

## ANALISIS SENTIMEN PEMINDAHAN IBU KOTA NEGARA DENGAN KLASIFIKASI NAÏVE BAYES UNTUK MODEL BERNOULLI DAN MULTINOMIAL

Nabila Surya Wardani<sup>1</sup>, Alan Prahutama<sup>2</sup>, Puspita Kartikasari<sup>3</sup>

<sup>1,2,3</sup>Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro  
[nabilasuryaw@gmail.com](mailto:nabilasuryaw@gmail.com)

### ABSTRACT

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Indonesian President affirmed that the capital would be moved to East Kalimantan on August 26, 2019. That planning would received pros and cons from public. Sentiment analysis is part of text mining that typically involves taking data from opinion, comment or response. Sentiment analysis is the choice to do in this topic to get result about public's opinion. As the most used social media in Indonesia, Youtube is able to be data source by crawling the comments on video uploaded by Kompas TV channel. Those comments were crawled on October, 15 2019 and selected 1500 latest comments (August 26 – October 12, 2019). The selected comments get transformed by using data pre-processing technique that involves case folding, removing mention, unescaping HTML, removing numbers, removing punctuation, text normalization, stripping whitespace, stopwords removal, tokenizing, and stemming. Labelling of sentiment class uses the sentiment scoring technique. The number of negative comments is 849, while the number of positive comments is 651. Ratio between training data and testing data is 80% : 20%. The classification method used to do sentiment analysis is Naive Bayes Classifier for Bernoulli and Multinomial model. Bernoulli model only uses occurrence information, whereas the multinomial model keeps track of multiple occurrences. The results show that Bernoulli Naïve Bayes has 93,45% level of sensitivity (recall) and Multinomial Naïve Bayes has 90,19% level of sensitivity (recall). It means that both Bernoulli and Multinomial have good result for this research.

**Keywords:** Text Mining, Relocation of Indonesia's Capital, Youtube, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Sensitivity (Recall).

### 1. PENDAHULUAN

Laman *We Are Social Hootsuite* melaporkan bahwa hingga Januari 2019, pengguna aktif media sosial di Indonesia sudah mencapai 56% atau sama dengan 150 juta orang dari seluruh total populasi sebesar 268,2 juta. *Platform* media sosial nomor satu yang paling banyak diakses oleh sebesar 88% pengguna aktif media sosial di Indonesia adalah Youtube (*Hootsuite*, 2019). Youtube merupakan situs web yang menyediakan fasilitas *video sharing* yang memungkinkan penggunanya untuk memperoleh informasi dengan baik dan *up to date*.

Data yang dipublikasikan oleh Badan Pusat Statistika menunjukkan jumlah populasi Pulau Jawa pada tahun 2015 adalah 145.143.000 jiwa dari total seluruh populasi di Indonesia sebesar 238.518.000 jiwa. Hal ini berarti sekitar 60,85% dari total populasi Indonesia terkonsentrasi di Pulau Jawa sedangkan di Pulau Kalimantan hanya 6% atau sekitar 16,23 juta penduduk. Padatnya penduduk di Pulau Jawa inilah yang menjadi salah satu alasan Presiden Joko Widodo untuk memindahkan ibu kota negara yang semula di Jakarta menjadi di Kalimantan Timur. Rencana pemindahan ibu kota tersebut tentu menuai berbagai opini dari masyarakat.

Analisis sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini (Rozi, et al., 2012). Terdapat banyak metode klasifikasi yang dapat digunakan untuk melakukan analisis sentimen, salah satunya adalah metode klasifikasi *Naïve Bayes*. Metode ini dipilih karena merupakan metode sederhana

yang dikembangkan berdasarkan aturan Bayes dengan melihat kondisi-kondisi yang ada dan peluang-peluang pada setiap kondisinya. Klasifikasi *Naïve Bayes* yang digunakan dalam pengolahan teks, yaitu *Multinomial Naïve Bayes* dan *Bernoulli Naïve Bayes* (Manning, et al., 2009). Model Multinomial dalam klasifikasi *Naïve Bayes* merupakan model yang memperhatikan jumlah kemunculan kata dalam dokumen, sedangkan model Bernoulli merupakan model yang hanya memperhatikan kemunculan kata pada dokumen dan mengabaikan jumlah kemunculannya (Manning, et al., 2009).

Berdasarkan latar belakang tersebut, penelitian analisis sentimen mengenai pemindahan ibu kota negara ini akan dilakukan menggunakan metode klasifikasi *Naïve Bayes* dengan model Bernoulli dan Multinomial. Penelitian tersebut menggunakan data komentar masyarakat yang terdapat dalam video pengumuman resmi mengenai rencana pemindahan ibu kota yang dipublikasikan oleh akun resmi Kompas TV di situs Youtube dengan judul “Tok! Presiden Putuskan Kalimantan Timur Jadi Ibu Kota Baru Indonesia, Ini Pertimbangannya”.

## **2. TINJAUAN PUSTAKA**

### **2.1. Pemindahan Ibu Kota Negara**

Presiden Joko Widodo secara resmi mengumumkan rencana pemindahan ibu kota Indonesia yang semula terletak di Jakarta menjadi di Kalimantan Timur pada tanggal 26 Agustus 2019. Faktor-faktor yang menjadi alasan pemerintah dalam pemindahan ibu kota ke luar Jawa, yaitu terdapat sekitar 57% penduduk Indonesia terkonsentrasi di Pulau Jawa, kontribusi ekonomi per pulau terhadap PDB Nasional, krisis ketersediaan air di Pulau Jawa terutama DKI Jakarta dan Jawa Timur, konversi lahan terbesar terjadi di Pulau Jawa, pertumbuhan urbanisasi yang sangat tinggi dengan konsentrasi penduduk terbesar di Jakarta dan Jabodetabekpunjur, dan meningkatnya beban DKI Jakarta (Bappenas, 2019).

Pemindahan ibu kota memiliki potensi untuk meningkatkan pertumbuhan ekonomi nasional dengan tingkat inflasi yang tetap rendah. Selain itu, pertumbuhan wilayah industri dengan rantai nilai hingga hilir akan meningkatkan pendapatan masyarakat yang pada gilirannya akan menurunkan ketimpangan ekonomi antara Pulau Jawa dan luar Pulau Jawa (Silalahi, 2019).

### **2.2. Media Sosial Youtube**

Youtube merupakan situs media sosial yang menyediakan layanan *video sharing* secara gratis yang dapat diakses oleh seluruh golongan masyarakat di dunia. Laman *We Are Social Hootsuite* menyatakan bahwa per periode Oktober 2019, Youtube menjadi media sosial yang menduduki peringkat nomor dua setelah Facebook dalam kategori *Top Social Platforms* dengan jumlah pengguna aktif sebesar 2000 miliar dari seluruh dunia (Hootsuite, 2019).

### **2.3. Text Mining**

*Text mining* merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dan *text mining* merupakan variasi dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar (Ratniasih, et al., 2017). *Text mining* membantu organisasi menemukan potensi wawasan bisnis yang berharga dalam dokumen perusahaan, *email* pelanggan, *log call center*, komentar survei verbalim, *posting* jaringan sosial, catatan medis dan sumber data berbasis teks lainnya.

### **2.4. Analisis Sentimen**

Sentimen adalah tentang menggali sikap, emosi, perasaan yang bersifat subyektif dari sebuah fakta (Batrinca dan Treleaven, 2014). Analisis sentimen secara umum bertujuan untuk menentukan sikap yang diungkapkan oleh penulis teks atau pembicara sehubungan dengan topik atau polaritas kontekstual keseluruhan dokumen. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau pendapat (Ling, et al., 2014). Polaritas mempunyai arti bahwa teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif. Analisis sentimen merupakan bagian dari *text mining* yang memanfaatkan opini, tanggapan, dan komentar berupa teks yang diolah secara komputasi.

## **2.5. Youtube Crawling**

*Youtube Crawling* adalah penerapan teknik *crawling* pada media sosial Youtube yang bertujuan untuk mengumpulkan informasi yang dibutuhkan, baik informasi komentar, data statistik mengenai jumlah *views*, *likes*, dan hal – hal lain di dalam situs Youtube. Hal yang perlu diperhatikan sebelum melakukan *Youtube Crawling* adalah memastikan kode API (*Application Programming Interface*) sudah didapatkan dari situs Google Api Console.

## **2.6. Data Pre-Processing**

Tahap *data pre-processing* merupakan proses untuk mempersiapkan data dalam bentuk tesktual yang mentah sehingga dapat diolah. *Data pre-processing* bertujuan untuk mengurangi ukuran data, menemukan hubungan antara data, menormalisasikan data, membuang *outliers*, dan mengekstraksi fitur pada data (Alasadi dan Bhaya, 2017). Oleh karena itu, tahapan *data pre-processing* penting dilakukan untuk menghindari data yang kurang sempurna, mengalami gangguan, dan data-data yang tidak konsisten. Terdapat tahapan-tahapan dalam *data pre-processing* yang harus dilakukan sesuai dengan keadaan data yang dimiliki. Penelitian ini melakukan beberapa tahapan, yaitu: *case folding*, *remove mention*, *unescape HTML*, *remove numbers*, *remove punctuation*, normalisasi kata, *strip whitespace*, *stopwords removal*, *tokenizing*, dan *stemming*.

## **2.7. Sentiment Scoring**

*Sentiment scoring* merupakan teknik yang digunakan untuk melabelkan suatu pernyataan yang digolongkan menjadi sentimen positif atau negatif. *Sentiment scoring* membutuhkan kamus *boosterwords* dan kamus kata negasi. *Boosterwords* adalah kata yang dapat meningkatkan atau mengurangi intensitas sentimen kata disebelahnya. Kata negasi merupakan kata yang terdapat dalam suatu kalimat yang dapat mengubah orientasi dari suatu opini. Misalnya, kata “nakal” merupakan kata yang bermakna negatif, akan tetapi jika kata “nakal” didahului kata negasi “tidak” maka susunan kata menjadi “tidak nakal” yang bermakna positif (Wahid dan Azhari, 2016). Kamus sentimen juga dibutuhkan dalam proses penentuan kelas sentimen pada suatu pernyataan. Kamus sentimen berisi kumpulan kata yang telah diberi bobot dengan kekuatan sentimen 1 (tidak memiliki sentimen positif) s.d. 5 (memiliki sentimen positif yang sangat kuat), dan -1 (tidak memiliki sentimen negatif) s.d. -5 (memiliki sentimen negatif yang sangat kuat) (Wahid dan Azhari, 2016).

## **2.8. Text Vizualisation**

*Text vizualization* merupakan teknik yang digunakan untuk menampilkan visualisasi data teks. Salah satu bentuk dari visualisasi teks adalah *word cloud*. *Word cloud* memberikan informasi yang relevan dengan menunjukkan istilah atau *term* yang penting dalam teks

(Tessem, et al., 2015). Konsep pembuatan *word cloud* yaitu kata-kata yang memiliki ukuran huruf terbesar merupakan kata yang paling penting. *Word cloud* bekerja dengan memanfaatkan frekuensi kata yang terdapat pada teks.

## 2.9. Pembobotan Kata

Perhitungan pembobotan kata ini memerlukan dua hal, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* merupakan banyaknya jumlah kata atau *term* tertentu yang ada dalam suatu dokumen. Sementara *Inverse Document Frequency* adalah frekuensi kemunculan kata atau *term* pada seluruh dokumen. Nilai IDF berbanding terbalik dengan jumlah dokumen yang mengandung *term* tertentu (Rahman, et al., 2017). Salton (1989) menyatakan bahwa TF-IDF dapat dihitung menggunakan persamaan sebagai berikut:

$$W_{j,i} = \frac{n_{j,i}}{\sum_k n_{k,i}} \cdot \log_2 \frac{D}{d_j} \quad (1)$$

Keterangan:

$W_{j,i}$  = Pembobotan TF-IDF untuk *term* ke j pada dokumen ke i.

$n_{j,i}$  = Jumlah kemunculan *term* ke j pada dokumen ke i.

$\sum_k n_{k,i}$  = Jumlah kemunculan seluruh *term* pada dokumen ke i.

$D$  = Banyaknya dokumen yang dibangkitkan.

$d_j$  = Banyaknya dokumen yang mengandung *term* ke j.

## 2.10. Klasifikasi Naïve Bayes

### 2.9.1. Bernoulli Naïve Bayes

Klasifikasi *Bernoulli Naïve Bayes* merupakan suatu model yang menetapkan bahwa dokumen diwakili oleh vektor atribut biner yang menunjukkan kata yang muncul dan tidak muncul dalam dokumen (McCallum dan Nigam, 1998). Model Bernoulli hanya memperhatikan kemunculan kata dalam dokumen dan mengabaikan frekuensi dari data yang muncul. Manning, et al. (2009) menyatakan bahwa pada kasus analisis teks klasifikasi *Bernoulli Naïve Bayes*, probabilitas dokumen  $d$  yang terletak di kategori  $c$  memiliki persamaan :

$$P(c|d) \propto P(c) \prod_{1 \leq i \leq M} P(d_i|c) \quad (2)$$

dengan  $P(d_i|c)$  adalah *conditional probability* dan  $P(c)$  merupakan *prior probability* dokumen yang berada pada kategori  $c$  (Manning, et al., 2009).

Kategori terbaik dalam klasifikasi *Naïve Bayes* metode Bernoulli adalah kategori yang memiliki nilai maksimal atau *maximum a posteriori* (MAP) class  $C_{map}$  :

$$C_{map} = \arg \max P(c|d) \quad (3)$$

berikut merupakan formula dari *prior probability* :

$$P(c) = \frac{N_c}{N} \quad (4)$$

$N_c$  merupakan jumlah dokumen dari kategori  $c$ , sedangkan  $N$  merupakan jumlah dokumen dari kategori keseluruhan. *Conditional probability* akan dihitung berapa dokumen latih pada kategori  $c$  yang mengandung *term* (di sembarang posisi dan dapat muncul beberapa kali) dengan persamaan :

$$P(d|c) = \prod_{1 \leq i \leq M} (B_{it}P(e_i|c) + ((1 - B_{it})(1 - P(e_i|c))) \quad (5)$$

dengan  $P(e_i|c)$  merupakan probabilitas bahwa dalam dokumen kategori  $c$ ,  $term$  akan muncul (di sembarang posisi dan dapat muncul beberapa kali).  $B_{it}$  memiliki nilai 0 jika  $term$  tidak muncul dan 1 jika  $term$  muncul dalam dokumen (McCallum dan Nigam, 1998). Ketidakhadiran suatu  $term$  dalam model Bernoulli menjadi faktor yang diperhitungkan sehingga akan mempengaruhi hasil klasifikasi (Manning, et al., 2009).

$$P(e_i|c) = \frac{df_i + 1}{df' + 2} \quad (6)$$

$df'$  adalah jumlah seluruh dokumen pada kategori  $c$  dan  $df_i$  adalah jumlah dokumen latih pada suatu kategori  $c$  yang mengandung  $term$ . Angka 2 merupakan nilai konstan yang menunjukkan kejadian muncul dan tidak munculnya  $term$ . Penambahan angka 1 berfungsi sebagai *Laplace Smoothing* agar terhindar dari *zero probability* (Manning, et al., 2009).

### 2.9.2. Multinomial Naïve Bayes

Klasifikasi *Multinomial Naïve Bayes* menentukan kategori dokumen tidak hanya berdasarkan kata yang muncul pada, namun juga berdasarkan jumlah kemunculannya (Witten, et al., 2011). Manning, et al. (2009) menyatakan bahwa probabilitas dokumen  $d$  yang terletak pada kategori  $c$  memiliki persamaan:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (7)$$

$P(t_k|c)$  adalah *conditional probability* dari  $t_k$  yang berada pada dokumen yang dimiliki kategori  $c$  dan  $P(c)$  merupakan *prior probability* dokumen yang berada pada kategori  $c$  (Manning, et al., 2009).

Kategori terbaik dalam klasifikasi *Naïve Bayes* metode Multinomial adalah kategori yang memiliki nilai maksimal atau *maximum a posteriori* (MAP) class  $C_{map}$ :

$$C_{map} = \arg \max P(c|d) \quad (8)$$

Persamaan dari *prior probability* model Multinomial sama dengan model Bernoulli yang dinyatakan pada persamaan (4). *Conditional probability* akan menghitung kata atau  $term$   $t_k$  pada seluruh dokumen latih pada kategori  $c$  dengan persamaan, yaitu :

$$P(t_k|c) = \frac{N_k + 1}{|V| + N'} \quad (9)$$

$|V|$  adalah jumlah seluruh  $term$  atau kata unik (jika berulang, tetap dihitung 1) pada data pelatihan.  $N_k$  adalah jumlah kemunculan  $t_k$  dalam dokumen latih pada suatu kategori  $c$  dan  $N'$  adalah jumlah total  $term$  yang terdapat pada  $c$  dokumen latih. Penambahan angka 1 berfungsi sebagai *Laplace Smoothing* (Manning, et al., 2009).

### 2.11. Evaluasi Performansi

Evaluasi performansi merupakan parameter yang berguna untuk mengukur performa dari model yang telah dibangun (Yulianto, et al., 2019). Tabel 1 merupakan tabel *confusion matrix*.

Tabel 1. *Confusion Matrix*

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

*Confusion matrix* adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan (Indriani, 2014). Parameter yang akan digunakan untuk mengukur performansi, yaitu *sensitivity* (*recall*). *Sensitivity*

adalah suatu bagian dari dokumen yang relevan yang diambil oleh sistem (Manning, et al., 2009). *Sensitivity* dipilih jika peneliti lebih memilih *False Positive* lebih baik terjadi daripada *False Negative* (Ghoneim, 2019).

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

### 3. METODE PENELITIAN

#### 3.1. Jenis dan Sumber Data

Data yang digunakan pada penelitian ini adalah data kualitatif berupa komentar pada video di Youtube. *Youtube Crawling* dilakukan pada tanggal 15 Oktober 2019 pada situs Youtube di akun resmi Kompas TV dengan judul, “Tok! Presiden Putuskan Kalimantan Timur Jadi Ibu Kota Baru Indonesia, Ini Pertimbangannya”. Terdapat 3.382 komentar yang didapatkan melalui tahapan *Youtube Crawling*. Komentar tersebut dipilih sebanyak 1.500 komentar terbaru (26 Agustus 2019 – 12 Oktober 2019) dalam kategori bahasa Indonesia.

#### 3.2. Langkah-Langkah Analisis Data

Analisis data pada penelitian ini dilakukan dengan menggunakan bantuan *software* RStudio 1.1.463 dan Microsoft Excel 2013. Adapun langkah-langkah analisis yang dilakukan adalah sebagai berikut :

1. Melakukan *Youtube Crawling*
2. Mempersiapkan data menggunakan teknik *data pre-processing*
3. Melabelkan data dengan *sentiment scoring*
4. Mengurangi dimensi data dengan *stopwords removal*
5. Memotong kata dan mentransformasi kata menjadi kata dasar menggunakan teknik *tokenizing* dan *stemming*
6. Membuat visualisasi data dengan teknik *text visualisation* dalam bentuk *word cloud*
7. Membobotkan kata menggunakan TF-IDF
8. Membangun model klasifikasi dan melakukan pengujian model
9. Mengevaluasi performansi model

### 4. HASIL DAN PEMBAHASAN

#### 4.1. *Youtube Crawling*

*Youtube Crawling* dilakukan dengan menggunakan *package tuber* pada tanggal 15 Oktober 2019 pada video yang diunggah oleh Kompas TV berjudul “Tok! Presiden Putuskan Kalimantan Timur Jadi Ibu Kota Baru Indonesia, Ini Pertimbangannya” dengan total komentar yang berhasil dikumpulkan adalah sebanyak 3382. Komentar yang sudah berhasil terkumpul akan disimpan secara otomatis dalam format CSV (*Comma Separated Value*). Komentar yang akan diolah adalah 1500 komentar terbaru (26 Agustus 2019 – 12 Oktober 2019) dengan kategori bahasa Indonesia.

#### 4.2. *Data Pre-Processing*

Proses ini membersihkan dan mentransformasi data agar siap untuk diolah.

1. *Case Folding*  
*Case folding* bertujuan untuk mentransformasi seluruh huruf pada dokumen menjadi huruf kecil (*lowercase*).
2. *Remove Mention*  
*Remove mention* digunakan untuk menghilangkan karakter “@” pada dokumen yang biasa digunakan untuk menyebutkan nama akun yang terlibat pada suatu perbincangan.

### 3. *Unescape HTML*

*Unescape HTML* merupakan tahapan yang digunakan untuk menghapus jejak HTML. HTML yang terdapat dalam dokumen pada penelitian merupakan *emoji encode* atau emoji yang mengandung karakter angka dan huruf.

### 4. *Remove Numbers*

*Remove numbers* bekerja untuk menghilangkan karakter berupa angka pada keseluruhan dokumen.

### 5. *Remove Punctuation*

*Remove punctuation* bertujuan untuk menghapus karakter tanda baca karena karakter tersebut tidak digunakan dalam analisis sentimen.

### 6. Normalisasi Kata

Normalisasi kata bertujuan untuk mentransformasi kata-kata yang tidak sesuai KBBI, seperti kata yang mengandung bahasa lokal, bahasa gaul, maupun singkatan.

### 7. *Strip Whitespace*

*Strip whitespace* digunakan untuk menghapus spasi yang berlebih di dalam dokumen komentar.

## 4.3. *Sentiment Scoring*

Pelabelan menggunakan teknik *sentiment scoring*, mendapatkan hasil bahwa kelas sentimen positif lebih dominan dibandingkan kelas sentimen negatif. Jumlah komentar yang masuk ke dalam kelas positif, yaitu sebanyak 873 komentar. Sedangkan komentar yang masuk ke dalam kelas negatif, yaitu sebanyak 627 komentar. Pelabelan menggunakan teknik *sentiment scoring* masih menimbulkan kesalahan sebanyak 78 atau 5,2% dari 1500 komentar. Kesalahan ini disebabkan karena skor akhir dari masing-masing komentar yang dihitung oleh program komputasi berbeda dengan penilaian yang diputuskan oleh manusia. Hal ini memang seringkali terjadi karena sentimen merupakan hal yang bersifat subjektif.

Kesalahan dalam pelabelan dengan *sentiment scoring* tersebut diganti dengan pelabelan manual untuk mendapatkan label yang benar pada seluruh komentar sehingga dapat memberikan performa yang optimal. Setelah dilakukan perbaikan, jumlah komentar positif menjadi sebanyak 849 dan komentar negatif sebanyak 651.

## 4.4. *Stopwords Removal*

*Stopwords removal* membutuhkan daftar kata yang akan dihapus, yakni *stoplist*. Kata-kata yang sering muncul dihilangkan untuk mengurangi dimensi data sehingga tersisa dimensi data yang diperlukan. Setelah melalui *stopwords removal*, jumlah kata pada seluruh dokumen adalah sebanyak 3288 *term* yang sebelumnya berjumlah sebanyak 4943 *term*.

## 4.5. *Tokenizing dan Stemming*

Proses *tokenizing* digunakan untuk memisahkan atau memotong kata per kata tiap komentar. *Tokenizing* juga membantu untuk memudahkan proses selanjutnya, yaitu *stemming*. *Stemming* dilakukan untuk mengubah kata-kata yang berimbuhan menjadi kata dasar.

## 4.6. *Text Vizualisation*

Visualisasi teks dalam bentuk *word cloud* hanya digunakan untuk eksplorasi secara visual yang selanjutnya akan dianalisis dengan pemodelan secara formal. Berikut merupakan hasil dari visualisasi teks menggunakan *word cloud*.



#### 4.7.2 Bernoulli Naïve Bayes

Evaluasi performansi yang digunakan adalah *confusion matrix* dengan *sensitivity* sebagai parameter. Tabel 4 merupakan hasil dari *confusion matrix*.

Tabel 4. Hasil *Confusion Matrix* Model Bernoulli

		<i>Predicted</i>	
		<i>Positive</i>	<i>Negative</i>
<i>Actual</i>	<i>Positive</i>	200	14
	<i>Negative</i>	30	56

Hasil dari perhitungan *sensitivity* yang telah diolah, didapatkan tingkat performa klasifikasi *Naïve Bayes* model Bernoulli adalah sebesar 93,45%. Hasil tersebut menyatakan bahwa sebesar 93,45% performa model klasifikasi dapat dengan tepat memprediksikan komentar dengan kelas positif dari seluruh data komentar yang memang benar masuk ke dalam kelas positif.

#### 4.7.3 Multinomial Naïve Bayes

Evaluasi performansi yang digunakan adalah *confusion matrix* dengan *sensitivity* sebagai parameter. Tabel 5 merupakan hasil dari *confusion matrix*.

Tabel 5. Hasil *Confusion Matrix* Model Multinomial

		<i>Predicted</i>	
		<i>Positive</i>	<i>Negative</i>
<i>Actual</i>	<i>Positive</i>	193	21
	<i>Negative</i>	26	60

Hasil dari perhitungan *sensitivity* yang telah diolah, didapatkan tingkat performa klasifikasi *Naïve Bayes* model Multinomial adalah sebesar 90,19%. Hasil tersebut menyatakan bahwa sebesar 90,19% performa model klasifikasi dapat dengan tepat memprediksikan komentar dengan kelas positif dari seluruh data komentar yang memang benar masuk ke dalam kelas positif.

### 5. KESIMPULAN

Masyarakat cenderung menuliskan komentar positif dalam menanggapi rencana pemindahan ibu kota berdasarkan pada pelabelan yang telah dilakukan. Pelabelan menggunakan teknik *sentiment scoring* masih menimbulkan kesalahan diakibatkan skor yang dihitung oleh program tidak sesuai dengan penilaian yang diputuskan oleh manusia. Hal ini disebabkan penentuan kelas sentimen yang melibatkan persepsi manusia. Klasifikasi menggunakan *Bernoulli Naïve Bayes* menghasilkan performa dengan tingkat sensitifitas 90,19%, sementara *Multinomial Naïve Bayes* menghasilkan performa dengan tingkat sensitifitas 93,45%. Sehingga, baik model Bernoulli maupun Multinomial memiliki performa yang baik dalam melakukan klasifikasi teks untuk kelas sentimen.

### DAFTAR PUSTAKA

- Alasadi, S. A., dan Bhaya, W. S. (2017). *Review of Data Preprocessing Techniques in Data Mining*. Journal of Engineering and Applied Science.
- Bappenas. (2019). *Dampak Ekonomi dan Skema Pembiayaan Pemindahan Ibu Kota Negara*. Bappenas: <https://www.bappenas.go.id/files/diskusi-ikn>. Diakses : 05 Februari 2020.

- Batrinca, B., dan Treleaven, P. (2014). *Social Media Analytics: A Survey of Techniques, Tools And Platforms*. AI dan Society Journal.
- BPS. (2015). *Profil Penduduk Indonesia Hasil Supas 2015*. Badan Pusat Statistika: <https://www.bps.go.id/publication/2016/11/30/63daa471092bb2cb7c1fada6/profil-penduduk-indonesia-hasil-supas-2015.html>. Diakses : 01 Februari 2020.
- Ghoneim, S. (2019). *Accuracy, Recall, Precision, F-Score dan Specificity, Which to Optimize On? Toward Data Science*: <https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124>. Diakses: 20 Februari 2020.
- Hootsuite, W. A. (2019). *Hootsuite. The Global State of Digital in 2019 Report*: <https://hootsuite.com/pages/digital-in-2019>. Diakses : 27 Januari 2020.
- Indriani, A. (2014). *Klasifikasi Data Forum Dengan Menggunakan Metode Naïve Bayes Classifier*. Seminar Nasional Aplikasi Teknologi Informasi (SNATI).
- Ling, J., Kencana, I. P. E. N. dan Oka, T. B., 2014. *Analisis Sentimen Menggunakan Metode Naive Bayes Classifier Dengan Seleksi Fitur Chi Square*. E-Jurnal Matematika, Volume 03.
- Manning, C., Raghavan, P., dan Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- McCallum, A., dan Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. Proceedings in Workshop on Learning for Text Categorization. Hal. 41-48.
- Rahman, A., Wiranto, dan Doewes, A. (2017). *Online News Classification Using Multinomial Naive*. ITSMART: Jurnal Ilmiah Teknologi dan Informasi. Volume. 06.
- Ratniasih, L. N., Sudarma, M. dan Gunantara, N., 2017. *Penerapan Text Mining dalam Spam Filtering untuk Aplikasi Chat*. Teknologi Elektro. Volume 16.
- Rozi, I. F., Pramono, S. H. dan Dahlan, E. A., 2012. *Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik*. Jurnal EECCIS. Volume 6.
- Salton, G., 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Saputra, P. Y., 2017. *Implementasi Teknik Crawling Untuk Pengumpulan Data Dari Media Sosial Twitter*. Jurnal Dinamika Dotcom. Volume 08.
- Silalahi, S. A. F., 2019. *Dampak Ekonomi dan Risiko Pemindahan Ibu Kota Negara*. Kajian Singkat Terhadap Isu Aktual dan Strategis. Volume 11.
- Tessem, B., Bjørnstad, S., Chen, W. & Nyre, L., (2015). *Word Cloud Visualisation of Locative Information*. *Journal of location Based services*, Volume 09, Hal. 254-272.
- Wahid, D. H. dan Azhari, S., 2016. *Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity*. IJCCS. Volume 10.
- Yulianto, A., Herdiani, A. dan Sardi, I. L., 2019. *Klasifikasi Keberpikahan Tweet Menggunakan Multinomial Naïve Bayes*. e-Proceeding of Engineering. Volume 06, Hal. 9078 - 9085.