JASIE "Jurnal Aplikasi Sistem Informasi dan Elektronika"
Vol. 3, No. 2, Desember 2021, Hal: 86-96
p-ISSN: 2714-612X, e-ISSN: 2798-1509

Analisis Big Data dengan Metode *Multinomial Naïve Bayes* terhadap Klasifikasi Media Pemberitaan

Asfik Alfain[1*], Hardian Oktavianto[2], Yeni Dwi Rahayu[3]
Program Studi Teknik Informatika, Universitas Muhammadiyah Jember[1,2,3]
Email: alfainasfik@gmail.com[1*], hardian@unmuhjember.ac.id[2], yenidwirahayu@unmuhjember.ac.id[3]

ABSTRAK

Media pemberitaan merupakan sebuah sebaran informasi berupa fakta yang disebarkan melalui media online, cetak, maupun siaran televisi dan radio. Saat ini media yang paling diminati adalah berbasis internet yang tersusun atas beberapa kategori berita contohnya kesehatan, seleb, news, olahraga, otomotif, travel, musik, dll. Pengklasifikasian artikel berita saat ini masih dilakukan secara manual sehingga memerlukan banyak waktu. Solusi yang diperlukan adalah sebuah sistem yang dapat mengklasifikasi artikel berita dengan otomatis. Pengklasifikasian ini menggunakan metode Multinomial *Naïve Bayes* dengan jumlah 19.200 *dataset* kemudian diukur dengan menggunakan confusion matrix dan diperoleh tingkat akurasi tertinggi sebear 74%, presisi 98%, *recall* 93%.

Kata Kunci: berita, klasifikasi, multinomial *Naïve Bayes*

ABSTRACT

News media is a distribution of information in the form of facts that are disseminated through online media, print, as well as television and radio broadcasts. Currently the most popular media is internet-based which is composed of several news categories such as health, celebrity, news, sports, automotive, travel, music, etc. Classification of news articles is still done manually so it takes a lot of time. The solution needed is a system that can classify news articles automatically. This classification uses the Multinomial Naïve Bayes method with a total of 19,200 datasets then measured using a confusion matrix and the highest accuracy rate is 74%, precision is 98%, recall is 93%.

Keywords: *news, classification, multinomial Naïve Bayes*

1. PENDAHULUAN

Pendahuluan berisi tentang latar belakang penelitian, tujuan penelitian, kontribusi /manfaat Berita (news) merupakan sebuah sajian utama dari media massa di samping views (opini) (Romli, 2018). Berdasarkan data dari perusahaan konsultan Maverick Indonesia yang melibatkan 453 responden di kawasan Jakarta dan Bandung menghasilkan beberapa klasifikasi berita yaitu pembaca dengan rentang usia 18 hingga 24 tahun tertarik pada topik bahasan edukasi yaitu sebesar 35% sedangkan rentang usia 25 hingga 34 tahun topik bahasan tertinggi adalah kesehatan (31%), keduanya memiliki topik bahasan terendah yaitu gaya hidup 4% dan 7% (Kasih, 2020). Tahun 2006 perkembangan dan pertukaran informasi telah berada di titik lebih dari 550 triliun dokumen dan 7,3 juta halaman baru pada internet setiap harinya. Sementara itu pengklasifikasian kategori pada berita masih dilakukan secara manual sehingga memerlukan waktu yang lama dan secara penataan bahasa masih sulit untuk dikategorikan (Ariadi & Fithriasari, 2015). Umumnya pada portal berita selalu dikelompokkan ke dalam beberapa kategori tertentu (Hand & Yu, 2001). Namun, pengelompokan tersebut masih dilakukan secara manual, artinya pengelompokan tersebut perlu membaca isi berita secara keseluruhan agar pengelompokan dapat dilakukan dengan optimal. Hal tersebut dinilai kurang efektif, terlebih jika data berjumlah banyak.

Salah satu solusi dalam menyelesaikan permasalah tersebut adalah dengan menggunakan klasifikasi terhadap text mining (Romli, 2018). Text mining termasuk salah satu variasi dari data mining dari kumpulan data teks dengan jumlah besar. Selain klasifikasi, text mining menjadi solusi clustering dan information extraction (Efendi & Mustakim, 2017).

Penelitian oleh Findra Kartika dan Tri Purnomo yang mengklasifikasi Berita dengan Metode Multinomial *Naïve Bayes* menggunakan Pembobotan kata TF-IDF dengan 7 kategori dan 10.500 berita menghasilkan akurasi, presisi, *recall* dan f1-*score* sebesar 96%, peneliti menyebutkan bahwa

sampel sangat sedikit dibandingkan dengan jumlah *dataset* dan perlu model klasifikasi dan pengujian terbaru. (Dewi & Aji, 2021). Kemudian Dio Ariadi dan Kartika Fithriasari dalam penelitiannya menjelaskan klasifikasi dengan metode *Multinomial Naïve Bayes* dengan menambahkan *confix stripping stemmer* dengan 1200 berita dan 12 kategori berita menghasilkan performa akurasi, presisi dan *recall* sebesar 82,2%, 83,9%, dan 82,2% tetapi berbeda saat menggunakan *Support Vector Machine* yaitu 88,1%, 89,1%, dan 88,1%.

Berdasarkan penelitian yang sudah dijelaskan di atas, penelitian ini memiliki kesamaan pada metode yang digunakan. Namun, pada penelitian ini data yang digunakan mencapai 19.200 dengan membandingkan 17 dan 14 kategori. Metode yang digunakan adalah Algoritma *Multinomial Naïve Bayes*, dan diuji menggunakan *K-Fold Validation*. Dari perbandingan jumlah kategori tersebut, dapat diketahui hal-hal yang berpengaruh terhadap nilai akurasi, presisi, dan *recall*. Berdasarkan uraian di atas maka peneliti mengambil judul "Analisis *Machine Learning* dengan Metode *Multinomial Naïve Bayes* terhadap Klasifikasi Media Pemberitaan."

2. KAJIAN PUSTAKA

A. Media Online

Website atau site (situs) merupakan sebuah halaman yang memiliki konten (media) dalam bentuk video, audio, teks, dan gambar. *Website* bisa diakses melalui internet dan alamat internal yang dikenal dengan URL (Uniform Resource Locator) yang diawali dengan kata www atau http:// (Hypertext Transfer Protocol) (Romli, 2018).

B. Berita

Berita *(news)* merupakan sebuah sajian utama dari media massa di samping views (opini). Mencari bahan pemberitaan kemudian menyusunnya merupakan tugas pokok wartawan dan bagian redaksi sebuah penerbitan pers (media massa) (Romli, 2018).

C. *Machine Learning*

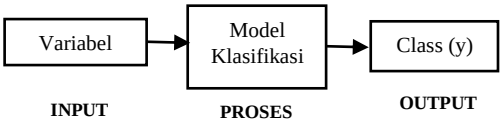
Machine Learning menjadi salah satu metode yang sesuai dalam mengklasifikasikan berita. *Machine Learning* dapat menerima hasil analisis pada pengklasifikasian, dimana jenis dan sifat data akan diolah sebelum mendapatkan hasil (Witten dkk., 2011).

D. *Data Mining*

Data mining adalah sebuah proses pengumpulan informasi penting dari data yang berjumlah besar. *Data mining* juga bisa disebut sebagai proses yang sistematik dengan tujuan menemukan nilai berupa pengetahuan yang belum bisa didapatkan secara manual dari kelompok data (Bustami, 2014).

E. Klasifikasi

Klasifikasi adalah proses dalam menemukan fungsi atau model untuk menggambarkan sebuah kelas atau konsep pada sebuah data. Proses tersebut dapat meramalkan kecenderungan di masa mendatang dengan menggunakan deskripsi data. Pemodelannya berupa aturan sebab akibat, pohon keputuran, atau formula matematis (Bustami, 2014).



Gambar 1. Tahapan klasifikasi

F. *Naïve Bayes*

Naïve Bayes adalah salah satu metode pengklasfikasian dengan menggunakan probabilitas dan statistik yang disampaikan oleh Thomas Bayes. Teorema Bayes miliknya dapat memprediksi peluang di masa mendatang dengan data kejadian di masa sebelumnya. Teori tersebut digabungkan dengan

Naïve yang mengasumsikan kondisi antar variabel bebas. Klasifikasi *Naïve Bayes* berasumsi bahwa ciri-ciri pada tiap kelas tidak ada korelasi dengan kelas lainnya (Bustami, 2014).

$$P(H|X)\frac{p\ (X|H).P(H)}{P(X)})$$

Keterangan:
X = Data dengan kelas yang belum dikenal.
H = Hipotesa data X adalah kelas spesifik.
P(H|X) = Probabilitas hipotesis H sesuai keadaan X (Posteriori probability)
P(X) = Probabilitas hipotesis H (Prior probability).
P(X|H) = Probabilitas X sesuai kondisi terhadap hipotesis H.
P(X) = Probabilitas X

G. *Multinomial Naïve Bayes*

Multinomial Naïve Bayes merupakan salah satu varian dari *Naïve Bayes*. Metode ini menyatakan seluruh atribut saling bergantung mengingat konteks pada sebuah kelas dan mengabaikan seluruh dependensi antar atribut (Saleh, 2015).

$$(C)\frac{count\ (c)+K}{N+K\ .|classes|}$$

P = Probabilitas variable c.
Count = Jumlah kemunculan dari sampel c.
K = Nilai parameter.
N = Jumlah total kejadian dari sampel c.
|classes| = Jumlah kelas pada sampel.

H. *Mean*

Nilai rata-rata pada seluruh nilai yang dijumlahkan pada sebuah data kemudian dibagi dengan banyaknya data. Persamaan mencari nilai mean dapat dilihat pada persamaan berikut ini.

$$X = \frac{\sum x}{N}$$

Keterangan:
X = *Mean*
∑x = Jumlah data
N = Banyaknya data

I. TF-IDF

Ekstraksi yang digunakan adalah TF-IDF Vectorizer (TF-IDFVec) dimana algoritma ini digunakan untuk menghitung kata yang digunakan. TF-IDF sering kali digunakan untuk memberikan karakteristik dalam dokumen (Sugiyama dkk., 2003). Perhitungan pada metode ini digunakan untuk memunculkan kata yang sering muncul.

$$tf_{t,d} = \frac{n_{t,d}}{max\ (tf)}$$

$$idf_d = log\left(\frac{D}{df(t)}\right)$$

$$tfidf_{t,d} = tf_{t,d} \times idf_d$$

Keterangan:
D = Dokumen ke-d

t = *term* ke-t dari dokumen
W = bobot dokumen ke-d terhadap *term* ke-t
tf = jumlah *term* i pada dokumen
idf = *Inversed Document Frequency*
df = banyak dokumen yang mengandung *term* i

J. *Confusion Matrix*

Confusion Matrix adalah sebuah media pengukuran yang berfungsi menilai akurasi dari hasil klasifikasi pada sebuah penelitian. *True Positive* (TP) dan *True Negative* (TN) digunakan saat klasifikasi benar, sedangkan *False Positive* (FP) dan *False Negative* (FN) digunakan saat klasifikasi menghasilkan kesalahan (Han dkk., 2011).

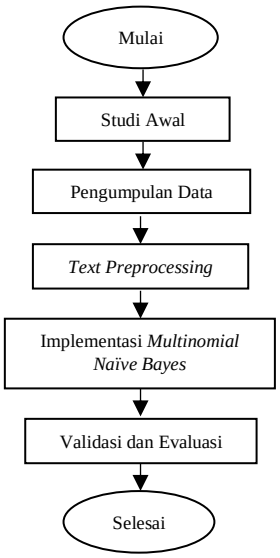
K. *Google Colab*

Google Colab adalah sebuah *platform* yang mengangkat *Jupyter Notebook* dengan sumber daya milik *Google*. Selain itu, *Google Colab* menjadi solusi atas pembatasan akses yang dimiliki oleh *Jupyter Notebook*, karena *Google Colab* mendukung fitur *sharing* layaknya *Google Drive* (Sengkey dkk., 2020).

3. METODE PENELITIAN

A. Diagram Alur Penelitian

Penelitian ini menggunakan langkah-langkah penelitian yang mendukung agar selesai dengan maksimal. Penelitian tentang Analisis *Data Mining* Menggunakan Metode *Multinomial Naïve Bayes* pada Media Pemberitaan menggunakan data dari *Website* unmuhjember.ac.id. Adapun penelitian yang dilakukan dalam beberapa tahap sebagai berikut.



Gambar 2. Diagram alur penelitian

B. Studi Awal

Literatur yang sesuai dalam penelitian. Tahap pertama dari penelitian yaitu mencari dan mempelajari permasalahan yang akan diteliti, kemudian menentukan ruang lingkup permasalahan, pendahuluan, dan mempelajari beberapa literatur yang berkaitan dengan permasalahan yang ditemukan dan mencari solusinya. Agar tujuan tercapai, penulis perlu mempelajari beberapa literatur yang digunakan, lalu diseleksi agar dapat menentukan.

C. Pengumpulan Data

Seluruh data diperoleh dari official *Website* Indozone.id dengan metode *Web Scrapping*. Pengumpulan data diambil dari teks berita berbahasa Indonesia dengan jumlah 19.200 data, dan dapat diakses melalui *link* indozone.id. *Dataset* terdiri dari variabel judul, dan kategori.

Tabel 1. Fitur data yang digunakan

No	Fitur Nama	Deskripsi
1	Judul	Teks singkat yang berisi tentang sebuah artikel atau karangan.
2	Kategori	Bagian dari isi teks yang telah terbagi.

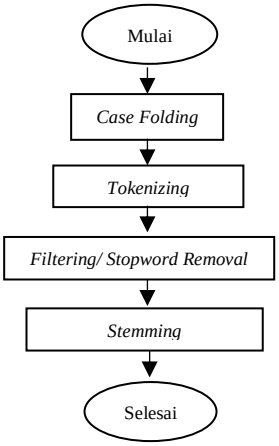
Tabel 2. Jumlah *dataset*

No	Label	Jumlah Data	
		17 Kategori	14 Kategori
1	Fyi	52	-
2	Fakta dan Mitos	1737	1737
3	Film	1234	1234
4	Game	1195	1195
5	Kecantikan	1044	1044
6	Kehidupan	1372	1727
7	Kesehatan	1794	1794
8	Kuliner	1168	1168
9	Musik	1077	1077
10	News	2343	2395
11	Olahraga	161	1226
12	Otomotif	1022	1022
13	Seleb	1252	1252
14	Sepakbola	1065	-
15	Teknologi	1227	1227
16	Travel	1102	1102
17	Videografi	355	-
	Jumlah	19.200	19.200

Penggabungan 17 kategori menjadi 14 kategori bertujuan untuk menyeimbangkan data pada masing-masing label. Oleh sebab itu, label Fyi, Sepakbola, dan Videografi digabungkan ke dalam label Kehidupan, *News*, dan Olahraga dengan memasukkan ketiga data ke dalam label tujuan.

D. *Text Pre-Processing*

Tahapan *teks preprocessing* adalah tahapan awal untuk membersihkan sebuah data dari *noise* pada teks. Adapun diagram alur *preprocessing* adalah sebagai berikut.



Gambar 3. Diagram alur *preprocessing*

E. Implementasi MNB

Setelah data melalui proses *preprocessing*, maka tahapan berikutnya adalah pembobotan kata dengan metode TF-IDF guna mengubah bentuk teks menjadi numerik dengan menggunakan persamaan (2.4), (2.5), (2.6). Pembobotan kata pada judul berita dibutuhkan untuk proses perhitungan *Multinomial Naïve Bayes*. Berikut contoh proses pembobotan dengan TF-IDF dengan 3 dokumen, yaitu marak judi *online* Komisi I DPR ingat masyarakat sanksi pidana (D1), tips cegah radang tenggorokan puasa laku sahur buka (D2), serta tips aman belanja baju lebaran *online shop* fokus merek perhati testimoni (D3).

F. Validasi dan Evaluasi

K-Fold Cross Validation akan digunakan untuk proses validasi dengan cara membagi *dataset* menjadi beberapa bagian dalam artian dua bagian terbagi antara data *training* dan data *testing*. Penentuan nilai *K-fold* yang akan digunakan pada penelitian ini adalah 2, 4, 5, dan 10. Tidak ada aturan resmi untuk penentuan nilai pada *K-Fold Cross Validation* (Kuhn & Johnson, 2013), maka penentuan nilai K diambil nilai yang habis terbagi, hal ini membuat tiap partisi akan memiliki nilai yang sama rata. Proses validasi akan dilakukan terhadap *dataset* yang tersedia kemudian mengevaluasi hasil klasifikasi yang dilakukan.

4. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan diambil dari situs *Website* berita online: Indozone.id tahun 2019-2022. Pengumpulan data tersebut diambil dengan menggunakan Teknik *Web Scraping*. Data yang dikumpulkan disalin kemudian dipindahkan ke dalam bentuk excel, lalu dikonversi menjadi *file* csv-UTF-8 dan diunggah ke dalam *folder* pada *google drive*.

Dataset berisi tentang judul berita dan kategorinya dengan rincian 19.200 judul berita dan 14 kategori. Setelah seluruh data terkumpul, data kemudian dimasukkan dan disimpan ke dalam *folder google drive*.

Tabel 3. Data judul berita

No	Judul	Kategori
1	A Perfect Pairing Resmi Tayang di Netflix, Simak Sinopsisnya	Film
2	Abaikan Barcelona Demi Pilih PSG, Wijnaldum Dinilai Salah Langkah	Olahraga
3	Abang L Anak Lesti Billang Pakai Outfit Gemes, Harga Bikin Mata Netizen Kunang-kunang	Kecantikan
…	…	…
19199	Zombie Terkuat Dalam Game Plants Vs Zombie	Game
19200	Zoom untuk *Chromebook* Bakal Dihentikan Agustus Mendatang	Teknologi

B. Text Pre-Processing

Proses *Processing* memiliki tiga empat tahapan, yaitu *case folding, tokenizing, filtering* dan stemming. Data yang dibersihkan adalah judul berita, karena judul berita termasuk data utama yang akan digunakan. Berikut adalah serangkaian tahapannya.

1) *Case Folding*

   Tahap ini bertujuan untuk mengubah huruf kapital menjadi huruf kecil. Gambar 1 menampilkan data sebelum *Case Folding*, sedangkan gambar 2 memaparkan hasil setelah dilakukan *Case Folding*.

```
data[['comment', 'Kategori']].head(11)
```

	comment	Kategori
0	Teknik Dasar Gerakan Senam Lantai yang Mudah D...	Olahraga
1	Jajanan di Pasar Lama Tangerang yang Paling La...	Kuliner
2	A Perfect Pairing Resmi Tayang di Netflix, Sim...	Film
3	Abaikan Barcelona Demi Pilih PSG, Wijnaldum Di...	Olahraga
4	Abang L Anak Lesti Billar Pakai Outfit Gemes, ...	Kecantikan
5	Ablutophobia Fobia Yang Takut Mandi	Kesehatan
6	Abramovich Jual Chelsea, Tuchel: Berharap Kris...	Olahraga
7	Abramovich Minta Uang RpTriliun Dibalikin, Che...	Olahraga

Gambar 4. Sebelum *case folding*

```
data[['comment', 'Kategori']].head(11)
```

	comment	Kategori
0	teknik dasar gerakan senam lantai yang mudah dilakukan	Olahraga
1	jajanan di pasar lama tangerang yang paling laris, cobain pancake souffle!	Kuliner
2	a perfect pairing resmi tayang di netflix, simak sinopsisnya	Film
3	abaikan barcelona demi pilih psg, wijnaldum dinilai salah langkah	Olahraga
4	abang l anak lesti billar pakai outfit gemes, harganya bikin mata netizen kunang-kunang	Kecantikan
5	ablutophobia fobia yang takut mandi	Kesehatan
6	abramovich jual chelsea, tuchel: berharap krisis kepemilikan ini tuntas	Olahraga
7	abramovich minta uang rptriliun dibalikin, chelsea terancam dicoret dari liga inggris	Olahraga
8	absen tahun, ini sejarah perjalanan motogp indonesia, dari sentul ke mandalika	Fakta dan Mitos

Gambar 5. Setelah *case folding*

2) *Tokenizing*

Tokenizing merupakan tahapan memecah kalimat menjadi kata per kata, sehingga setiap kata memiliki nilai masing-masing. Proses ini memanfaatkan *package* dari sastrawi.

```
data[['comment', 'Kategori']].head(11)
```

	comment	Kategori
0	[teknik, dasar, gerakan, senam, lantai, yang, mudah, dilakukan]	Olahraga
1	[jajanan, di, pasar, lama, tangerang, yang, paling, laris,, cobain, pancake, souffle!]	Kuliner
2	[a, perfect, pairing, resmi, tayang, di, netflix,, simak, sinopsisnya]	Film
3	[abaikan, barcelona, demi, pilih, psg,, wijnaldum, dinilai, salah, langkah]	Olahraga
4	[abang, l, anak, lesti, billar, pakai, outfit, gemes,, harganya, bikin, mata, netizen, kunang-kunang]	Kecantikan
5	[ablutophobia, fobia, yang, takut, mandi]	Kesehatan
6	[abramovich, jual, chelsea,, tuchel:, berharap, krisis, kepemilikan, ini, tuntas]	Olahraga
7	[abramovich, minta, uang, rptriliun, dibalikin,, chelsea, terancam, dicoret, dari, liga, inggris]	Olahraga
8	[absen, tahun,, ini, sejarah, perjalanan, motogp, indonesia,, dari, sentul, ke, mandalika]	Fakta dan Mitos
9	[absen, tahun,, kanada, lolos, ke, piala, dunia, usai, bantai, jamaika]	Olahraga
10	[abu, bakar, baasyir, bebas, murni, dan, tidak, dikenakan, wajib, lapor]	News

Gambar 6. Hasil dari *tokenizing*

3) *Filtering*

Tahap *filtering* akan mengambil kata-kata penting dari hasil token, dan menghapus penggunaan kata penghubung.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

	comment	Kategori
0	[teknik, dasar, gerakan, senam, lantai, mudah]	Olahraga
1	[jajanan, pasar, tangerang, laris,, cobain, pancake, souffle!]	Kuliner
2	[a, perfect, pairing, resmi, tayang, netflix,, simak, sinopsisnya]	Film
3	[abaikan, barcelona, pilih, psg,, wijnaldum, dinilai, salah, langkah]	Olahraga
4	[abang, l, anak, lesti, billar, pakai, outfit, gemes,, harganya, bikin, mata, netizen, kunang-kunang]	Kecantikan
5	[ablutophobia, fobia, takut, mandi]	Kesehatan
6	[abramovich, jual, chelsea,, tuchel:, berharap, krisis, kepemilikan, tuntas]	Olahraga
7	[abramovich, uang, rptriliun, dibalikin,, chelsea, terancam, dicoret, liga, inggris]	Olahraga
8	[absen, tahun,, sejarah, perjalanan, motogp, indonesia,, sentul, mandalika]	Fakta dan Mitos
9	[absen, tahun,, kanada, lolos, piala, dunia, bantai, jamaika]	Olahraga
10	[abu, bakar, baasyir, bebas, murni, dikenakan, wajib, lapor]	News

Gambar 7. Hasil *filtering*

4) *Stemming*

Tahap terakhir yaitu *stemming*, tahapan ini adalah tahapan untuk mengembalikan sebuah kaya yang memiliki kata imbuhan menjadi kata dasar. Proses ini menggunakan *package* yang disediakan oleh sastrawi.

```
data_clean.head(50)
```

```
zoom chromebook henti agustus
```

	No.	comment	Kategori
0	1	teknik dasar gera senam lantai mudah	Olahraga
1	2	jajan pasar tangerang laris cobain pancake souffle	Kuliner
2	3	a perfect pairing resmi tayang netflix simak sinopsis	Film
3	4	abai barcelona pilih psg wijnaldum nilai salah langkah	Olahraga
4	5	abang l anak lesti billar pakai outfit gemes harga bikin mata netizen nang	Kecantikan
5	6	ablutophobia fobia takut mandi	Kesehatan
6	7	abramovich jual chelsea tuchel harap krisis milik tuntas	Olahraga
7	8	abramovich uang rptriliun dibalikin chelsea ancam coret liga inggris	Olahraga
8	9	absen tahun sejarah jalan motogp indonesia sentul mandalika	Fakta dan Mitos
9	10	absen tahun kanada lolos piala dunia bantai jamaika	Olahraga

Gambar 8. Hasil *Stemming*

C. Implementasi MNB

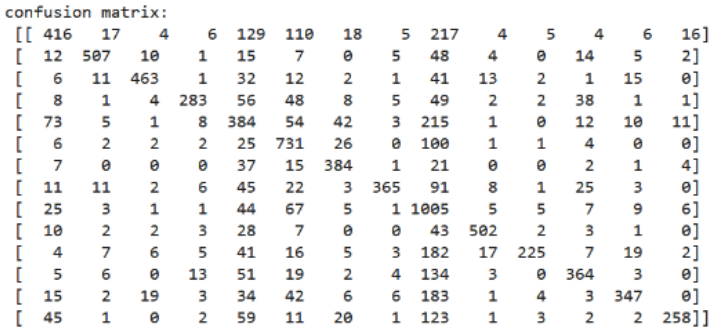
Pada tahap pengujian, penulis menggunkana metode *Multinomial Naïve Bayes* untuk mengklasifikasi judul berita. Penulis menggunakan *library* dari *Scikit Learn*. Untuk dapat melakukan pengujian tersebut, peneliti melakukan tahapan-tahapan yaitu pembangunan model *Pipeline* untuk mempermudah proses pengujian, pembagian data dengan train_test_split dari *Scikit Learn*, dan memvalidasi dengan *K-Fold Validation*.

```
kf = KFold(n_splits=2)
X_array = x_tfidf.toarray()
def cross_val(estimator):
    acc = []
    pcs = []
    rec = []
```

Gambar 9. Data split dengan *K-fold*

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
cross_val(nb)
```

Gambar 10. Model MNB



Gambar 11. *Heatmap confusion matrix*

D. Hasil Evaluasi dan Validasi

Dari serangkaian proses pengujian terhadap klasifikasi data pemberitaan dengan mengambil judul berita dengan menggunakan 14 dan 17 kategori, maka hasilnya dapat dirangkum dalam tabel di bawah ini.

Tabel 4. Hasil evaluasi

Uji Coba	*K-Fold*	Tahap ke-	Akurasi		Presisi		*Recall*		Data *Training*
			Kategori						
			14	17	14	17	14	17	
1	2	1	65%	64%	76%	62%	64%	54%	9600
		2	66%	65%	77%	62%	65%	54%	9600
2	4	1	70%	68%	78%	63,6%	69%	57%	4800
		2	70%	69%	77%	62%	71%	59%	4800
		3	69%	69%	77%	62%	68%	57%	4800
		4	69%	67%	78%	63%	69%	57%	4800
3	5	1	69%	67%	78%	63%	69%	57%	3840
		2	70%	70%	76%	63%	70%	59%	3840
		3	69%	68%	76%	62%	69%	57%	3840
		4	71%	70%	78%	63%	69%	58%	3840
		5	69%	68%	78%	63%	70%	58%	3840
4	10	1	72%	71%	79,1%	**64%**	71%	59%	1920
		2	69%	68%	77%	62%	70%	58%	1920
		3	73%	**73%**	78%	63%	71%	59%	1920
		4	70%	70%	77%	63%	72%	**60%**	1920
		5	72%	70%	77%	62%	70%	59%	1920
		6	68%	68%	75%	61%	68%	58%	1920
		7	70%	69%	78%	63%	68%	57%	1920
		8	74%	72%	79%	63,8%	71%	59%	1920
		9	70%	71%	78%	63,7%	70%	58%	1920
		10	70%	68%	78%	63%	72%	59,6%	1920

Pada tabel 4 menjelaskan beberapa nilai akurasi, *recall*, dan presisi pada klasifikasi judul berita dengan 19.200 data menggunakan metode algoritma *Multinomial Naive Bayes*. Didapatkan hasil uji coba bahwa nilai akurasi tertinggi terhadap 17 kategori akurasi tertinggi mencapai 73%, presisi 64%, dan *recall* sebesar 60%. %. Sementara pada 14 kategori diraih pada percobaan ke 8 dengan nilai K=10 sebesar 74%, presisi pada percobaan ke-1 dengan nilai K=10 sebesar 79,1% dan *recall* pada percobaan ke-10 dengan nilai K=10 sebesar 72%. Maka rata-rata yang didapatkan dari pengujian menggunakan *K-Fold Validation* dapat dilihat pada tabel 5 berikut ini:

Tabel 5. Nilai rata-rata Uji *K-Fold validation*

K-Fold	14 Kategori			17 Kategori		
	Akurasi	Presisi	*Recall*	Akurasi	Presisi	*Recall*
2	65 %	74 %	64%	65%	62%	55%
4	70 %	76 %	69 %	68%	63%	58%
5	70 %	76 %	69 %	69%	63%	58%
10	**71 %**	**77 %**	**71 %**	70%	63%	59%

Berdasarkan tabel 5 di atas dapat dianalisa bahwa salah satu faktor terjadinya tingkat kesalahan prediksi bisa saja terjadi akibat ketidakseimbangan data label yang diperoleh. Kemudian, karena semakin banyak label yang diuji maka dari perbandingan kategori tersebut membuat performa *Multinomial Naïve Bayes* kurang optimal. Sehingga, peneliti mencoba untuk mengurangi jumlah label, yaitu label Sepakbola ke dalam label Olahraga, label FYI ke dalam *News*, dan label Videografi ke dalam label Kehidupan. Hal tersebut peneliti lakukan untuk membandingkan performa tiap kategori dan menyeimbangkan data pada masing-masing label. Akurasi yang dicapai oleh 2 perbandingan mempunyai selisih yang cukup kecil, berbeda dengan selisih presisi dan *recall* yang bisa mencapai lebih dari 10%.

Kategori yang kompleks membuat pemrosesan kata yang tersimpan menjadi sulit untuk dilakukan oleh *Machine Learning*. Pada tabel *confusion matrix* terdapat angka-angka yang tidak terprediksi dengan baik, sehingga terdapat sebaran data yang memiliki nilai besar tetapi tidak terletak di kelas aktual, hal tersebut menyebabkan presisi dan *recall* bernilai kecil. Namun pada 14 kategori nilai pada kelas prediksi dan kelas aktual lebih optimal.

5. KESIMPULAN

A. Kesimpulan

Berdasarkan dari uji coba kelas aktual dan kelas prediksi pada data judul berita dengan algoritma *Multinomial Naïve Bayes*, maka dapat ditarik kesimpulan bahwa, banyaknya data pada tiap-tiap label tidak seimbang, dan berpengaruh terhadap nilai akurasi, presisi dan *recall*. Salah satunya pengaruhnya adalah akurasi yang dicapai rendah karena tidak ada data yang cukup untuk memprediksi, artinya semakin besar nilai *True Positif* (TP) maka semakin tinggi tingkat akurasi, semakin kecil nilai *False Positif* (FP), maka nilai presisi meningkat, dan semakin rendah nilai *False Negatif* (FN) maka nilai *recall* akan membesar.

Hasil pengukuran pada 14 kategori pemberitaan lebih unggul dengan tingkat akurasi sebesar 71%, nilai presisi sebesar 77% dan *recall* sebesar 71%, hal tersebut berbeda pada 17 kategori pemberitaan dengan tingkat akurasinya jauh lebih rendah yaitu selisih 1%, presisi 14%, dan *recall* 12%.

B. Saran

Penelitian ini terdapat beberapa bagian yang kurang sempurna, oleh karena itu penulis berharap adanya pengembangan dari peneliti berikutnya. Adapun beberapa aspek yang perlu dikembangkan adalah dapat menambahkan metode XGBoost untuk menaikkan tingkat akurasinya, dapat menambahkan metode tambahan pada teks *pre-processing* untuk memperkuat pelabelan, serta dapat membandingkan dengan algoritma lain seperti SVM.

6. DAFTAR PUSTAKA

Ariadi, D., & Fithriasari, K. (2015). Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer. *Jurnal Sains Dan Seni ITS*, *4*(2).

Bustami. (2014). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *Jurnal Informatika*, *8*(1), 884–898.

Dewi, F. K. S., & Aji, T. P. (2021). Klasifikasi Berita Menggunakan Metode *Multinomial Naïve Bayes*. *SCAN Jurnal Teknologi Informasi Dan Komunikasi*, *XVI*(3).

Efendi, Z., & Mustakim. (2017). Text Mining Classification Sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi. *Seminar Nasional Teknologi Informasi Komunikasi Dan Industri*, 235–242.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes - Not so stupid after all? *International Statistical Review*, *69*(3).

Kasih, A. P. (2020). *Kompas.com Jadi Portal Berita Online Pilihan Generasi Y dan Z*. Https://Www.Kompas.Com/Edu/Read/2020/12/15/200323471/Kompascom-Jadi-Portal-Berita-Online-Pilihan-Generasi-y-Dan-z?Page=all.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Romli, A. S. M. (2018). *Jurnalistik Online: Panduan Mengelola Media Online*. Penerbit Nuansa Cendekia.

Saleh, A. (2015). Implementasi Metode Klasifikasi *Naïve Bayes* Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Creative Information Technology Journal (CITEC JOURNAL)*, *2*(3).

Sengkey, D. F., Kambey, F. D., Lengkong, S. P., Joshua, S. R., & Kainde, H. V. F. (2020). Pemanfaatan Plaform Pemrograman Daring dalam Pembelajaran Probabilitas dan Statistika di Masa Pandemi COVID-19. *Jurnal Teknik Informatika*, *15*(4).

Sugiyama, K., Hatano, K., Yoshikawa, M., & Uemura, S. (2003). Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. *HYPERTEXT '03: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, 198–207.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Elsevier Inc.