



## **ANALISIS SENTIMEN OPINI MASYARAKAT TERHADAP PINDAHNYA IBU KOTA INDONESIA DENGAN MENGGUNAKAN KLASIFIKASI NAÏVE BAYES**

**Kurnia Ardiansyah Lubis<sup>1)</sup>, Mhd. Theo Ari Bangsa<sup>2)</sup>, Andreo Yudertha<sup>3)</sup>**

<sup>123</sup> Sistem Informasi / Sains dan Teknologi, Universitas Islam Negri Sulthan Thaha Saifuddin Jambi

<sup>123</sup> Jl. Jambi Ma. Bulian KM 16 Sei. Duren Kec. Jaluko Kab. Muaro Jambi, Jambi-Indonesia

Email: <sup>1</sup> Kurniaardiansyah81@gmail.com , <sup>2</sup> theoaribangsa@uinjambi.ac.id , <sup>3</sup> andreo@uinjambi.ac.id

### **Abstract**

*Based on the issues discussed by the government in 2019, various tweets emerged, such as #capitalmove and #relocationofthecity. From the many responses of the Twitter user community to the issue, which is one of the trending topics, it is necessary to conduct a sentiment analysis to find out whether the Indonesian people respond to this issue with a positive, neutral, or even negative response. So that with this research we can see an overview of public opinion regarding government policies that plan to relocate the capital city of Indonesia and can also be used as material for evaluating the government in making decisions, this sentiment analysis makes it easier for us to see data and information on large social media quickly. Collecting data for this study is done by pulling data sets from social media. The data set that is pulled is tweets from the community, and the tweets will be grouped with positive and negative labels using the Naive Bayes Classifier (NBC) method. Analysis using the Naive Bayes method began with crawling data from Twitter on March 6, 2023, including as many as 4533 raw files with the keywords IKN and MOVING CAPITAL CITY. After crawling, data cleaning or preprocessing consists of cleansing, stopwords, and casefolding, leaving 1449 clean records. The data is divided into two categories: training data and test data. Training data is used to provide knowledge to the Rapid Miner in classifying data so that the test data can be predicted by the Rapid Miner application. In the first manual test (69%: 31%), you get an accuracy of 74.61%, and in the second (83%: 17%), you obtain an accuracy of 76.30%. Then, testing by the system with a comparison (50%: 50%) obtained an accuracy of 70.30%, and the second (70%: 30%) test data obtained an accuracy of 73.79%. Regarding the government's policy of wanting to move the capital city of Indonesia, it elicited various positive and negative responses from the public; the most frequent sentiments were positive sentiments of 1073 (74%), and negative sentiments of 376 (26%). Suggestions for further research to add neutral sentiment to further research.*

**Keyword:** Capital of Indonesia, Data mining, Naive Bayes

### **Abstrak**

Berdasarkan isu yang di bahas pemerintah pada tahun 2019, timbulah beragam tweet yang bermunculan seperti #ibukotapindah, #pemindahanibukota. Dari banyaknya respon masyarakat pengguna Twitter terhadap isu yang menjadi salah satu trending topik tersebut maka perlu diadakannya analisis sentimen untuk mengetahui apakah masyarakat Indonesia merespon isu ini dengan respon positif, netral, atau malah negatif. Sehingga dengan adanya penelitian ini kita bisa melihat gambaran pendapat masyarakat terkait kebijakan pemerintah yang merencanakan pemindahan ibu kota negara Indonesia dan juga bisa jadi bahan evaluasi pemerintah dalam pengambilan keputusan, dengan adanya analisis sentimen ini memudahkan kita untuk melihat informasi data di media sosial yang besar dengan cepat. Pengumpulan data pada penelitian ini adalah dengan melakukan penarikan data set dari media sosial. Data set yang di tarik merupakan tweet dari masyarakat dan tweet itu akan di kelompokkan dengan label positif dan negatif menggunakan metode Naive Bayes Classifier (NBC). Analisis dengan metode Naive Bayes di mulai dengan crawling data dari Twitter pada tanggal 6 maret 2023 sebanyak 4533 data mentah dengan 2 kata kunci IKN dan IBU KOTA PINDAH. Setelah crawling selanjutnya pembersihan data atau preprocessing yang terdiri dari cleansing, stopword, dan casefolding sehingga menyisakan data yang bersih sebanyak 1449 data. data di bagi menjadi dua yaitu data latih dan data uji data latih digunakan untuk memberikan pengetahuan kepada rapid miner dalam menggolongkan suatu data sehingga data uji tersebut dapat terprediksi oleh aplikasi Rapid Miner, dalam pengujian manual pengujian pertama (69%:31%) di dapatkan Acuraccy 74.61% dan yang kedua (83%:17%) didapatkan Acuraccy 76.30% selanjutnya pengujian oleh sistem dengan perbandingan (50%:50%) di dapatkan Acuraccy 70.30% dan yang kedua (70%:30%) data uji, didapatkan Acuraccy sebesa 73.79%. Atas kebijakan pemerintah yang ingin memindahkan ibu kota indonesia menimbulkan



beragam respon positif dan negatif dari masyarakat, sentiment yang paling banyak yaitu sentiment positif sebesar 1073 (74%), sentiment negatif ditemukan sebanyak 376 (26%). Saran untuk penelitian selanjutnya untuk menambahkan sentiment netral pada penelitian selanjutnya.

**Kata Kunci:** Ibu kota Indonesia, *Data Mining*, *Naive Bayes*

## 1. PENDAHULUAN

Menurut Bappenas sebanyak 56,56% dari total penduduk Indonesia yang tinggal atau menetap di pulau Jawa. Sedangkan 7,33% saja penduduk yang berada di pulau Kalimantan Indonesia. Dari data persentase tersebut terlihat jelas bahwa pulau Jawa sangat padat penduduk, maka pemerintah berinisiatif untuk memindahkan ibu kota ke Kalimantan. Pemerintah juga mempertimbangkan ketersediaan air bersih dan konversi lahan yang kurang merata [1]. Hampir setiap tahun terjadi banjir yang besarnya bervariasi. Banjir yang terjadi tahun 2007 merupakan yang terbesar, hampir mencakup 70% wilayah Jakarta. Akibat banjir 2007 tersebut menimbulkan pemikiran atau gagasan untuk memindahkan Ibu Kota Jakarta [2].

Perencanaan tentang pemindahan Ibu kota sebenarnya sudah dibahas beberapa tahun yang lalu, namun pada tahun 2019 rencana ini baru mulai terealisasi. Dari rencana pemerintah atas pemindahan ibukota ini menimbulkan banyak pro dan kontra bagi penduduk Indonesia di berbagai media sosial. Masyarakat berpendapat dari anggaran pemindahan Ibu kota baru lebih baik digunakan untuk hal yang lebih penting seperti mengurangi bencana *covid-19* yang terjadi di Indonesia [3]. Atas rencana pemerintah tersebut untuk memindahkan ibu kota, hal ini menimbulkan reaksi positif dan negatif dari masyarakat Indonesia melalui beberapa media sosial seperti Twitter dan Facebook, tidak hanya masyarakat beberapa toko publik pun juga memberikan pro dan kontra terhadap rencana pemerintah tersebut.

Salah satu platform media sosial seperti Twitter dapat kita perhatikan khusus karena para pemakainya sangat dengan mudah mendapatkan maupun mengunggah informasi tentang pendapat mereka tentang masalah tertentu melalui pesan publik yang disebut tweet. Diluar itu, pejabat pemerintah dan tokoh politik menggunakan Twitter untuk memberi informasi kepada publik tentang kegiatan mereka [4]. Pengguna Twitter yang banyak dengan hak yang bebas berpendapat di negara Indonesia, berbagai informasi atau isu politik umumnya menjadi trending topik di Twitter, dan salah satu isu yang sempat gempar dan trending topic adalah rencana pemerintah atas pemindahan ibu kota negara. Dari isu yang di bahas pemerintah pada tahun 2019, timbulah beragam tweet yang bermunculan seperti *#ibukotapindah*, *#pemindahanibukota*. Dari banyaknya respon masyarakat pengguna Twitter terhadap isu yang menjadi salah satu trending topik tersebut maka perlu diadakannya analisis sentimen untuk mengetahui apakah masyarakat Indonesia merespon isu ini dengan respon positif, netral, atau malah negatif. Sehingga dengan adanya penelitian ini kita bisa melihat gambaran pendapat masyarakat terkait kebijakan pemerintah yang merencanakan pemindahan ibu kota negara Indonesia dan juga bisa jadi bahan evaluasi pemerintah dalam pengambilan keputusan, dengan adanya analisis sentimen ini memudahkan kita untuk melihat informasi data di media sosial yang besar dengan cepat.

Analisis sentimen adalah cabang ilmu pada *data mining* yang di pakai untuk menganalisis, mengolah dan memperoleh data tekstual dari entitas semacam layanan, produk, individu, organisasi, dan topik tertentu. Analisis ini berguna buat mendapatkan sebetulnya informasi dari sebuah himpunan data yang ada [5]. Pada analisis sentimen terdapat beberapa metode klasifikasi *mechine learning* seperti *Naïve Bayes*, *Support Vector Machine*, *Logistic Resgion* dan *Lexion Based*, yang digunakan untuk mencari hasil yang terbaik. Namun pada penelitian ini peneliti memilih untuk menggunakan metode *Naïve Bayes Clasifier* (NBC) dikarenakan metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian[6]. Membandingkan data latih dan data uji yang berdasarkan probabilitas kata kuncinya merupakan metode klasifikasi teks *Naive Bayes Classifier*[4], [7], [8].

Terdapat beberapa penelitian terdahulu yang terkait dengan penelitian ini yaitu. Penelitian yang di lakukan oleh (Samsir, et al.,2021) Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode *Naïve Bayes* didapatkan hasil selama penelitian tersebut menunjukkan 30% perasaan positif, 69% perasaan negatif, dan 1% perasaan netral. [10] Analisis Sentimen Pengguna Twitter terhadap Perpanjangan PPKM Menggunakan Metode K-Nearest Neighbor didapatkan hasil Hasil akurasi 69,5%, recall 69,5%, dan presisi 68,7%. Hasilnya kurang memuaskan karena diperlukan ketelitian yang tinggi agar valid ketika digunakan dalam evaluasi kebijakan. [11] Analisis Sentimen Review Film Menggunakan TF-IDF dan SVM didapatkan hasil nilai accuracy 0,85, nilai presisi 1,0, nilai recall 0,7 dan nilai skor F1 0,82. Hal ini dikategorikan baik. [12] Analisis Sentimen Pembelajaran Campuran Pada Twitter Data Menggunakan Algoritma *Naïve Bayes* didapatkan hasil 44,51% positif, 45,80% negatif, dan 9,69 % netral. positif menunjukkan bahwa siswa, guru, dan institusi penyelenggara metode pembelajaran campuran. [13] Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024, mendapat



rata-rata opini negatif hanya 0,1%, netral 97,25, dan positif 2,55%. Pengguna Twitter pada umumnya tidak agresif saat membahas topik yang mengarah pada pencalonan Puan Maharani di Pilpres 2024.

## 2. METODE PENELITIAN

### 2.1 Tahapan Penelitian

Pengumpulan data pada penelitian ini adalah dengan melakukan penarikan data set dari media sosial atau di sebut dengan *Internet Searching*. Data set yang di tarik merupakan tweet dari masyarakat dan tweet itu akan di kelompokkan dengan label positif dan negatif menggunakan metode *Naïve Bayes Clasifier (NBC)*.



Gambar 1 Tahapan Penelitian Menggunakan Klasifikasi *Naïve Bayes*

### 2.2 Data Crawling

Data crawling adalah proses otomatis untuk mengumpulkan dan mengindeks data dari berbagai sumber seperti situs web, database, atau dokumen, pengumpulan data dilakukan dengan crawling data *tweet* menggunakan RapidMiner dengan Keyword *#ibukotapindah, #ikn*.

Gambar 2 Tahapan *Crawling*

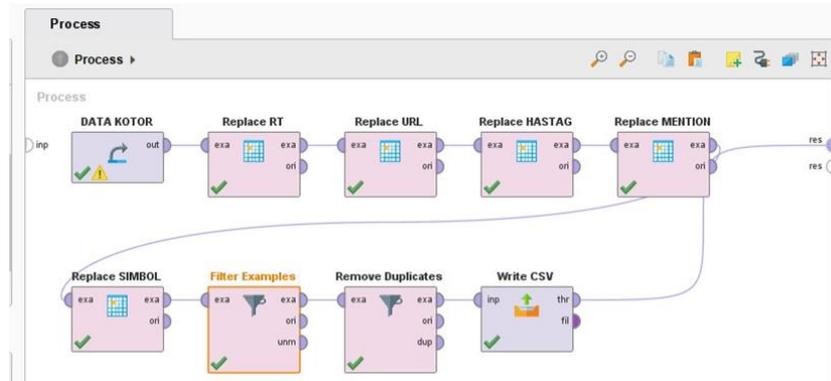
tahapan *Crawling* data pada tanggal 6 maret 2023 dengan menggunakan dua kata kunci yaitu IKN dan IBU KOTA PINDAH, dimana data di tarik menggunakan operator *Search Twitter* menggunakan pengaturan data populer atau yang terbaru, kata kunci IKN sebanyak 3866 data mentah dan IBU KOTA PINDAH sebanyak 667 data mentah. Adapun tampilan beberapa data mentah terdapat pada tabel di bawah:

Tabel 1 Tampilan beberapa data mentah hasil *Crawling*

No	Text
1	@rahmawaadityan Kagak warga ibu kota jg, di kota lain yg rumahnya pinggir rel atau yg pakai lahan punya TNI kalau disuruh pindah karena tanahnya atau rumahnya atau lahannya mau di pakai pasti ngamuk-ngamuk nggak terima. yg KAI malah harus kasih kompensasi sesuai bangunan
2	@sociotalker Seharusnya mumpung ada rencana pindah ibu kota ke Kalimantan, sekalian aja reset NPP buat dibangun di Kalimantan yang notabeneanya 'aman' dari aktivitas seismik. Toh Jepun yg jelas2 di RoF aja berani bikin NPP...
3	@daffanacom ibu kota mau pindah ke ikn. tp proyek unggulan kayak gini aja bisa mangkrak. yakin gak malu maluin indonesia ?
4	sekarang Ada istilah ibu kota di pindah

### 2.3 Preprocessing

Adapun tahap ini adalah mempersiapkan data teks sebelum digunakan pada proses lainnya[14]. Pada tahap ini akan mengubah data teks menjadi bentuk yang lebih baik sehingga menghasilkan informasi teks dengan kualitas yang baik dan siap di gunakan untuk proses selanjutnya.



**Gambar 3** Tahapan pembersihan data

Gambar 4 dimana proses pembersihan data kotor dengan menggunakan operator *Replace* RT, URL, HASTAG, MENTION, SIMBOL. Lalu menggunakan operator *Filter Examples* yang gunanya untuk menghapus data yang kosong dan untuk menghapus data yang sama atau dubel digunakan operator *Remove Duplicates*. Adapun tahapannya dapat di lihat sebagai berikut :

1. *Cleansing*

*Cleansing* adalah tahapan dimana karakter dan tanda baca yang tidak diperlukan dihilangkan dari teks, yang mana salah satu fungsi *cleansing* itu sendiri yaitu dapat mengurangi *noise* pada dataset. Contoh karakter yang di hilangkan berupa URL, tag (#), tanda baca seperti titik (.), koma (,) dan karakter lainnya yang tidak digunakan.

a. *Cleansing RT*

*Cleansing RT* adalah pembersihan kata yang mengandung RT (*retweet*) dimana teks tersebut tidak mempunyai makna dalam analisis *sentiment* sehingga dalam analisis *sentiment* dilakukan *Cleansing RT* dengan menggunakan operator *Replace RT* menggunakan kode **RT @.\***

b. *Cleansing URL*

*Cleansing URL* adalah contoh pembersihan kata yang mengandung URL dimana URL tersebut tidak mempunyai makna dalam analisis *sentiment* sehingga dalam analisis *sentiment* dilakukan *Cleansing URL* dengan operator *Replace URL* menggunakan kode **http.\***

c. *Cleansing HASTAG*

*Cleansing HASTAG* adalah contoh pembersihan kata yang mengandung HASTAG dimana HASTAG tersebut tidak mempunyai makna dalam analisis *sentiment* sehingga dalam analisis *sentiment* dilakukan *Cleansing HASTAG* dengan operator *Replace HASTAG* menggunakan kode **#.\*?**

d. *Cleansing MENTION*

*Cleansing MENTION* adalah contoh pembersihan kata yang mengandung MENTION dimana MENTION tersebut tidak mempunyai makna dalam analisis *sentiment* sehingga dalam analisis *sentiment* dilakukan *Cleansing MENTION* dengan operator *Replace MENTION* menggunakan kode **@.\*?**

e. *Cleansing SIMBOL*

*Cleansing SIMBOL* adalah contoh pembersihan kata yang mengandung SIMBOL dimana SIMBOL tersebut tidak mempunyai makna dalam analisis *sentiment* sehingga dalam analisis *sentiment* dilakukan *Cleansing SIMBOL* dengan operator *Replace SIMBOL* menggunakan kode **[!@#%&\*()\_+ = / , . > < ; " : ]**

f. *Cleansing data kosong*

Tahapan selanjutnya setelah melakukan pembersihan data dengan melakukan *Cleansing* data kosong yaitu tahapan menghapus data yang kosong dari data yang sudah melalui tahapan *Cleansing* dengan menggunakan operator yang bernama *Filter Examples*.

g. *Cleansing data duplikat*

Setelah melakukan tahapan penghapusan data yang kosong maka ada satu tahap lagi yaitu menghapus data yang duplikat dengan menggunakan operator yang bernama *Remove Duplicates* data yang mengandung lebih dari satu data yang sama maka akan di hapus supaya mempermudah pada saat melakukan analisis *sentiment*.

h. *Case Folding*

*Case Folding* adalah untuk menyamaratakan penggunaan huruf kapital. Didalam penulisan tweet pasti terdapat



perbedaan bentuk huruf maka perlu dilakukan yang namanya *Case Folding*.

i. *Tokenizing*

Tokenisasi adalah memecah kalimat menjadi pisah-pisah supaya kata-kata tersebut dapat diproses lebih lanjut dalam analisis teks.

j. *Filtering*

*Filtering* adalah langkah menghilangkan kata-kata yang muncul dalam jumlah banyak tetapi tidak dianggap penting (*stopwords*). Pada dasarnya *stopwords* adalah sekumpulan kata yang banyak digunakan dalam berbagai bahasa. Alasan untuk menghapus kata-kata *stopwords* adalah karena terlalu sering muncul, memungkinkan pengguna untuk fokus pada kata-kata yang lebih penting.

**2.4 Ekstrasi Fitur**

merupakan proses penting pada klasifikasi teks untuk mengubah format tekstual yang tidak terstruktur menjadi terstruktur sehingga dapat diproses oleh algoritma *machine learning* untuk mengklasifikasikan ke class yang telah ditentukan sehingga mempermudah klasifikasi *Naïve Bayes*. Selanjutnya TF-IDF (*Term frequency Inverse Document Frequency*) digunakan untuk menentukan nilai frekuensi sebuah kata dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata dalam sebuah dokumen.

**2.5 Pelabelan Manual**

Sebelum melanjutkan ketahapan berikutnya adapun data yang telah melalui proses yang telah disebutkan di atas, data tersebut di labelkan secara manual dengan sentiment positif dan negatif, apabila terdapat data yang bernada netral maka data tersebut dilihat dulu jika tidak ada kata-kata seperti mengejek, menghina atau cenderung tidak setuju maka di golongan positif, barulah dilanjutkan ke tahap berikutnya, menggunakan Majority rule, yaitu suara terbanyak yang ditentukan melalui voting untuk mengambil keputusan tanpa mempertimbangkan argumentasi.

O1 = Orang pertama, O2 = Orang kedua, O3 = Orang ketiga, N = Negatif, P = Positif

**Tabel 2** Data pelabelan manual

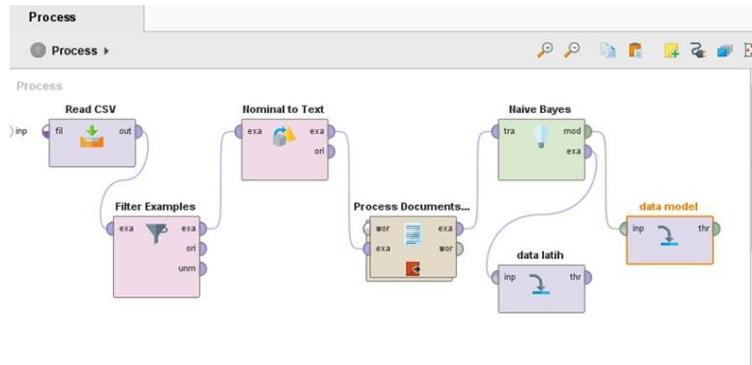
No	Text	O1	O2	O3	Sentimen
1	iya lah klo distop otomatis ini banyak ngalamin kendala masalah baru apalagi ikn loh yg mau dijadiin ibu kota pikir pikir lagi deh ibu kota pindah	N	N	N	Negatif
2	ibu kota mau pindah ke ikn tp proyek unggulan kayak gini aja bisa mangkrak yakin gak malu maluin indonesia ?	N	N	N	Negatif
3	Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Abdullah Azwar Anas mengklaim rencana pemindahan aparatur sipil	P	P	P	Positif
4	Ga ada urgensinya ibu kota pindah	N	N	N	Negatif

Tabel 2 dapat kita lihat di mana beberapa data yang telah dilakukan proses pelabelan secara manual, pada pelabelan manual peneliti melakukan pelabelan dengan menggunakan 3 orang untuk melakukan pelabelan tersebut, jika 2 orang menganggap positif dan 1 orang mengatakan negatif maka *sentiment* positif begitupun sebaliknya.

**2.6 Clasificasion Naïve Bayes**

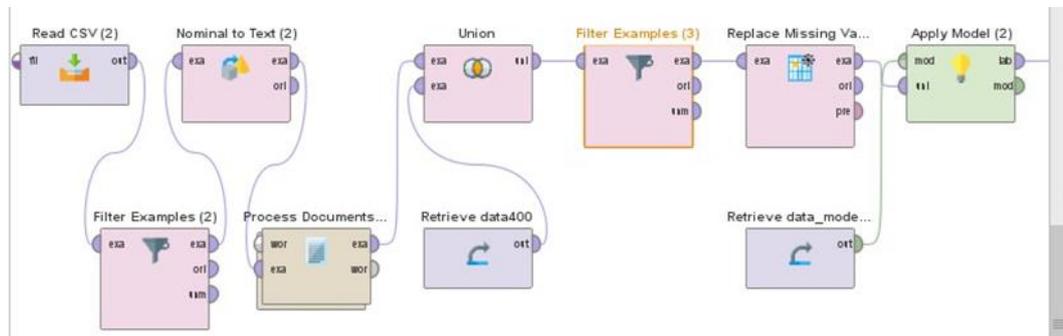
Data dalam bentuk tweet di ambil dari Twitter lalu disimpan dalam bentuk file csv. Data dibagikan menjadi dua set data, yaitu data latih dan data uji. Pelabelan akan disediakan untuk membedakan antara tweet negatif dan positif. Metode Naïve Bayes digunakan pada tahap klasifikasi *sentimen* dan interpretasi hasil analisis *sentiment*, selanjutnya data di *splitting* dengan perbandingan data *training* dan data *testing* yang kemudian dikomputasi menggunakan metode *Naïve Bayes* sehingga menghasilkan luaran.

Tahapan awal dalam menggunakan metode *Naïve bayes* pada Rapid Miner adalah membuat model dari metode *Naïve bayes* tersebut yaitu dapat kita lihat pada gambar dibawah ini :



**Gambar 4** Membuat model *Naive bayes* pada Rapid Miner

Selanjutnya yaitu tahapan mengisi data yang sentimentnya masi kosong dengan menggunakan model *Naive bayes* yang kita buat pada gambar 3.5 tahapannya dapat kita lihat pada gambar dibawah :



**Gambar 5** Proses pelabelan dengan model dan data latihan *Naive Bayes*

### 2.7 Confusion Matrik

Confusion Matrik adalah metode untuk mengukur suatu kinerja klasifikasi, pada cara kerjanya Confusion Matrik bekerja untuk membandingkan atau untuk melihat hasil klasifikasi yang dibuat oleh sistem dengan hasil klasifikasi yang dibuat. Saat pengukuran kinerja menggunakan confusion matrix, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut adalah True Positive (TP), True Negative (TN), False Positive (FP) dan False Negative (FN). Nilai True Negative (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan False Positive (FP) merupakan data negatif namun terdeteksi sebagai data positif [15].

### 2.8 Evaluasi

Tahapan ini adalah tahapan terakhir dalam penelitian ini yaitu evaluasi adalah proses memeriksa secara kritis hasil klasifikasi, kegiatan, kebijakan atau sejenisnya. Ini termasuk mengumpulkan informasi tentang kegiatan dan hasil klasifikasi. Tujuannya adalah untuk melakukan penilaian tentang suatu klasifikasi, meningkatkan keefektifannya, dan menimbang keputusan. Dari proses tersebut nantinya kita dapat mengetahui apakah metode yang digunakan sudah sesuai prediksinya.

#### a. *Split Validation*

*Split validation* merupakan bentuk klasifikasi *Naive Bayes* yang berkerja untuk membagi dua data set yaitu data latih dan data tes dengan jumlah data latih lebih besar dari data tes, data latih digunakan sebagai pelatih data dan data tes di gunakan untuk data yang di uji [15].



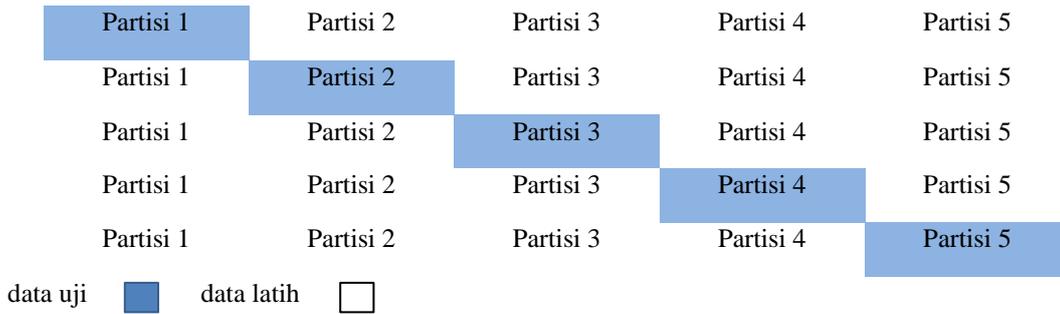
**Gambar 6** Ilustrasi *Split Validation*



b. *Cross Validation*

Cross Validagion dilakukan untuk pengujian klasifikasi yang menggunakan 5 partisi atau disebut dengan  $k = 5$ , pengujian ini berkerja 4 sebagai data latih 1 sebagai data uji, yang disebut dengan 5 skema cross Validation.

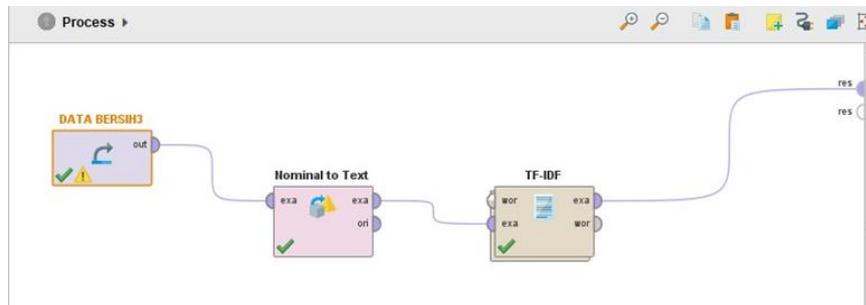
**Tabel 3** Ilustrasi skema *5-fold cross validation*



### 3. HASIL DAN PEMBAHASAN

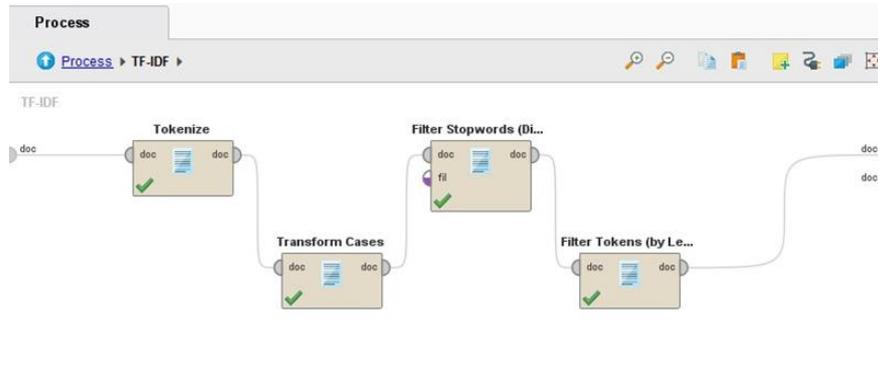
#### 3.1 Pembahasan dan hasil dari Ekstraksi Fitur

Setelah melalui tahapan *Cleansing* maka data tersebut memasuki tahahapan Exstrasi Fitur yang mana di dalam proses tersebut ada istilah yang dinamakan TF-IDF (*Term frequency Inverse Document Frequency*) adapun proses atau tahap TF-IDF dalam Rapid di bawah ini contoh proses tf idf pada Gambar 4.1 berikut :



**Gambar 7** Proses TF-IDF di Rapid Miner

Sebelum melakukan TF-IDF kita perlu mengubah data set yang sudah di *Cleansing* karena data set yang sudah di *Cleansing* tersebut masih bersifat Nominal sedangkan yang dibutuhkan untuk proses dokumen ini adalah tipe data teks maka data set tersebut harus kita ubah dari Nominal ke Teks dengan menggunakan operator yang bernama *Nominal to Text* seperti yang dapat kita lihat di Gambar 7, setelah itu barulah data tersebut bisa di lakukan TF-IDF dengan menggunakan operator yang bernama *Process Documents from Data*, di dalam operator TF-IDF tersebut kita harus memasukkan beberapa operator lagi seperti yang dapat kita lihat pada gambar di bawah :



**Gambar 8** Proses didalam operator TF-IDF

Pada Gambar 4.2 adalah gambar proses di dalam operator TF-IDF di mana terdapat beberapa operator lagi yang pertama adalah *Tokenize* yang mana operator ini berfungsi untuk memisahkan menjadi bagian- bagian kalimat yang terpisah atau kata-kata dengan tujuan untuk proses analisis teks selanjutnya adapun gambaran dari proses ini dapat kita lihat pada gambar sebagai berikut :

Row No.	ABUD	ADA	AI	AIR	AJA	AKAN	AKBP	ALAMI	AMDAL	AMPERA	ANAK	A
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0

**Gambar 9** Gambar hasil dari data yang di *Tokenize*

adapun beberapa contoh kata dari hasil pembobotan kata atau TF-IDF dapat kita lihat pada gambar sebagai berikut :

Word	Attribute Name	Total Occure... ↓	Document Occurences
IKN	IKN	885	824
di	di	477	403
yg	yg	239	183
dan	dan	228	204
kota	kota	181	149
Presiden	Presiden	164	156
yang	yang	160	140

**Gambar 10** Contoh kata-kata hasil pembobotan TF-IDF

Pada Gambar 10 terdapat contoh kata kata yang sudah melalui tahap pembobotan kata atau TF-IDF di mana pada gambar di atas dapat kita lihat kata IKN muncul sebanyak 885 kali dalam 824 data atau dokumen *tweet*, begitu juga dengan kata-kata yang lain yang berhasil di TF-IDF.



Selanjutnya di karnakan kata kata tersebut masi bercampur antara huruf besar dan huruf kecil maka perlu di samaratakan sehingga menjadi huruf kecil semua supaya menjadi kata yang seragam dengan menggunakan operator yang bernama *Transform Cases*.

Setelah menyeragamkan huruf tahapan selanjutnya yaitu menghapus kata- kata yang tidak penting dan tidak mempunyai makna atau bisa disebut dengan istilah *Stopwords* seperti kata-kata, yang, dan, di, adalah dan masi banyak lagi, untuk menghapus kata-kata tersebut kita membutuhkan operator yang bernama *Filter Stopwords(dictionary)* dimana nantinya didalam operator tersebut kita isikan kumpulan kata kata yang tidak penting tersebut. Yang terkhir dalam proses ini adalah menghilangkan kata kata yang singkat seperti kata-kata yang muncul 2 huruf (yg) kata kata seperti itu di hapus menggunakan operator yang bernama *Filter Tokens (by length)* dan berikut gambar kata kata yang telah dilakukan proses *Stopwords* dan *Filter Tokens (by length)* :

Word	Attribute Name	Total Occurences ↓	Document Occurences
ikn	ikn	977	914
kota	kota	282	234
presiden	presiden	216	203
pembangunan	pembangunan	186	176
pindah	pindah	175	162
nusantara	nusantara	160	157
jokowi	jokowi	158	147
indonesia	indonesia	104	103
negara	negara	99	89
menteri	menteri	95	84
bangun	bangun	77	70

**Gambar 11** Kata-kata yang sudah di *Stopwors* dan *Filter Tokens*

Pada Gambar 11 kita dapat melihat di mana kata-kata di gambar tersebut sudah melalui berbagai proses sehingga hanya menampilkan kata-kata yang berharga dan juka menjadi trending topik dalam kedua kata kunci IKN dan IBU KOTA PINDAH. Supaya tampilannya lebih menarik maka peneliti akan menampilkan sebuah visualisasi yang cukup tren di analisis *sentiment* yaitu berupa *Wordcloud* :



**Gambar 12** *Wordcloud*



Pada Gambar 12 di atas peneliti mengambil 30 contoh kata teratas kecuali kata IKN karena kata IKN adalah kata kunci, kita dapat melihat sebuah kumpulan kata-kata yang banyak muncul atau trending di dalam teks twiiter di mana semakin besar kata-kata pada gambar *Wordcloud* maka semakin besar bobot kata-kata tersebut begitupun sebaliknya. Pada gambar juga terdapat kata-kata lain yang mana kata-kata yang muncul seperti contohnya kata-kata seperti presiden Jokowi ternyata menjadi trending topik atau banyak di sebut dalam Twitter juga.

**3.2 Hasil Permodelan Metode Naïve Bayes**

Dari data keseluruhan yang tersisa setelah melalui berbagai tahap yang telah dipaparkan sebelumnya tersisalah data sebanyak 1.449 data. Dari data 1.449 tersebut dilakukan 2 kali pengujian yang pertama yang mana 1000 sebagai data latih dan 449 sebagai data uji sedangkan pengujian yang kedua 1.200 sebagai data latih 249 sebagai data uji baru di dapatlah hasil *Accuracy*, presisi dan *Recall* yang dihitung menggunakan rumus :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$Presisi = \frac{TP}{FP+TP} * 100\%$$

$$Recall = \frac{TP}{FN+TP} * 100\%$$

Dimana :

- TP adalah True positif, yaitu jumlah data yang terklasifikasi dengan benar oleh sistem.
- TN adalah True negatif, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN adalah false negatif, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP adalah false positif, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

Untuk hasil pengujian manual dapat kita lihat pada tabel dan penjelasan dibawah:

**Tabel 4** Hasil uji data uji 449 dan data latih 1000

	True negatif	True positif
<b>pred.negatif</b>	67	70
<b>pred.positif</b>	44	268
<b>Total</b>		449

$$Accuracy = \frac{268+67}{268+67+70+44} * 100\%$$

$$= 74.61\%$$

$$Presisi = \frac{268}{70+268} * 100\%$$

$$= 79.28\%$$

$$Recall = \frac{268}{44+268} * 100\%$$

$$= 85.89\%$$

**Tabel 5** Hasil uji data uji 249 dan data latih 1200

	True negatif	True positif
<b>pred.negatif</b>	41	39
<b>pred.positif</b>	20	149
<b>Total</b>		249

$$Accuracy = \frac{149+41}{149+41+39+20} * 100\%$$

$$= 76.30\%$$

$$Presisi = \frac{149}{39+149} * 100\%$$

$$= 79.25\%$$

$$Recall = \frac{149}{20+149} * 100\%$$

$$= 79.25\%$$

Berdasarkan pengujian data di atas di mana yang pertama dengan jumlah data latih 1000 dan data uji 449 atau 69% : 31% di dapatkan nilai *Acuraccy* sebesar 74.61%, Presisi 79.28%, *Recall* 85.89% dan pengujian yang kedua



dengan data latih 1200 dan data uji 249 atau 83% : 17% didapatkan nilai *Acuracy* sebesar 76.30%, *Presisi* 79.25%, *Recall* 79.25%.

### 3.3 Hasil pengujian *Split Validation* oleh sistem

Untuk hasil pengujian dari sistem dapat kita lihat pada gambar dan penjelasan dibawah :

accuracy: 70.30%

	true negatif	true positif	class precision
pred. negatif	115	142	44.75%
pred. positif	73	394	84.37%
class recall	61.17%	73.51%	

**Gambar 13** Pengujian data 50% : 50%

Berdasarkan hasil pengujian yang dapat kita lihat pada Gambar 4.7 di mana pengujian di lakukan dengan membagi data uji dan data latih dengan perbandingan 50% :50% di dapatkan nilai *Acuracy* sebesar 70.30%, *Presisi* 73.51% dan *Recall* 84.37%.

accuracy: 73.79%

	true negatif	true positif	class precision
pred. negatif	78	79	49.68%
pred. positif	35	243	87.41%
class recall	69.03%	75.47%	

**Gambar 14** Pengujian data 70% : 30%

Hasil pengujian pada Gambar 4.8 di mana pengujian dilakukan dengan membagi data uji dan data latih dengan perbandingan 70% data latih dan 30% data uji, didapatkan nilai *Acuracy* sebesar 73.79%, *Presisi* 75.47%, *Recall* 87.41%.

Berdasarkan kedua pengujian di atas menunjukkan bahwa semakin besar data latih yang digunakan untuk tahap pengklasifikasian maka semakin besar pula nilai *Acuracy* yang di dapat dengan dibuktikan dua tahapan pengujian *Split Validation* di atas yaitu pengujian secara manual dan pengujian secara otomatis oleh sistem Rapid Miner.

### 3.4 Hasil pengujian *Cross Validation*

Setelah melakukan pengujian dengan *Split validation* selanjutnya pengujian dengan menggunakan *Cross Validation* di mana pada pengujian ini semua data 1449 di dapatkan *accuracy* sebesar 71.98% yang mana dapat kita lihat pada Gambar 4.9 dibawah ini.

accuracy: 71.98% +/- 1.71% (micro average: 71.98%)

	true negatif	true positif	class precision
pred. negatif	248	278	47.15%
pred. positif	128	795	86.13%
class recall	65.96%	74.09%	

**Gambar 15** pengujian *Cross Validation*

### 3.5 Hasil percobaan data baru

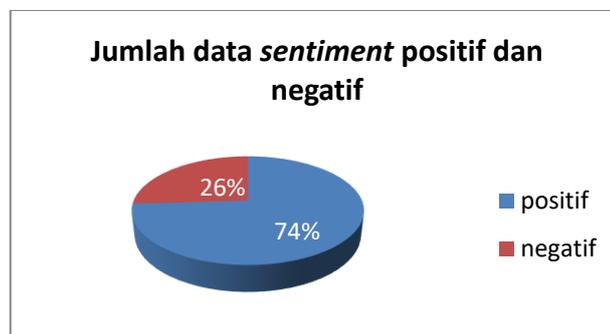
Selanjutnya yaitu peneliti mencoba memasukkan data baru kedalam klasifikasi ini untuk melihat bagaimana hasil prediksinya adapun data dan prediksinya dapat dilihat pada tabel di bawah :



**Tabel 6** Percobaan tambah data baru

No	Text	Prediksi	Hasil
1	Indonesia makin maju karena ikn	Positif	Benar
2	Semangat untuk pembangunannya	Positif	Benar
3	Saya bangga jadi warga indonesia	Negatif	Salah
4	Wiih ibu kota baru pasti mantapni	Negatif	Salah
5	Ikn itu gak terlalu penting menghabiskan uang saja	Negatif	Benar
6	Ikn gak penting	Negatif	Benar
7	Indonesia bisa bobrok kalau maksa bangun ikn	Negatif	Benar
8	Menyusahkan rakyat	Negatif	Benar

Berdasarkan data di atas kita dapat melihat dari 8 data di atas yang betul terprediksi dengan benar adalah 6 data dan 2 data salah.



**Gambar 16** Jumlah sentiment positif dan negatif

Berdasarkan gambar diagram di atas menampilkan jumlah data yang bernada positif dan negatif dengan keseluruhan data 1449, sentiment yang bernada positif berjumlah 1073 atau 74% dan sentiment negatif berjumlah 376 atau 26% ini menunjukkan bahwa sentiment yang benada positif lebih dominan tinggi.

### 3.6 Evaluasi

**Tabel 7** Hasil Evaluasi

Perbandingan Pengujian	Acuraccy	Precision	Recall
Manual, data uji 449 dan data latih 1000	74.61%	79.28%	85.89%
Manual, data uji 249 dan data latih 1200	76.30%	79.25%	79.25%
Split Validation, 50% : 50%	70.30%	73.51%	84.37%
Split Validation, 70% : 30%	73.79%	75.47%	87.41%
Cross Validation keseluruhan data	71.98%	74.09%	86.13%

## 4. KESIMPULAN

Metode Naïve Bayes dapat digunakan sebagai metode untuk mengklasifikasi sentiment. Analisis dengan metode Naïve Bayes di mulai dengan crawling data dari Twitter pada tanggal 6 maret 2023 sebanyak 4533 data mentah dengan 2 kata kunci IKN dan IBU KOTA PINDAH. Setelah crawling selanjutnya pembersihan data atau preprocessing yang terdiri dari cleansing, stopword, dan casefolding sehingga menyisakan data yang bersih sebanyak 1449 data. Tahapan terakhir



yaitu mengklasifikasikan data dengan metode *Naive Bayes* dengan menggunakan aplikasi Rapid Miner, pelabelan manual diperlukan sebelum menggunakan metode selanjutnya data di bagi menjadi dua yaitu data latih dan data uji data latih digunakan untuk memberikan pengetahuan kepada rapid miner dalam menggolongkan suatu data sehingga data uji tersebut dapat terprediksi oleh aplikasi Rapid Miner, dari pengujian yang sudah di lakukan dapat disimpulkan bahwa semakin banyak data latih yang digunakan semakin tinggi *Accuracy*. Atas kebijakan pemerintah yang ingin memindahkan Ibu kota Indonesia menimbulkan beragam respon positif dan negatif dari masyarakat, *sentiment* yang paling banyak yaitu positif sebesar 1073 (74%), *sentiment* negatif ditemukan sebanyak 376 (26%). Banyaknya *sentiment* positif dengan jumlah *sentiment* negatif ini menunjukkan bahwa masyarakat Indonesia tidak menutup kemungkinan menyetujui atas kebijakan pemerintah dalam pemindahan Ibu kota Indonesia.

## DAFTAR PUSTAKA

- [1] H. B. Rochmanto and E. M. P. Hermanto, "Analisis Sentimen Pengguna Twitter terhadap Kebijakan Pemindahan Ibukota Indonesia Menggunakan Metode NBC dan SVM," *Pros. Semin. Nas. Stat. IX 2020*, vol. 9, no. 2020, pp. 1–11, 2020.
- [2] M. Yahya, "Pemindahan Ibu Kota Negara Maju dan Sejahtera," *J. Stud. Agama dan Masy.*, vol. 14, no. 1, p. 21, 2018, doi: 10.23971/jsam.v14i1.779.
- [3] M. P. Agustina and H. Hendry, "Sentimen Masyarakat Terkait Perpindahan Ibukota Via Model Random Forest dan Logistic Regression," *Aiti*, vol. 18, no. 2, pp. 111–124, 2021, doi: 10.24246/aiti.v18i2.111-124.
- [4] W. Yulita *et al.*, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naive Bayes Classifier," *Jdmsi*, vol. 2, no. 2, pp. 1–9, 2021.
- [5] V. Kevin, S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization ( Online Transportation Sentiment Analysis Using Support Vector Machine Based on Particle Swarm Optimization )," vol. 9, no. 2, pp. 162–170, 2020.
- [6] D. Alita and R. B. A. Shodiqin, "Sentimen Analisis Vaksin Covid-19 Menggunakan Naive Bayes Dan Support Vector Machine," *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 1–12, 2023.
- [7] T. Krisdiyanto, "Analisis sentimen opini masyarakat Indonesia terhadap kebijakan PPKM pada media sosial Twitter menggunakan Naive bayes classifiers," *J. CoreIT*, pp. 32–37, 2021.
- [8] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.
- [9] R. W. Samsir<sup>1</sup>, Ambiyar<sup>2</sup>, Unung Verawardina<sup>3</sup>, Firman Edi<sup>4</sup> and 1, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 1, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.
- [10] A. Asro'i and H. Februariyanti, "Analisis Sentimen Pengguna Twitter Terhadap Perpanjangan Ppkm Menggunakan Metode K-Nearest Neighbor," *J. Khatulistiwa Inform.*, vol. 10, no. 1, pp. 17–24, 2022, doi: 10.31294/jki.v10i1.12624.
- [11] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.
- [12] R. Watrianthos, M. Giatman, W. Simatupang, R. Syafriyeti, and N. K. Daulay, "Analisis Sentimen Pembelajaran Campuran Menggunakan Twitter Data," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 166, 2022, doi: 10.30865/mib.v6i1.3383.
- [13] D. A. Vonega, A. Fadila, and D. E. Kurniawan, "Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024," vol. 6, no. 2, 2022.
- [14] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," *J. Tekno Kompak*, vol. 15, no. 1, pp. 131–145, 2021.
- [15] Karsito and Susanti Santi, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naive Bayes Di Perumahan Azzura Residence," *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.