Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit

Sarika Afrizal[1], Helena Nurramdhani Irmanda*[2], Noor Falih[3], Ika Nurlaili Isnainiyah[4]
Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta

Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
email:
[1]sarika.afrizal@upnvj.ac.id, [2]helenairmanda@upnvj.ac.id, [3]falih@upnvj.ac.id ,
[4]nurlailika@upnvj.ac.id
Jl. Rs. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia

Abstrak. - Kegiatan riset ini bertujuan untuk menganalisis animo masyarakat Indonesia khususnya warga Jakarta atas munculnya transportasi massa umum MRT yang di resmikan oleh Pemerintah di bulan Maret 2019. Tahapan penelitian diawali proses crawling tweet dengan menggunakan tweetscrapper dari python. Kemudian dilakukan Preprocessing sehingga didapatkan data tweet yang siap untuk diproses pada pemisahan data yaitu data training dan data testing. Data training dilakukan proses pembobotan dengan TF-IDF, dan proses pembelajaran dengan naive bayes. Proses ini disebut dengan proses training yang bertujuan untuk menghasilkan model klasfikasi. Model klasifikasi digunakan untuk data testing melakukan proses klasifikasi yang menghasilkan label sentimen (positif/negatif). Proses ini dinamakan dengan proses testing. Hasil testing akan dilakukan perhitungan akurasi dari model yang sudah dibuat. Luaran dari penelitian ini berupa analisis sentimen animo warga Jakarta pada media sosial Twitter terhadap kehadiran layanan transportasi publik MRT, dan akurasi yang dihasilkan oleh metode naïve bayes yang diimplementasikan pada analisis sentimen

Kata kunci: transportasi umum, MRT, Jakarta, Analisis Sentimen, Twitter.

1 PENDAHULUAN

Menurut hasil survei penetrasi dan perilaku pengguna Internet Indonesia di tahun 2017 yang dilakukan APJII (Asosiasi Penyelenggara Jasa Internet Indonesia), ada sekitar 143,26 juta pengguna internet di antara 262 juta populasi di Indonesia. Dan sebanyak 87,13 % menggunakan sosial media. Jumlah tersebut terus meningkat dari tahun sebelumnya. *Twitter* merupakan lima teratas media sosial yang digunakan di Indonesia [8]. Menurut *MIT Technology Review*, Indonesia menempati Negara ketiga penyumbang twit terbanyak dengan jumlah 1 milyar twit, di bawah Amerika serikat (3,7 milyar) dan Jepang (1,8 milyar). Bahkan, Jakarta menjadi kota dengan jumlah twit terbanyak dan teraktif di dunia [4]. *Twitter* seringkali dijadikan tempat untuk menyampaikan opini terhadap layanan publik maupun produk dari sebuah perusahaan[7], salah satu layanan tranportasi publik yang akan di tawarkan di Jakarta adalah *Mass Rapid Transit* (MRT).

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

Banyak warga jakarta yang menyampaikan opini mereka terhadap kehadiran MRT, baik berupa opini positif, negatif maupun netral. Apabila opini tersebut di teliti lebih lanjut maka akan dihasilkan sebuah sentimen yang dapat berguna untuk mengetahui animo warga Jakarta terhadap sebuah layanan, sehingga dapat menjadi bahan evaluasi agar dapat meningkatkan kualitas layanannya.

Analisis sentimen adalah studi yang bertujuan untuk menganalisis opini, sentimen dan emosi yang terdapat pada dokumen atau data[3]. Tugas dasar dari analisis sentimen adalah untuk mengelompokkan sifat dari teks yang ada di dalam kalimat maupun pendapat, biasanya terbagi menjadi 3 kelas yaitu negatif, positif dan netral [5].

Metode *Naïve bayes* mempunyai dua tahap proses klasifikasi teks, yaitu tahap *training* dan tahap klasifikasi. Pada tahap *training* dilakukan proses analisis terhadap sampel data berupa pemilihan kata yang mungkin muncul dalam koleksi dokumen sampel yang menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi.

Lebih konkritnya jika diasumsikan koleksi dokumen adalah D={di | i=1,2,…|D|}={d1,d2,…,d|D|} dan koleksi kategori V={vj|j=1,2,…|V|}={v1,v2,…,v|V|}. Klasifikasi NBC dilakukan dengan mencari probabilitas P(V=vj | D=di), yaitu probabilitas category vj jika diketahui dokumen di. Dokumen di dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu <a1, a2,…,an>, yang frekuensi kemunculannya diasumsikan sebagai variable random dengan distribusi probabilitas Bernoulli [2]. Selanjutnya klasifikasi dokumen adalah mencari nilai maksimum dari :

$$V_{MAP} = \text{argmax}\ P(v_j \mid a_1, a_2, ..., a_n)\ vj \varepsilon V \quad (1)$$

Teorema Bayes menyatakan tentang probabilitas bersyarat menyatakan :

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (2)$$

Dengan menerapkan teorema Bayes persamaan (1) dapat ditulis :

$$V_{MAP} = \arg\max_{vj \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j)P(v_j)}{P(a_1, a_2, ..., a_n)} \quad (3)$$

Karena nilai P(a1,a2,...,an) untuk semua vj besarnya sama maka nilainya dapat diabaikan, sehingga persamaan (3) menjadi :

$$V_{MAP} = \arg\max_{vj \in V} P(a_1, a_2, ..., a_n \mid v_j)P(v_j) \quad (4)$$

Dengan mengasumsikan bahwa setiap kata dalam <a1, a2,…,an> adalah independent, maka P(a1,a2,...,an | v j) dalam persamaan (4) dapat ditulis sebagai :

$$P(a_1, a_2, ..., a_n \mid v_j) = \prod_i P(a_i \mid v_j) \quad (5)$$

Sehingga persamaan (4) dapat ditulis :

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

$$V_{MAP} = \arg\max_{vj \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (6)$$

Nilai P(vj) ditentukan pada saat *training*, yang nilainya didekati dengan : docJ

$$P(v_j) = \frac{|doc_J|}{|Contoh|} \quad (7)$$

dimana docJ adalah banyaknya dokumen yang memiliki kategori j dalam *training*, sedangkan Contoh banyaknya dokumen dalam contoh yang digunakan untuk *training*.
Untuk nilai P(wk | v j) , yaitu probabilitas kata wk dalam kategori j ditentukan dengan :

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |vocabulary|} \quad (8)$$

Dimana nk adalah frekuensi munculnya kata wk dalam dokumen yang ber kategori vj , sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori vj, dan *vocabulary* adalah banyaknya kata dalam contoh *training*,

Beberapa penelitian membahas tentang perbandingan beberapa algoritma pada tahap *pre-processing* (*url elimination, word normalization, and stop words elimination*), pemilihan atribut (*CfsSubsetEval dan ClassifierSubsetEval*), klasifikasi (*Naïve bayes Classifier* dan *Support Vector Machine*), dengan menggunakan 949 *tweet* di kota Bandung. Hasil klasifikasi terbaik adalah 89,04%, dicapai dengan kombinasi penggunaan *word normalization* pada tahap *Preprocessing*, CfsSubsetEval pada tahap pemilihan atribut dan algoritma *Naïve bayes* pada tahap klasifikasi [1].

Penelitian lain membahas tentang perbandingan 16 teknik *pre-processing* (*Remove unicode strings and noise, Replacing URLs and user mentions, Replacing slang and abbreviations, Replacing contractions, Removing numbers, Replacing repetitions of punctuation, Replacing negations with antonyms, Removing punctuation, Handling capitalized words, Lowercasing, Removing stopwords, Replacing elongated words, Spelling correction, Part-of-Speech (POS) tagging, Lemmatization, Stemming*) dengan dua set data *Twitter* (SS *Twitter*, SemEval) untuk Analisis Sentimen, menggunakan empat algoritma machine learning, yaitu, *Linear SVC, Bernoulli Naïve bayes, Regresi Logistik, dan Convolutional Neural Networks.* Hasilnya teknik *pre-processing lemmatization, removing numbers, replacing repetitions of punctuation, dan replacing contractions*, meningkatkan tingkat akurasi [6].

Sudah banyak penelitian yang membahas menganai topik sentimen analisis dengan mengambil data sample di *Twitter* [9][10][11], namun belum ada penelitian yang membahas tentang animo warga Jakarta terhadap kehadiran layanan transportasi publik MRT dan seberapa besar akurasi yang dihasilkan oleh metode *naïve bayes* yang diimplementasikan pada analisis sentimen tersebut. Sehingga tujuan penelitian ini yaitu untuk mengetahui seberapa besar animo warga Jakarta pada media sosial *Twitter* terhadap kehadiran layanan transportasi publik MRT, dan untuk mengetahui seberapa besar akurasi yang dihasilkan oleh metode *naïve bayes* yang diimplementasikan pada analisis sentimen.

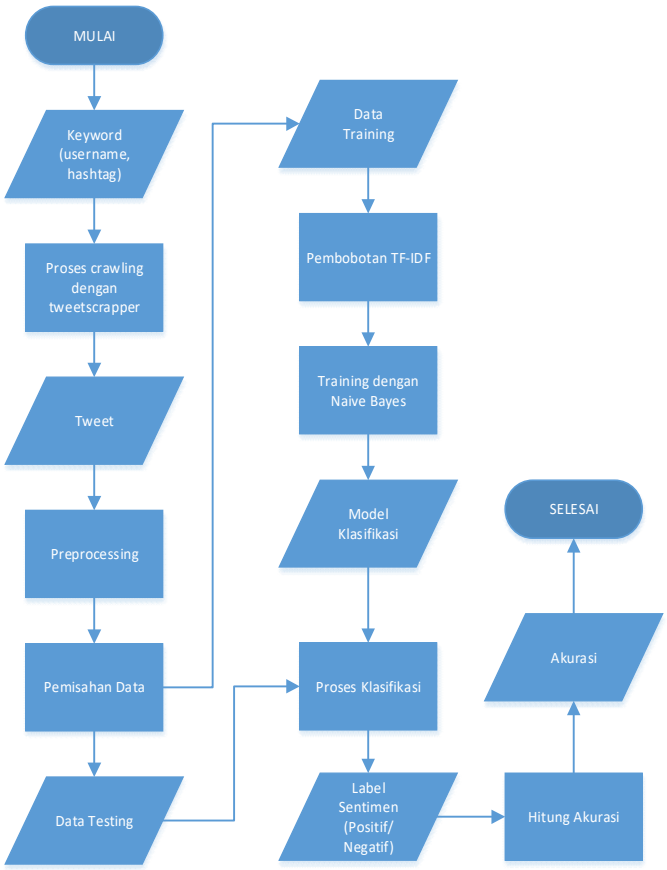
ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

2 METODOLOGI PENELITIAN

Penelitian ini merupakan jenis penelitian kuantitatif yang terdiri dari lima tahapan. Adapun tahapan-tahapannya sebagai berikut:

1. Perumusan Masalah dan Tujuan : Pada tahap ini dilakukan observasi terhadap pengaduan dan tanggapan masyarakat terhadap MRT jakarta melalui *Twitter*.
2. Pengumpulan Data dan Informasi : Pada tahapan ini dilakukan studi literatur mengenai keilmuan mengenai sentimen analisis dan pengolahan data.
3. Pengolahan Data : Pada tahap ini dilakukan proses crawling *Twitter* mengenai keyword MRTJakarta. Selanjutnya, dilakukan proses pembelajaran untuk mendapat model klasifikasi yang akan digunakan pada tahap *testing* data.
4. Analisis Data :Pada tahap ini dilakukan analisis data dari hasil pengolahan data untuk menghasilkan label sentimen dan akurasi hasil *testing*.
5. Kesimpulan dan Saran : Tahapanan ini bertujuan untuk mendokumentasikan keseluruhan hasil penelitian dan penarikan kesimpulan serta saran yang dapat diberikan untuk penelitian selanjutnya.

Pada penelitian ini akan dibangun suatu sistem untuk melakukan sentiment analysis untuk melihat animo masyarakat terhadap kehadiran *Mass Rapid Transit* Jakarta MRT Jakarta. Sistem ini terdiri dari proses crawling, pelabelan data, *Preprocessing*, proses klasifikasi dan evaluasi. Flowchart untuk rancangan sistem yang akan dibangun dapat dilihat pada Gambar 1.



Gambar 1. Metode Penelitian

Input dari sistem ini adalah keyword yang berupa *username* @mrtjakarta maupun hashtag mengenai MRT Jakarta #mrtjakarta, tahapan selanjutnya adalah proses crawling *tweet* dengan menggunakan *tweet*scrapper dari python. Kemudian dilakukan *Preprocessing* sehingga didapatkan data *tweet* yang siap untuk diproses pada pemisahan data yaitu data *training* dan data *testing*. Data *training* dilakukan proses pembobotan dengan TF-IDF, dan proses pembelajaran dengan naive bayes. Proses ini disebut dengan proses *training* yang bertujuan untuk menghasilkan model klasfikasi. Model klasifikasi digunakan untuk data *testing* melakukan proses klasifikasi yang menghasilkan label sentimen (positif/negatif). Proses ini dinamakan dengan proses *testing*. Proses perhitungan akurasi pada data *testing* dilakukan dengan confussion matrix dari model yang sudah dibuat. nilai akurasi merupakan hasil bagi dari jumlah data *testing* yang benar yang terdiri dari *true positive* (TP) dan *true negative* (TN) dengan jumlah data *testing* keseluruhan.

$$Akurasi = \frac{TP + TN}{TP + TN + FN + FP} \quad (9)$$

Keterangan
TP: *True positive*
TN: *True negative*
FP: *False positive*
FN: *False negative*

Selain menggunakan akurasi, penelitian ini akan menggunakan *sensitivity* dan *specificity.* Sensitivity dihitung dari jumlah *True positive* (prediksi positif yang benar) dibagi dengan keseluruhan kelas positif. Sedangkan *specifity* dihitung dari keseluruhan true negatif dibagi dengan keseluruhan jumlah kelas yang salah. [12].

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

$$SP = \frac{TN}{TN + FP} \quad (11)$$

3 HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data (Crawling dan Scraping)

Data didapatkan dari sosial media yaitu *Twitter* dengan keyword “MRTJakarta” selama masa uji coba public MRT yaitu dari tanggal 5 – 23 maret 2019. *Tweet* yang diambil sebanyak 1000 *tweet*. Dari *tweet* yang terkumpul dilakukan pelabelan dengan 5 anotator. Tediri dari 2 jenis label / kelas yaitu positif dan negatif. Setiap anotator melabeli 1000 *tweet*. Sehingga terdapat 5 label untuk setiap *tweet*. Cara pemilihan label yang digunakan yaitu dengan mengambil label yang merupakan mayoritas label yang diberikan oleh Anotator (*voting*). Hasil dari pengumpulan data digambarkan pada Tabel 1.

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

Tabel 1. Sampel pelabelan data

No	*Tweet*	Kelas
1	Baru jg jadi MRT, sdh ada demo, moga gak dirusak massa ya	Negatif
2	Mim tolong ditindak ya kalo ada yg mau merusak fasilitas. Soalnya saya blm nyobain MRT.	Positif
3	Salut min buat mrt yang tetap bisa beroperasi sampai Bundaran HI semangat min	Positif
4	Semoga Jkt segera kondusif aman selamat sehat makmur sejahtera semuanya.	Positif
5	berarti masih beroperasi ya sampai Bundaran HI, thanks	Positif
6	sejak kemarin sore arah ke stasiun MRT Bundaran HI ditutup meskipun untuk pejalan kaki, saya dihimbau naik melalui stasiun dukuh atas. semoga sekarang sudah beroperasi kembali	Positif
7	Td dr Bunderan HI ke arah Lebak Bulus masih dibuka...	Positif
8	Stay safe yaa. Aku khawatir bgt sama kamu MRT	Positif
9	Yang HI ko tutup	Negatif
10	Masih tuuh, nyaman2 aja mrt 8.15 pic.twitter.com/dOIDq8yhSY	Positif
11	Oke terima kasih infonya. Semoga situasi selalu kondusif supaya bisa tetap beroperasi normal	Positif
12	Salut tetap bertugas. Semoga aman dan selamat semuanya.	Positif

Berdasarkan hasil *crawling* dan pelabelan manual diperoleh sentiment positif sebanyak 585 *tweet* dan sentiment negatif sebanyak 415 *tweet*.

3.2. Data Cleansing / *Preprocessing*

Setelah memberikan label pada setiap *tweet* dilakukan *Preprocessing tweet*, pada penelitian ini dilakukan berapa proses yaitu:

1. Mengubah huruf menjadi huruf kecil, menghilangkan tanda baca (punctuation) dan *username*.
   Contoh :
   Input teks : Bahagia banget hari ini merasakan uji coba @mrtjakarta intinya perjalanan akan semakin cepat dan nyaman sampai tujuan,,, https?://www.pic.*Twitter*.com/zwpH2Fi7xi
   Output : bahagia banget merasakan uji coba intinya perjalanan cepat nyaman tujuan https www pic *Twitter* com zwph2fi7xi
2. Menghilangkan situs website pada *tweet*.
   Contoh:
   Input teks : bahagia banget merasakan uji coba intinya perjalanan cepat nyaman tujuan https www pic *Twitter* com zwph2fi7xi
   Output : bahagia banget hari ini merasakan uji coba intinya perjalanan akan semakin cepat dan nyaman sampai tujuan
3. Menghilangkan stopword

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

Fungsi ini bertujuan untuk menghilangkan kata yang dirasa tidak penting pada sistem tersebut. Seperti kata yang sering keluar atau imbuhan yang dianggap tidak memiliki makna. Contoh stopwords : kami, aku, kalau,dan lain lain.
Contoh:
Input teks : bahagia banget hari ini merasakan uji coba intinya perjalanan akan semakin cepat dan nyaman sampai tujuan
Output : bahagia banget merasakan uji coba intinya perjalanan cepat nyaman tujuan.

3.3. Stemming

Fungsi ini bertujuan untuk menghilangkan imbuhan sehingga didapatkan kata dasar dari kata berimbuhan tersebut.
Contoh:
Input teks : bahagia banget merasakan uji coba intinya perjalanan cepat nyaman tujuan
Output : bahagia banget hari ini rasa uji coba inti jalan akan makin cepat dan nyaman sampai tuju.
Dari proses Data Cleansing ini didapatkan output yang digambarkan pada tabel 2.

Tabel 2. Sampel data yang sudah dibersihkan

Row	Kelas	clean_text
0	Negatif	jg mrt sdh demo moga gak rusak massa
1	Positif	mim tolong tindak kalo yg rusak fasilitas blm ...
2	Positif	salut min mrt operasi bundar hi semangat min
3	Positif	moga jkt kondusif aman selamat sehat makmur se...
4	Positif	operasi bundar hi thanks
5	Positif	kemarin sore arah stasiun mrt bundar hi tutup ...
6	Positif	td dr bunderan hi arah lebak bulus buka
7	Positif	stay safe yaa khawatir bgt mrt
8	Negatif	hi ko tutup
9	Positif	tuuh nyaman2 aja mrt 15 pic twitter com doidq8...
10	Positif	oke terima kasih info moga situasi kondusif op...

3.4. Vektorisasi

Vektorisasi bertujuan untuk mengubah teks dalam *tweet* menjadi angka yang bisa diolah dan dicari polanya pada saat proses klasifikasi. Terdapat dua proses dalam vektorisasi yaitu tokenisasi dan pembobotan setiap kata dengan metode TD/IDF.

1. Melakukan tokenisasi
   Fungsi ini bertujuan untuk mengubah kalimat menjadi kata/token token. Contoh:
   Input teks : Bahagia banget hari ini
   Output : {Bahagia, banget, hari, ini}
2. Melakukan perhitungan TF/IDF
   Metode Term Frequency-Inverse Document Frequency (TF-IDF) merupakan suatu metode yang bertujuan memberi bobot hubungan suatu kata (term) terhadap dokumen/*tweet*. TF-IDF serta mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen. Perhitungan TF -IDF menggunakan *library* di Sklearn python yaitu TfidfVectorizer().

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

Contoh :
Input teks :
documentA ="salut mrt tetap operasi bundar semangat"
documentB = "moga jakarta segera kondusif aman selamat sehat makmur sejahtera semua arti operasi bundar thanks"
documentC = "kemarin sore arah stasiun mrt bundar tutup jalan kaki dihimbau naik stasiun dukuh moga operasi kembali"
Ouput :

Tabel 3. Sampel data yang sudah dibersihkan

	aman	arah	arti	bundar	dihimbau	dukuh	jakarta
0	0.00	0.00	0.00	0.29	0.00	0.00	0.00
1	0.29	0.00	0.29	0.17	0.00	0.00	0.29
2	0.00	0.25	0.00	0.15	0.25	0.25	0.00

3.5. Proses Klasifikasi

Proses klasifikasi dilakukan pada traning set dengan tujuan untuk membuat model, dan kemudian model ini digunakan untuk *testing*. Tahapan yang pertama yaitu memisahkan data *training* dan data *testing*. Sebesar 80% untuk *training* (800 *tweet*) dan 20% untuk data *testing* (200 *tweet*). Setelah melakukan pemisahan dilakukan proses klasifikasi dengan menggunakan *naïve bayes*. Pada penelitian ini proses klasifikasi menggunakan *library naïve bayes scikit learn* python. Terdapat tiga jenis distribusi yang umum digunakan pada *scikit learn*, yaitu Bernoulli, Multinomial, dan Gaussian. Ketiga jenis distribusi tersebut disesuaikan dengan tipe data yang ditangani. Distribusi yang digunakan adalah Naïve Bayers Bernoulli, karena NB Bernoulli ini cocok untuk tipe data boolean dimana dalam penelitian ini label hanya terdiri dari 2 kelas yaitu positif dan negatif. Berikut hasil *training* ditunjukan pada gambar 2.

```
Positive Positive
Positive Positive
Positive Positive
Negative Positive
Positive Positive
Positive Negative
Negative Positive
Positive Positive
Positive Negative
Positive Positive
Negative Positive
Positive Negative
Positive Positive
Negative Positive
Negative Positive
Positive Positive
Positive Positive
Negative Positive
Positive Positive
```

Gambar 2. Sampel hasil *training* data

Output dari proses *training* ini yaitu model yang dapat dilihat pada gambar 3. Model ini akan disimpan dan digunakan pada proses selanjutnya yaitu *testing* sehingga dapat disimpulan akurasinya.

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

```
In [25]: hasil=model.predict(x_train)

In [26]: sentiment = []
n = 0
for i in hasil:
  n = n + 1
  if i == "baik":
    a = 'Positive'
  else:
    a = 'Negative'
  sentiment.append(a)

sentiment_sebenarnya = []
for i in y_test:
  if i == "baik":
    b = 'Positive'
  else:
    b = 'Negative'
  sentiment_sebenarnya.append(b)


i = 1
while i < n:
  print(sentiment[i]+ " " + sentiment_sebenarnya[i])
  i += 1
```

Gambar 3.*Source code model*

3.5. Evaluasi

Model yang didapatkan dari hasil *training* digunakan untuk *testing* dengan jumlah data *testing* sebanyak 200 *tweet* (125 data *tweet* positif 75 data *tweet* negatif). Berikut hasil *training* ditunjukan pada gambar 4.

```
Prediksi Sebenarnya
Negative Negative
Positive Positive
Positive Negative
Positive Positive
Positive Negative
Positive Positive
Negative Negative
Positive Negative
Negative Negative
Negative Negative
```

Gambar 4. Sampel hasil *testing* data

Dari data testing Hasil evaluasi digambarkan dalam *confussion matrix* pada Tabel 4

Tabel 3. Sampel data yang sudah dibersihkan

Pred **Actual**	**Negatif**	**Positif**
Negatif	39	36
Positif	13	112

Bedasarkan Confussion matrix Tabel 4 diperoleh 23 *tweet* sentiment negatif yang diprediksi tepat, 75 sentimen positif yang diprediksi tepat, 59 *tweet* bersentimen positif yang diprediksi menjadi *tweet* bersentimen negatif, dan 43 *tweet* dengan sentiment negatif namun diprediksi oleh classifier sebagai *tweet* bersentimen positif. Dari confussion matrix tersebut dapat dihitung nilai akurasinya yaitu sebesar 75%, sensitivity sebesar 90%, specificity sebesar 52%. Pengukuran nilai akurasi, sensitivity, dan specificity bertujuan untuk mengetahui kemampuan
model dalam mengidentifikasi sentiment pada tweet terkait MRT. Nilai sensitivity sebesar 90% berarti kemampuan model untuk meberikan hasil prediksi berupa

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

sentimen positif bagi tweet terkait MRT adalah sebesar 90%. Sementara itu nilai Nilai specificity sebesar 52% berarti kemampuan model untuk meberikan hasil prediksi berupa sentimen negatif bagi tweet terkait MRT adalah sebesar 52%. Penelitian ini menunjukan hasil akurasi yang cukup baik, sensitivity yang baik, namun specificity yang masih rendah. Hal ini dikarenakan data tweet masih banyak mengandung spam dan terdapat kata-kata *slang* dan singkatan. Banyaknya spam juga membuat, jumlah data tweet yang digunakan untuk training dan testing belum memadai.

4 KESIMPULAN

Penelitian Implementasi Metode Naïve Bayes Untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit (MRT) menggunakan data dari sosial media yaitu *Twitter* dengan keyword "MRTJakarta" yang dilakukan selama masa uji coba public MRT yaitu dari tanggal 5 – 23 maret 2019. *Tweet* yang diambil sebanyak 1000 *tweet* (800 *tweet* untuk *training* dan 200 *tweet* untuk *testing*).

Dalam peneliltian ini naive bayes dapat memprediksi sentimen dari *tweet* yang sudah dikumpulkan terkait animo masyarakat terhadap MRTJakarta dengan akurasi sebesar 75%. Saran untuk penelitian berikutnya yaitu menambah data latih dan data uji, serta menggunakan teknik praproses untuk mengubah kata slang dan singkatan menjadi kata baku, dengan demikian diharapkan nilai akurasi dari model yang telah dibuat akan lebih baik..

Ucapan Terimakasih

Penulis menyampaikan ucapan terima kasih kepada LPPM Universitas Pembangunan Nasional Veteran Jakarta yang telah mendanai penelitian ini melalui skim Penelitian Dosen Pemula.

REFERENSI

[1] Akbarisanto, R., Danar, W., & Purwarianti, A. (2016). Analyzing Bandung Public Mood Using *Twitter* Data, 4(c).

[2] Mccallum, A., & Nigam, K. (1997). A Comparison of Event Models for Naive Bayes Text Classification.

[3] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications : A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[4] https://thenextweb.com/asia/2016/07/01/study-shows-37-of-tweets-originate-from-asia-is-that-right/, 2016. Study shows 37% of tweets originate from Asia. Is that right? [diakses tanggal 24 januari 2019]

[5] Avanco, L. V., & Nunes, M. G. (2014). Lexicon-based Sentiment Analysis for Reviews of Products in Bazillian Portuguese. Brazillian Conference on Intelligent System, 277-281.

[6] Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of *pre-processing* techniques and their interactions for *Twitter* sentiment analysis. Expert Systems With Applications, 110, 298–310. https://doi.org/10.1016/j.eswa.2018.06.022

[7] Pratama, I. D. (2017). Bahasa Komplain di Media Sosial Twitter. Transformatika: Jurnal Bahasa, Sastra, dan Pengajarannya, 1(1), 35-56.

[8] https://apjii.or.id/survei2017, 2018. Penetrasi & perilaku pengguna internet Indonesia [diakses tanggal 24 januari 2019]

[9] A.Jabbar Alkubaisi, G. A., Kamaruddin, S. S., & Husni, H. (2018). Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. *Computer and Information Science*, *11*(1), 52. https://doi.org/10.5539/cis.v11n1p52

[10] Kunal, S., Saha, A., Varma, A., & Tiwari, V. (2018). Textual Dissection of Live Twitter Reviews using Naive Bayes. *Procedia Computer Science*, *132*(Iccids), 307–313. https://doi.org/10.1016/j.procs.2018.05.182.

[11] Clark, E. M., James, T., Jones, C. A., Alapati, A., Ukandu, P., Danforth, C. M., & Dodds, P. S. (2018). *A Sentiment Analysis of Breast Cancer Treatment Experiences and Healthcare Perceptions Across Twitter*. Retrieved from http://arxiv.org/abs/1805.09959.

[12] Adinugroho, S., & Sari, Y. A. (2018). Implementasi Data Mining Menggunakan Weka. Universitas Brawijaya Press.

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)

ISSN : 2655-139X (ONLINE)
ISSN : 0216-4221 (PRINT)