Sementara itu dalam penelitian yang dilakukan oleh [1] dengan judul "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes Dengan Enhanced Confix Stripping Stemmer" diungkan bahwa berita dapat dikelompokkan secara manual oleh manusia, akan tetapi hal tersebut membutuhkan waktu yang lama untuk melakukan kategorisasi. Saat ini dalam waktu yang tergolong singkat dokumen berita olahraga dalam bentuk web memiliki jumlah yang sangat besar. Untuk kemudahan akses dokumen perlu melakukan pengelompokan dokumen berita kedalam beberapa kategori. Hal tersebut bertujuan agar berita olahraga tersusun sesuai dengan kategori yang ditentukan. Metode klasifikasi diusulkan dalam penelitian ini untuk melakukan pengkategorian secara otomatis dokumen berita. Tujuan dilakukannya klasifikasi adalah untuk mempercepat dan mempermudah dalam pemberian kategori, sehingga dapat meningkatkan efisiensi waktu. Pada penelitian ini digunakan metode klasifikasi Naïve Bayes Classifier. Sebelum dilakukan klasifikasi, proses *preprocessing* dilakukan dengan menggunakan *Enhanced Confix Striping Stemmer*. Hal ini bertujuan untuk mengembalikan sebuah kata ke bentuk dasarnya, sehingga data berkurang dan proses komputasi menjadi lebih efisien. Pengujian dilakukan menggunakan 18 berita olahraga yang dipilih secara acak oleh *user* atau *tester*, dari 18 berita yang diujikan terdapat 14 berita yang bernilai benar atau relevan dengan analisis yang dilakukan *user* atau tester pada berita uji. Dari penelitian ini dapat

disimpulkan bahwa Aplikasi Klasifikasi Berita Olahraga menggunakan Metode Naïve Bayes dengan Enhanced Confix Striping Stemmer mampu mengklasifikasi berita olahraga sesuai dengan kategori masing-masing, seperti Sepak Bola, Basket, Raket, Formula 1, Moto GP dan olahraga lainnya dengan keakuratan sebesar 77%.

2. TINJAUAN PUSTAKA

Penelitian dengan topik pengelompokkan berita yang dilakukan ini merupakan bagian dari *text mining* untuk klasifikasi konten berita vaksin covid-19 dan konten berita pembelajaran tatap muka yang telah memiliki label berdasarkan kategori berita dan kata kunci yaitu "Kesehatan" dan "Pendidikan". Data mining merupakan suatu langkah dalam *Knowledge Discovery in Databases* (KDD). *Knowledge discovery* sebagai suatu proses terdiri atas pembersihan data (*data cleaning*), integrasi data (*data integration*), pemilihan data (*data selection*), transformasi data (*data transformation*), data mining, evaluasi pola (*pattern evaluation*) dan penyajian pengetahuan (*knowledge presentation*). Data mining mengacu pada proses untuk menambang (*mining*) pengetahuan dari sekumpulan data yang sangat besar [5].

Jika pada *data mining* data yang diekstrasi adalah berupa dokumen, maka *text mining* adalah cara yang digunakan untuk ekstraksi informasiyang lebih berkualitas dari dataset yang tersedia. Penelitian ini mengusulkan metode klasifikasi dengan algoritma Naïve Bayes Classification (NBC) [4]. Banyak metode *data mining* yang dapat diterapkan untuk klasifikasi. Algoritma klasifikasi yang populer adalah *Decision Trees, Neural Networks, K-Nearest Neighbours, Naive Bayes,* dan algoritma Genetik [6].

Dalam ilmu *data mining*, terdapat 2 data yang digunakan. Yaitu *data training* dan *data testing*. *Data training* atau data latih adalah data yang digunakan sebagai acuan atau model sebelum melakukan proses uji atau testing. Proses melatih data ini dilakukan secara manual dengan melalui tahapan-tahapan yang terdapat pada *preprocessing*. Setiap kata pada dokumen teks akan melewati tahap pra proses untuk kemudian dibersihkan menjadi data yang bersih (terstruktur) untuk melihat apakah di dalam dokumen teks tersebut terdapat kata kunci atau kata lainnya yang berkaitan dengan kategori yang telah ditentukan sebelumnya. Kemudian dilakukan proses perhitungan dengan menggunakan algoritma *naïve bayes* untuk mendapatkan nilai probabilitas pada setiap kata atau kalimatpada dokumen yang ada sehingga data dapat dikelompokkan ke dalam kategori yang telah ditentukan.

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decesion tree dan neural network [7]. *Naïve Bayes Classifier* atau disebut juga dengan *Bayesian Classification* merupakan metode pengklasifikasian statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas.

Pada algoritma *Naïve Bayes* setiap dokumen dipresentasikan dengan masukan *"a1, a2, a3,…,an"* dimana *a1* adalah kata pertama dan berikutnya sampa a_n (kata ke-n), sedangkan V yaitu label kategori. Selanjutnya yaitu mencari nilai tertinggi dari kategoriteks yang diujikan (VMAP). Persamaan VMAP yaitu sebagai berikut:

$$V_{MAP} = \frac{argmax}{v_j \in V} \mathrm{P(v_j)} \prod_i P(a_i \mid v_j)$$

Nilai P(vj) dihitung pada saat data latih, dengan rumus sebagai berikut:

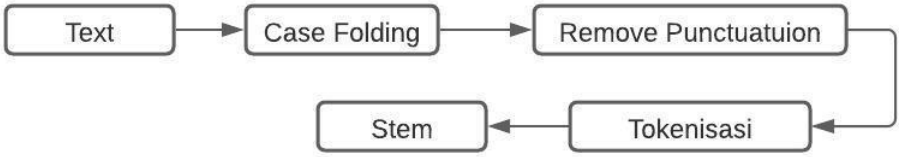
$$P(v_j) = \frac{|docs\ j|}{|training|}$$

Dimana:
|docs j| adalah dokumen yang memiliki kategori j pada dokumen latih.
|training| adalah jumlah dokumen latih.

$$P(a_i \mid v_j) = \frac{|n_i + 1|}{|n + kosakata|}$$

Dimana:
1. n_i adalah jumlah kemunculan kata a_i pada dokumen yang berkategori v_j
2. n adalah jumlah seluruh kata pada dokumen yang berkategori v_j
3. kosakata adalah jumlah kata pada seluruh dokumen latih

Setelah *dataset* terkumpul, maka selanjutnya adalah proses untuk memulai pengolahan data, yaitu proses *preprocessing* [2]. Tujuan dilakukannya *preprocessing* berita adalah untuk menghilangkan *noise*, menyeragamkan bentuk kata dan mengurangi volume kata. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Pada tahap *preprocessing* terdiri dari proses *Case Folding, Tokenizing, Stopword Removal, Stemming*.



Gambar 1: Tahap Preprocessing

Case folding proses *case folding* untuk menyeragamkan bentuk huruf menjadi huruf kecil. Hal ini dilakukan untuk mempermudah pencarian, tidak semua dokumen teks konsisten dalam penggunaan huruf kapital.
Contoh:
- Kalimat asli: Vaksin Covid telah tersedia pada bulan Januari
- Setelah Case Folding: vaksin covid telah tersedia pada bulan januari

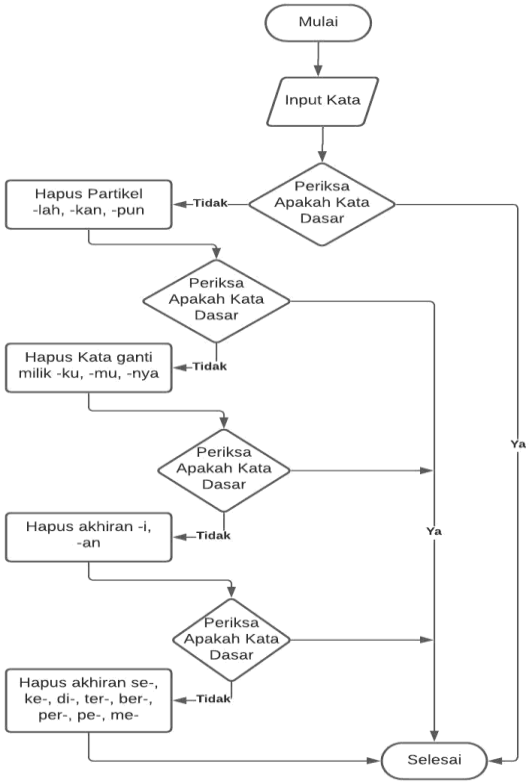
Tokenisasi pada proses tokenisasi ini, semua kata yang ada di dalam tiap dokumen dipisahkan dan dihilangkan tanda bacanya, serta dihilangkan jika terdapat simbol atau apapun yang bukan huruf.
Contoh:
- Kalimat asli: vaksin covid telah tersedia pada bulan januari
- Setelah Case Folding: vaksin, covid, telah, tersedia, pada, bulan, januari.

Stopword removal Pada tahap ini, kata-kata yang tidak relevan akan dihapus, kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dengan sentimen.
Contoh yang sering digunakan, yaitu:
- Dan, atau, maka, di, ke, dari, walaupun, meskipun, yang, ini, itu, disini.

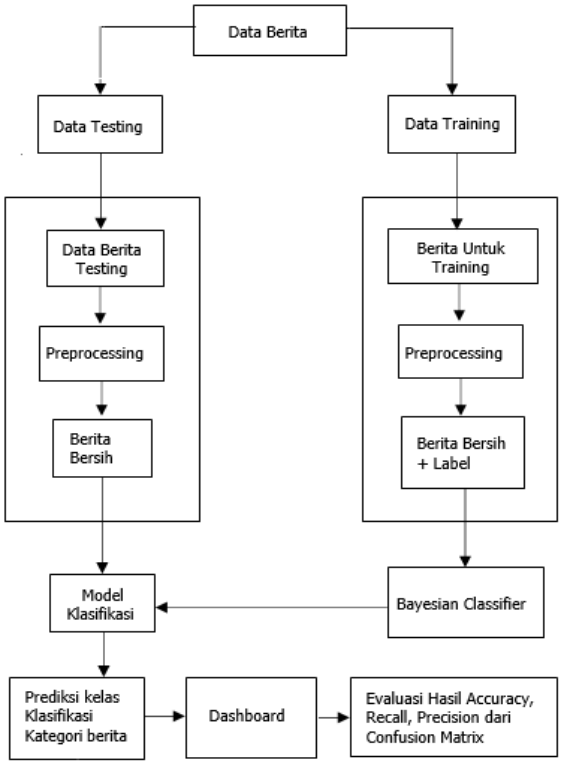
Stemming merupakan proses untuk mendapatkan kata dasar dari kata asli pada suatukalimat. Kata asli dapat mengandung imbuhan yang dipisahkan berdasarkan aturan tertentu, misalnya

kata makanan, dimakan, memakan yang memiliki satu kata dasaryang sama, yaitu makan. Dalam bahasa Indonesia, salah satu algoritme stemming dikembangkan oleh Nazief dan Adriani. Algoritme ini menggunakan kamus kata dasar sebagai acuan dalam proses stem kata. Jika semakin banyak kata dasar yang tersedia dalam kamus, maka algoritme ini akan semakin akurat [8].

- Kalimat asli: vaksin, covid, telah, tersedia, pada, bulan, januari.
- Setelah *stemming*: vaksin, covid, telah, sedia, pada, bulan, Januari



Gambar 2: Alur *Stemming*



Gambar 3: Alur Proses Klasifikasi

3. METODE PENELITIAN

Pada Penelitian ini yang akan diambil adalah objek pada kasus vaksin corona dan pembelajaran tatap muka di tengah pandemi ini, dengan mengambil data berita dengan menggunakan *Google Alerts* pada tanggal 03 Desember 2020 sampai tanggal 10 Desember 2020. Gambaran umum alur proses klasifikasi dijelaskan pada gambar 3.

Tabel 1: Data Berita *Google Alerts*

Keyword	Text	Sumber
Vaksin Covid	Vaksin Covid telahtersedia pada bulanJanuari 2020	Liputan6.com
Pembelajaran Tatap Muka	Saat ini Seluruh Siswa dan Siswi Se jabotabek akan melakukan Daring atau Pembelajaran tatap muka pada awal December	Kompas.com

Sumber Data

Pengumpulan data berita menggunakan fitur *Feed RSS* yang disediakan oleh*Google Alerts* berupa *file XML*. Data berita yang digunakan berasal dari berbagai sumbermedia berita online melalui *Google Alerts* yang memberitakan berita-berita terkait“Kesehatan” dan “Pendidikan”.

Berita-berita yang sudah terkumpul akan dilakukan proses *training* atau data latih secara manual dengan mengumpulkan beberapa narasumber agar melakukan voting pada data training dan kata-kata kunci yang sebelumnya ditentukan topik apa saja yang akan diangkat ke dalam penelitian terlebih dahulu.

Pada penelitian ini, berita yang dikumpulkan adalah berita dari hasil pencarian katakunci Vaksin Covid atau Pembelajaran Tatap Muka dari *Engine Google Alerts*. Pengambilan data berita yang akan digunakan dalam penelitian berasal dari berbagai macam sumber berita serta web yang diambil secara acak seperti indozone, iNews, Liputan6.com, Kompas.com, tribunnews.com, kumparan.com dan lain-lain yang diambilsecara *realtime.* Pengumpulan data dari *Google Alerts* ini dilakukan terhitung mulai tanggal 03 Desember 2020 hingga tanggal 10 Desember 2020 dan berhasil mendapatkan data sebanyak 295 *data* berita. Pada Tabel 3.1 terdapat *profiling* data berita yang bersumber dari web portal berita dan non portal berita dalam hasil 10 besar pada tabel
3.1 terdapat *profiling* data berita yang bersumber dari web portal berita dan non portal berita dalam 10 besar hasil pengumpulan *data*.

4. DISKUSI

4.1. Data Training

Penelitian ini menggunakan data berupa berita khusus berbahasa Indonesia yang terdapat dalam *Google Alerts* dengan topik Kesehatan dan Pendidikan. Data berita yang diambil berjumlah 295 data dan mengambil 236 total data sampel dengan pembagian 80:20. Penulis membuat database *mySQL* yang bernama kkp dan untuk menyimpan data tersebut, yang nantinya akan dipakai untuk proses *Stemming, Case Folding, Remove Punctuation* dan *Tokenizing*. Di bawah ini adalah tabel contoh *data training* yang tertera pada Tabel 2.

Tabel 2: Contoh Data Training

Sumber	Sumber	Text	Kategori	Kata Kunci
Google Alerts	Liputan6.com	Vaksin Covidtelah tersedia pada bulan Januari 2020	Kesehatan	Vaksin Covid
Google Alerts	Tribunnews.com	Saat ini Seluruh Siswa dan Siswi Se jabotabek akanmelakukan Daring atau Pembelajaran tatap muka padaawal December	Pendidikan	Pembelajaran Tatap Muka
Google Alerts	Kompas.com	Debat dua qosim alif terpapar visi misi dan programbangun lanjut di seluruh pelosok untuk pendidikan	Pendidikan	Pembelajaran Tatap Muka

4.2. Data Testing

Penelitian ini menggunakan data berupa berita khusus berbahasa Indonesia yang terdapat dalam Google Alerts dengan topik Kesehatan dan Pendidikan. Data berita yang diambil berjumlah 295 data dan mengambil 59 total *data testing* dengan pembagian 80:20. Penulis membuat database *mySQL* yang bernama kkp dan untuk menyimpan data tersebut, yang nantinya akan dipakai untuk proses *Stemming, Case Folding, Remove Punctuation* dan *Tokenizing*. Di bawah ini adalah tabel contoh *data training* yang terterapada Tabel 3.

Tabel 3: Contoh Data Testing

Sumber	Sumber	Text	Kategori	Nilai Probabilitas
Google Alerts	Liputan6.com	Vaksin covid percobaab di kabupaten penajam paser utara	Kesehatan	Kesehatan : 0.9898 Pendidikan : 0.5232

Google Alerts	Tribunnews.com	Sekolah sejabotabek hampir sebulan,siswa sdn karyabakti tetapsemangat belajar meski belajar tatap muka	Pendidikan	Kesehatan : 0.3203 Pendidikan : 0.8994
Google Alerts	Kompas.com	debat dua qosim alif terpapar visi misi dan program bangun lanjut di seluruh pelosok untuk pendidikan	Pendidikan	Kesehatan : 0.5349 Pendidikan : 0.5232

Proses *Data Training* dilakukan secara otomatis untuk mengetahui seberapa akuratsistem dalam mengklasifikasikan objek tertentu. Proses tersebut dilakukan dengan cara sistem membaca data baru yang digunakan sebagai *data training* yang terdapat pada Tabel 4 lalu memprediksi pola kata dari model (*data training*) dan menghitung nilai probabilitas darisetiap kata yang ada pada berita.

Tabel 4: Text Uji Yang Sudah Diprocessing

Text	Kategori
catat ini dating jadwal vaksin corona di ri	Kesehatan
kiat ajar tatap muka bupati pati Haryanto saya tidak maumurid jadi korban covid	Pendidikan

Pada Tabel 4 terdapat data baru yang akan digunakan sebagai data testing untuk kemudian dilakukan perhitungan nilai probabilitas menggunakan algoritma *naïve bayes classifier*. Berikut perhitungan nilai probabilitasnya. Pertama mencari probabilitas setiap kategori menggunakan persamaan pada rumus berikut ini:

$$P(Kesehatan) = \frac{1}{2} = 0.5$$

$$P(Pendidikan) = \frac{1}{2} = 0.5$$

1. *P* adalah probabilitas, *Vj* adalah kategori
2. *docs* adalah masih-masing dari kategori
3. *training* adalah keseluruhan dari kategori

Tabel 5: Text Uji Yang Sudah Diprocessing Dan Distemming

Text Uji
Catat, bupati

Mencari nilai probabilitas setiap kata uji pada kesehatan menggunakan persamaan pada rumus yaitu sebagai berikut:

$$P(catat|\ kesehatan) = \frac{1+1}{8+22} = 0{,}09090$$

$$P(canggih|\ kesehatan) = \frac{0+1}{8+22} = 0{,}04545$$

1. *ni* adalah kalimat yang terdapat di kategori
2. *N* adalah jumlah keseluruhan kosakata pada kategori
3. Kosakata adalah jumlah keseluruhan kosakata pada setiap kategori

Mencari nilai probabilitas setiap kata uji pada pendidikan menggunakanpersamaan pada rumus yaitu sebagai berikut:

$$P(catat|\ pendidikan) = \frac{0+1}{14+22} = 0{,}02777$$

$$P(canggih|\ pendidikan) = \frac{1+1}{14+22} = 0{,}05555$$

Lalu setelah mengetahui hasil probabilitas pada setiap kata uji selanjutnya melakukan proses perhitungan pada Tabel 6 yaitu sebagai berikut:

Tabel 6: Proses Perhitungan *Naïve Bayes*

Kategori	P(vj) = $\frac{\|docs\ j\|}{training}$	**P(ai | vj)** = $\frac{\|n_i+1\|}{\|n+kosakata\|}$	**P(ai | vj) P(vj)**	**VMAP** = $\underset{vj \in V}{argmax}$ **P(vj)** $\prod_i$ **P(ai | vj)**
Kesehatan	0.5	$P(catat\|\ kesehatan)$ = 0,09090 $P(canggih\|\ kesehatan)$ = 0,04545	0,09090 ∗ 0,04545 ∗ 0.5 = 0,0020657025	
Pendidikan	0.5	$P(catat\|\ pendidikan)$ = 0,02777 $P(canggih\|\ pendidikan$ = 0,05555	0,02777 ∗ 0,05555 ∗ 0.5 = 0,000771311	0,00077131175

Pada Tabel 6 proses perhitungan pada *Naïve Bayes* bahwa probabilitas yang mempunyai nilai terbesar adalah kategori pendidikan dengan nilai probabilitas 0,00077131175.

4.3. Hasil Pengujian

Pada pengujian lebih lanjut untuk *data testing*, data yang digunakan yaitu sebanyak 295 data yaitu 236 data *training* dan 59 untuk data *testing*, kemudian data yang telah di *training* sebelumnya dijadikan acuan untuk mengukur akurasi pada *data testing* ini dan menghasilkan akurasi sebesar 74.58% dimana hasilnya dapat dilihat pada tabel 7 berikut.

Tabel 7: Pengujian Akurasi Data Testing

accuracy :74.58%			
	true kesehatan	true pendidikan	class precision
kesehatan	38	15	71.70%
pendidikan	0	6	100.00%
class recall	100.00%	28.57%	

Pada hasil pengujian untuk *data testing* dan *data training* yang digunakan yaitu sebanyak 295 data, kemudian data yang telah di *training* sebelumnya dijadikan acuan untuk mengukur akurasi pada *data testing* ini dan menghasilkan akurasi sebesar 74.58% yang menghasilkan perhitungan *precision* dengan berita kesehatan memiliki nilai precision 71.70%, berita pendidikan memiliki nilai *precision* 100% dan hasil dari nilai *recall* berita kesehatan memiliki nilai *recall* sebesar 100%, berita pendidikan memiliki nilai *recall* sebesar 28.57% .

5. KESIMPULAN

Berdasarkan analisis dan pengujian yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut:

1. Aplikasi ini mampu melakukan klasifikasi untuk pengkategorian beritasecara otomatis.
2. Proses klasifikasi semakin akurat jika data latih yang digunakan dalam pembelajaran berjumlah banyak, akan tetapi dapat juga mengurangi keakuratan jika kata-kata yang terdapat pada berita tersebut mengalamibias atau bermakna ganda.
3. Akurasi yang didapatkan dengan menggunakan total 295 data sampel dengan pembagian 80% *data training* dan 20% *data testing* menghasilkan akurasi sebesar 74.58%.
4. Pada penelitian ini, dokumen teks yang dapat diproses hanya teks berbahasa Indonesia.

DAFTAR PUSTAKA

[1] Y. D. Pramudita, S. S. Putro & N. Makhmud, Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes Dengan Enhanced Confix Stripping Stemmer, *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 5, no. 3, pp. 269-276, 2018.

[2] B. S. Prakoso, D. Rosiyadi & H. S. Utama, Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting, *Jurnal Resti*, vol. 3, no.2, pp.227-232, 2019.

[3] D. Susandi and U. Sholahudin, Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN danFungsi Similarity Cosine untuk Pembobotan IDF dan WIDFpada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia, *Jurnal ProTekInfo*, vol. 3, no.1, pp.22-29, 2016.

[4] D. N. Chandra, G. Irawan & I. N. Sukajaya, Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram, *Jurnal Ilmi Komputer (JIKI)* vol. 4, no. 2, pp. 10-20, 2016.

[5] D. Firdaus, Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer, *Jurnal Format*, vol. 6, no. 2, pp.91-97, 2017.

[6] A. Saifudin, Metode Data Mining Untuk Seleksi Calon Mahasiswa Pada Penerimaan Mahasiswa Baru Di Universitas Pamulang, *Jurnal Teknologi*, vol. 10, no. 1, pp. 25-36, 2017.

[7] H. Annur, Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes, *Jurnal ILKOM*, vol. 10, no.2, pp.160-165, 2018.

[8] H. K. Wardana, I. Swanita and B. W. Yohanes, Sistem Pemeriksa Pola Kalimat Bahasa Indonesiaberbasis Algoritme Left-Corner Parsing dengan Stemming, *JNTETI*, vol. 8, no.3, pp. 211-217, 2019.