TUGAS AKHIR

ANALISIS SENTIMEN TWITTER MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* DENGAN SELEKSI FITUR *MUTUAL INFORMATION*

Oleh:

Maria Arista Ulfa
F1D 013 060

Telah diperiksa dan disetujui oleh Tim Pembimbing

1. Pembimbing Utama

Dr. Eng. Budi Irmawati, S.Kom., M.T.
NIP. 19721019 199903 2 001

Tanggal : 30 April 2018

2. Pembimbing Pendamping

Ario Yudo Husodo, S.T., M.T.
NIP.19901218 201212 1 002

Tanggal : 30 April 2018

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Teknik
Universitas Mataram

Dr. Eng. Budi Irmawati, S.Kom., M.T.
NIP. 19721019 199903 2 001

TUGAS AKHIR

ANALISIS SENTIMEN TWITTER MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* DENGAN SELEKSI FITUR *MUTUAL INFORMATION*

Oleh:

Maria Arista Ulfa
F1D 013 060

Telah dipertahankan di depan Dewan Penguji
Pada tanggal 28 Februari 2018
dan dinyatakan telah memenuhi syarat mencapai derajat Sarjana S – 1
Program Studi Teknik Informatika

Ususan Tim Penguji

1. Penguji I

Ida Bagus Ketut Widiartha, ST., M.T.
NIP. 19700514 199903 1 002

Tanggal : 27 April 2018

2. Penguji II

Royana Afwani, S.T., M.T.
NIP: 19850707 201404 2 001

Tanggal : 26 April 2018

3. Penguji III

Fitri Bimantoro, S.T., M.Kom.
NIP: 19860622 201504 1 001

Tanggal : 24 April 2018

Mataram, April 2018
Dekan Fakultas Teknik
Universitas Mataram

Akmaluddin, S.T., M.Sc.,(Eng)., Ph.D.
NIP. 19681231 199412 1 001

INTISARI

Analisis sentimen merupakan suatu teknik idetifikasi terhadap emosi yang diekspresikan melalui teks. Tujuan analisis sentimen adalah menentukan apakah suatu pendapat dalam kalimat atau dokumen termasuk kategori positif atau negatif. *Twitter* merupakan salah satu media sosial yang sering digunakan dalam menyampaikan pendapat. *Twitter* memungkinkan penggunanya (*user*) untuk menulis pendapat mereka mengenai berbagai topik dalam sebuah *tweet*. Data *twitter* dalam penelitian ini di*download* melalui *twitter Application Programming Interface* (API). Data *twitter* tersebut terdiri dari 500 *tweet* tentang pariwisata Lombok dengan *hashtag* #lombok dan #woderfullombok. Fitur informasi dari setiap tweet diseleksi menggunakan metode *Mutual Information* dan dianalisis menggunakan model klasifikasi *Naïve Bayes* (*Naïve Bayes Classifier*). Hasil pengujian klasifikasi sentimen *twitter* pada kategori positif dan negatif menggunakan 10-*fold cross validation* memperoleh akurasi rata-rata sebesar 97,9%.

Kata kunci : *Analisis Sentimen, Twitter, Naïve Bayes Classifier, Mutual Information*

ABSTRACT

Sentiment analysis is an identification technique of emotion expressed in texts. The sentiment analysis goal is to determine a negative or positive opinion within a sentence or a document. Twitter is one of social medias to convey an opinion. The twitter allows its users to write opinions related to a specific topic in a tweet. The twitter data used in this research was downloaded using the twitter Application Programming Interface (API). It consisted 500 tweets about Lombok tourism that contained #lombok and #woderfullombok hashtags. The features extracted from the twitter data were selected using the Mutual Information (MI) method then they were analyzed using the Naïve Bayes Classifier (NBC) model. The evaluation of sentiment analysis on the Lombok tourism twitter data in a 10-folds cross validation resulted 97.9% accuracy.
Key words : *Sentiment Analysis, Twitter, Naïve Bayes Classifier, Mutual Information.*

Analisis Sentimen Twitter Menggunakan *Naïve Bayes Classifier* dengan Seleksi Fitur *Mutual Information*

Twitter Sentiment Analysis Using Naïve Bayes Classifier with Mutual Information Feature Selection

Maria Arista Ulfa[1], Budi Irmawati[1], Ario Yudo Husodo[1]
[1]Program Studi Teknik Informatika Fakultas Teknik Universitas Mataram
Jl. Majapahit No. 62, Mataram, Lombok NTB, INDONESIA
Email: maria.arista95@gmail.com, *Email:* budi-i@unram.ac.id , *Email:* ario@ti.ftunram.ac.id

Abstract **Sentiment Analysis is an identification technique of emotion expressed in texts. The sentiment analysis goal is to determine a negative or positive opinion within a sentence or a document. Twitter is one of social medias to convey an opinion. The twitter allows its users to write opinions related to a specific topic in a tweet. The twitter data used in this research was downloaded using the twitter Application Programming Interface (API). It consisted 500 tweets about Lombok tourism that contained #lombok and #woderfullombok hashtags. The features extracted from the twitter data were selected using the Mutual Information (MI) method then they were analyzed using the Naïve Bayes Classifier (NBC) model. The evaluation of sentiment analysis on the Lombok tourism twitter data in a 10-folds cross validation resulted 97.9% accuracy.**

Key words**: Sentiment Analysis, Twitter, Naïve Bayes Classifier, Mutual Information.**

I. PENDAHULUAN

Analisis sentimen adalah studi komputasi dari opini dan emosi yang diekspresikan dalam teks. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif. Analisis sentimen banyak digunakan untuk mengetahui opini masyarakat terhadap suatu produk, layanan ataupun isu politik [1].

Salah satu media sosial yang digunakan untuk menyampaikan opini yaitu twitter. Twitter merupakan media komunikasi yang memungkinkan penggunanya untuk berbagi status kepada pengguna lain dengan jumlah *tweet* maksimal 140 karakter. Data twitter dapat diambil melalui twitter *Application Programming Interface* (API).

Analisis data twitter tidak dapat diproses secara langsung seperti data teks pada umumnya karena banyaknya penggunaan bahasa yang tidak normal akibat keterbatasan jumlah karakter yang dapat ditulis untuk setiap *tweet*. Selain itu, twitter belum mampu mengekstraksi *tweet* menjadi sebuah kesimpulan apakah termasuk kedalam opini positif atau negatif. Oleh karena itu, klasifikasi perlu dilakukan dalam menganalisis suatu opini. Salah satu model yang digunakan dalam klasifikasi yaitu *Naïve Bayes Classifier* (NBC). NBC adalah model klasifikasi yang menerapkan Teorema Bayes di mana model ini sangat cepat dalam pelatihan dan sederhana. Walaupun merupakan model yang sederhana namun NBC mampu menghasilkan akurasi yang cukup tinggi. Kemudahan implementasi juga merupakan keuntungan besar dari NBC [2]. Namun, klasifikasi teks memiliki masalah terhadap ribuan fitur sehingga akan digunakan *Mutual Information* (MI) sebagai penyeleksi fiturnya.

II. TINJAUAN PUSTAKA

Utami [3] menggunakan *Support Vector Machine* dan *K-Nearets Neighbor* dengan seleksi fitur *Particle Swarm Optimization* dalam menentukan sentimen publik mengenai berita kebakaran hutan dengan jumlah data sebanyak 360 data. Dari hasil penelitian tersebut diperoleh nilai akurasi untuk SVM sebesar 86,11% dan nilai akurasi untuk K-NN sebesar 73,06%.

Ernawati [4] menggunakan *Naïve Bayes Classifier* dengan seleksi fitur *Particle Swarm Optimization Square* dalam menentukan sentimen terhadap *review* penjualan *online* sebuah restoran dengan banyak data sebanyak 400 data. Dari hasil penelitian tersebut diperoleh nilai akurasi sebesar 86,88%.

Darma dkk. [5] menggunakan *Support Vector Machine* dalam menganalisis acara televisi dengan penambahan seleksi fitur Algoritma Genetika dengan banyak data sebanyak 160 *tweet* yang terbagi atas 80 *tweet* positif dan 80 *tweet* negatif. Dari hasil penelitian tersebut diperoleh nilai akurasi sebesar 90,50%.

Gupta dan Parveen [6] menggunakan *Naïve Bayes Classifier* dengan seleksi fitur TF-IDF (*Term Frequency-Inverse Document Frequency)* dan penambahan optimalisasi fitur *Gain Ratio* dalam menentukan sentimen terhadap *review* film dengan banyak data sebesar 500 data. Dari hasil penelitian tersebut diperoleh nilai akurasi sebesar 94%.

Dari beberapa penelitian yang telah dilakukan sebelumnya didapatkan hasil akurasi yang berbeda-beda tergantung dari model klasifikasi dan metode seleksi fitur yang digunakan di mana pada penelitian yang dilakukan oleh Gupta dan Parveen dengan menggunakan model klasifikasi *Naïve Bayes Classifier* dengan seleksi fitur TF-IDF dan penambahan optimalisasi fitur *Gain Ratio* memperoleh hasil yang cukup tinggi daripada yang lain. Hal ini terjadi karena adanya penambahan optimalisasi fitur *Gain Ratio* yang membantu meningkatkan akurasi. Tanpa adanya penambahan optimalisasi fitur, akurasi yang dihasilkan hanya sebesar 78%. Berdasarkan model klasifikasi SVM, K-NN dan NBC dengan penambahan seleksi fitur tanpa adanya optimalisasi fitur maka dapat dilihat bahwa pada penelitian yang dilakukan oleh Darma dkk., SVM memperoleh nilai akurasi yang cukup tinggi.

SVM pada prinsipnya bekerja baik pada kasus klasifikasi biner dan cocok digunakan pada dataset yang berdimensi tinggi namun SVM memiliki waktu komputasi yang cukup lama. SVM sensitif terhadap dataset yang tidak seimbang [7]. Salah satu alasan menurunnya kinerja SVM pada dataset yang tidak seimbang adalah munculnya bias pada saat menemukan *hyperplane* dengan *soft margin* di mana hasilnya akan lebih cenderung mengarah kepada kelas mayoritas [8].

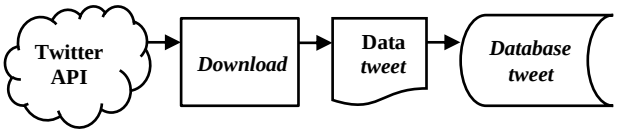
Selain model klasifikasi SVM, model klasifikasi NBC menunjukkan nilai akurasi yang cukup baik. Walaupun modelnya sederhana namun sudah terbukti efektif, cepat dan menghasilkan akurasi yang cukup tinggi dalam penelitian mengenai analisis sentimen. NBC juga bekerja cukup baik dalam mengklasifikasikan data pada kelas tidak seimbang [9]. Oleh karena itu dalam penelitian kali ini akan digunakan NBC sebagai model klasifikasi.

Pada proses seleksi fitur dalam penelitian sebelumnya ada beberapa metode yang digunakan yaitu PSO, AG dan TF-IDF. Dari ketiga metode ini, fitur yang dilibatkan hanyalah fitur yang hadir atau yang muncul di satu kelas tanpa mempertimbangkan kemunculannya di kelas lain. Hal ini kadang menghasilkan fitur yang kurang tepat untuk suatu kelas. Oleh karena itu digunakan metode *Mutual Information* (MI) di mana setiap fitur akan diukur kehadiran dan ketidakhadirannya untuk menghasilkan fitur yang tepat pada suatu kategori atau kelas.

III. METODE PENELITIAN

A. Pengumpulan Data

Pengambilan data twitter diawali dengan penarikan data menggunakan fasilitas *Application Programming Interface* (API) yang telah disediakan oleh twitter. Data tersebut diambil sesuai dengan jumlah kebutuhan yaitu 500 data. Data yang diambil berupa kumpulan *tweet* pengguna yang kemudian disimpan ke dalam *database* (format txt). Proses pengambilan data twitter ditunjukkan pada Gambar 1.



Gambar 1. Tahap Pengambilan Data Twitter

B. *Pre-processing*

Pre-processing merupakan proses pengolahan data awal dengan cara mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem.

Contoh *tweet* :

OMG!! Lombok is my favorite and most beautiful 6 beaches :* #lombok #beaches #gili #indonesia #travel #ttot https://t.co/xjuRrO5fGK

1. *Cleaning*

Membersihkan kata-kata yang tidak diperlukan untuk mengurangi derau. Kata yang dihilangkan adalah URL, *hashtag (#)*, *username* (@), *retweet* (RT) dan email.

OMG!! Lombok is my favorite and most beautiful 6 beaches :*

2. *Convert Emoticon*

Merubah simbol *emoticon* pada *tweet* dengan kata yang mencerminkan emotikon tersebut. Daftar emotikon dapat dilihat pada Tabel 1.

TABEL I. Konversi Emotikon

Emotikon	Makna
>:] ;) :-) :) :o) :] :3 :c) :> =] 8) =) :} :^) (: ;-) >:D :-D :D 8D x-D xD ;* :* =-D =D =-3 =3 ;D :P :-P <3	*happy*
>:[:-(:(:-c :c :-< :< :-[:[:{ >.> <.< >.< =(>:\ >:/ :-/ :/ :\ =/ =\ :S :s >:o :O :-O 8-0	*bad*

OMG!! Lombok is my favorite and most beautiful 6 beaches happy

3. *Case Folding*

Merubah kata-kata agar menjadi huruf kecil (*lowercase*).

omg!! lombok is my favorite and most beautiful 6 beaches happy

4. *Tokenization*

Memisahkan semua kata beradsarkan spasi dan menghilangkan tanda baca maupun angka.

omg	***lombok***	***is***	***my***	***favorite***	***and***
most	***beautiful***	***beaches***	***happy***		

5. *Slang Lookup*

Merubah kata singkatan atau kata gaul ke dalam bentuk standar.

Oh	***my***	***god***	***lombok***	***is***	***my***
favorite	***and***	***most***	***beautiful***	***beaches***	
happy					

6. *Stopword Removal*

Menghapus kata-kata yang tidak memiliki makna jika dipisahkan dari kata yang lain dan biasanya berupa kata ganti atau kata sambung.

god	***favorite***	***beautiful***	***beaches***	***happy***

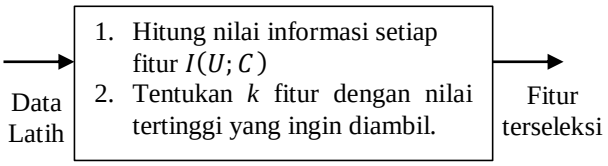
7. *Stemming*

Mengembalikan kata ke dalam bentuk dasar (*root word*). *Stemming* dilakukan berdasarkan aturan morfologi bahasa Inggris.

god	***favorite***	***beautiful***	***beach***	***happy***

C. Seleksi Fitur *Mutual information* (MI)

Seleksi fitur dapat membuat proses klasifikasi menjadi lebih efisien dan efektif dengan mengurangi jumlah data yang dianalisis. *Mutual Information* (MI) merupakan salah satu metode seleksi yang menunjukkan seberapa banyak informasi ada atau tidaknya sebuah *term* memberikan kontribusi dalam membuat keputusan klasifikasi secara benar atau salah [10]. Proses seleksi fitur dengan MI dapat dilihat pada Gambar 2.



Gambar 2. Tahap Seleksi MI

Perhitungan nilai MI untuk setiap fitur disimbolkan dengan notasi *I* pada persamaan 1.

$$I(U;C) = \frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_1.N_{.1}} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_0.N_{.1}} + \frac{N_{10}}{N} log_2 \frac{NN_{10}}{N_1.N_{.0}} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_0.N_{.0}} \quad (1)$$

di mana :

N = Jumlah dokumen yang memiliki e_t dan e_c atau (*N = N00 + N01 + N10 + N11).*

N_1. = Jumlah dokumen yang memiliki e_t atau (*N_1. = N10 + N11).*

$N_{.1}$ = Jumlah dokumen yang memiliki e_c atau (*$N_{.1}$ = N01 + N11).*

N_0. = Jumlah dokumen yang tidak memiliki e_t atau (*N_0.= N01 + N00).*

$N_{.0}$ = Jumlah dokumen yang tidak memiliki e_c atau (*$N_{.0}$ = N10 + N00).*

Contoh *tweet* sebelum *pre-processing* dapat dilihat pada Tabel 2.

TABEL II Tabel *Tweet*

Tweet	**Kelas**
OMG!! Lombok is my favorite and most beautiful 6 beaches : #lombok #beaches #gili #indonesia #travel #ttot https://t.co/xjuRrO5fGK*	*Positive*
All you need is love, happy and beach! : #gilitrawangan #lombok #island*	*Positive*
Love this place #gilitrawangan #lombok #ntb https://t.co/l9Xn5Nkv3x	*Positive*
so bad !! I found a lot of rubbish on #senggigi #savesenggigi #beach #dirty	*Negative*
Over Troubled Water :/ #bnw #bw #slowshutter #slowshutterspeed https://t.co/0IBayn	*Negative*

Contoh *tweet* setelah *pre-processing* dapat dilihat pada Tabel 3.

Tabel III. Tabel Kemunculan Fitur

Tweet	**Fitur (Kemunculan)**	**Kelas**
T1	*god*[1], *favorite* [1], *beautiful*[1], *beach*[1], *happy*[1]	*Positive*
T2	*love*[1], *beach*[1], *happy*[2]	*Positive*
T3	*love*[1]	*Positive*
T4	*bad*[1], *rubbish*[1]	*Negative*
T5	*trouble*[1], *water*[1], *bad*[1]	*Negative*

Penerapan perhitungan MI pada fitur adalah sebagai berikut :

	$e_c = 1$	$e_c = 0$
$e_t = 1$	2	0
$e_t = 0$	1	2

$$I_{happy}(U;C) = \frac{2}{5}\log_2 \frac{2.5}{(1+2).(2+0)} + \frac{1}{5}\log_2 \frac{1.5}{(2+1).(1+2)} + \frac{0}{5}\log_2 \frac{0.5}{(0+2).(2+0)} + \frac{2}{5}\log_2 \frac{2.5}{(1+2).(2+0)} = 0{,}41997$$

Perhitungan fitur dengan menggunakan persamaan 1 akan dilakukan untuk semua fitur pada setiap kelas. Hasil seleksi nilai MI semua fitur dapat dilihat pada Tabel 4.

TABEL IV. Tabel Seleksi Fitur

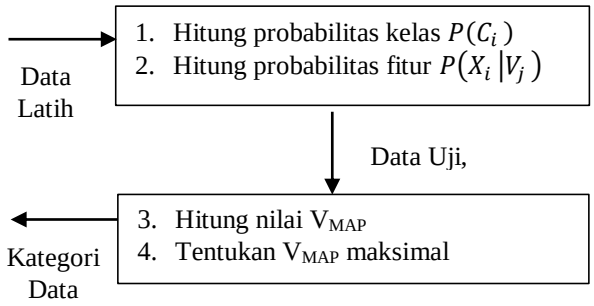
Fitur	**MI**
god	0,17095
favorite	0,17095
beautiful	0,17095
beach	0,41997
happy	0,41997
love	0,41997
rubbish	0,32192
trouble	0,32192
water	0,32192
bad	0,97095

→

Fitur	**MI**
bad	0,97095
beach	0,41997
happy	0,41997
love	0,41997
rubbish	0,32192

D. Klasifikasi Naïve Bayes Classifier (NBC)

Pada klasifikasi NBC akan dilakukan 2 tahap yaitu tahap pelatihan dan klasifikasi. Proses pelatihan dan klasifikasi NBC ditunjukkan pada Gambar 3.



Gambar 3. Tahap Klasifikasi NBC

1. Probabilitas Kelas

Perhitungan probabilitas dari setiap kelas dilakukan dengan menggunakan persamaan 2.

$$p(C_i) = \frac{fd(C_i)}{|D|} \quad (2)$$

di mana :
$fd(C_i)$ = Jumlah *tweet* yang termasuk kelas C_i
$|D|$ = Jumlah data latih

Contoh perhitungan probabilitas kelas ditunjukkan pada Tabel 5.

TABEL V. Tabel Probabilitas Kelas

Kelas	***Tweet***					$fd(C_j)$	$p(C_j)$
	1	**2**	**3**	**4**	**5**		
Positif	1	1	1	0	0	3	$\frac{3}{5}$
Negatif	0	0	0	1	1	2	$\frac{2}{5}$

2. Probabilitas Fitur

Perhitungan probabilitas dari setiap fitur dilakukan dengan menggunakan persamaan 3.

$$p(W_k|C_i) = \frac{f(W_k|C_i) + 1}{f(C_i) + |W|} \quad (3)$$

di mana :
$f(W_k|C_i)$ = Nilai kemunculan fitur pada kelas C_i
$f(C_i)$ = Jumlah fitur pada kelas C_i
$|W|$= Jumlah keseluruhan dari fitur (tanpa *duplicate* fitur)

Contoh perhitungan probabilitas fitur ditunjukkan pada Tabel 6.

TABEL VI. Tabel Probabilitas Fitur

Data $f(W_k\|C_i)$	**Kelas**	
	Positif	**Negatif**
bad	$\frac{0+1}{10+10} = \frac{1}{20}$	$\frac{2+1}{5+10} = \frac{3}{15}$
beach	$\frac{2+1}{10+10} = \frac{3}{20}$	$\frac{0+1}{5+10} = \frac{1}{15}$
happy	$\frac{3+1}{10+10} = \frac{4}{20}$	$\frac{0+1}{5+10} = \frac{1}{15}$
love	$\frac{2+1}{10+10} = \frac{3}{20}$	$\frac{0+1}{5+10} = \frac{1}{15}$
rubbish	$\frac{0+1}{10+10} = \frac{1}{20}$	$\frac{1+1}{5+10} = \frac{2}{15}$

3. Probabilitas V_{MAP}

Perhitungan untuk mencari V_{MAP} dilakukan dengan menggunakan persamaan 4.

$$V_{MAP} = \underset{Vj\ e\ V}{\arg max} \prod_{i=1}^{n} P(X_i|V_j)P(V_j) \quad (4)$$

di mana :
V_j = Kategori *tweet* yaitu sentimen positif dan sentimen negatif.
$P(X_i|V_j)$ = Probabilitas X_i pada kategori V_j
$P(V_j)$ = Probabilitas dari V_j

Contoh *tweet* data uji :

Rubbish everywhere!! Mataram city – Lombok https://www.instagram.com/p/BChXtu

Setelah tahap *pre-processing tweet* menjadi :

rubbish	*mataram*	*city*	*lombok*

Penerapan perhitungan V_{MAP} untuk data uji :
V_{MAP} ("positif") = P("positif") . P("*rubbish*"|" positif") = $\frac{3}{5} x \frac{1}{20} = 0{,}03$
V_{MAP} ("negatif") = P("negatif") . P("*rubbish*"|" negatif") = $\frac{2}{5} x \frac{2}{15} = 0{,}05$

Dari hasil perhitungan V_{MAP} di atas didapatkan bahwa nilai V_{MAP} negatif > V_{MAP} positif sehingga dapat disimpulkan bahwa *tweet* tersebut diklasifikasikan ke dalam sentimen negatif.

E. Validasi dan Evaluasi

Validasi akan dilakukan dengan menggunakan *k-fold cross validation* dan pengukuran akurasi dilakukan dengan *confusion matrix*. Tujuan dari *confusion matrix* yaitu melihat kinerja dari model klasifikasi [11]. Perhitungan *confusion matrix* dapat dilihat pada Tabel 7.

TABEL VII Tabel *Confusion Matrix*

Data Class	**Prediksi P**	**Prediksi N**
Positive	TP	FN
Negative	FP	TN

di mana :
TP (*True Positive*) = Kelas yang diprediksi positif dan benar.
TN (*True Negative*) = Kelas yang diprediksi negatif dan benar.
FP (*False Positive*) = Kelas yang diprediksi positif dan salah.
FN (*False Negative*) = Kelas yang diprediksi negatif dan salah.

Rumus untuk mendapatkan nilai akurasi, *precision* dan *recall* dinotasikan pada persamaan 5 hingga 9.

$$\text{Akurasi} = \frac{tp + tn}{tp + fp + fn + tn} \quad (5)$$

$$Precision\ Positive = \frac{tp}{tp + fp} \quad (6)$$

$$Precision\ Negative = \frac{tn}{tn + fn} \quad (7)$$

$$Recall\ Positive = \frac{tp}{tp + fn} \quad (8)$$

$$Recall\ Negative = \frac{tn}{tn + fp} \quad (9)$$

IV. IMPLEMENTASI DAN PENGUJIAN

A. Implementasi *Pre-processing*

Hasil implementasi *pre-processing* dapat dilihat pada Gambar 4.

```
Make your swimming more enjoyable, floating on
the water and relax! ?? ily :* #lombok
https://t.co/ukM23ePUE4

                    ↓

make swimm enjoyable float water relax love
happy
```

Gambar 4. Implementasi *Pre-Processing*

B. Implementasi Seleksi Fitur MI

Pada implementasi MI kali ini, ada beberapa variasi nilai fitur yang digunakan yaitu 10 fitur, 20 fitur, 30 fitur, 40 fitur dan 50 fitur di mana didapatkan hasil bahwa fitur dengan nilai 10 merupakan nilai yang paling optimal dengan akuasi tertinggi. Walaupun nilai fitur cukup kecil namun informasi dari 10 fitur ini sudah cukup besar untuk menentukan suatu kategori. Hasil dari implementasi setelah dilakukan seleksi fitur MI dapat dilihat pada Gambar 5.

```
paradise = [4.0, 223.0, 2.0, 19.0] =
0.008919897158994475
sunset = [12.0, 215.0, 0.0, 21.0] =
0.0063382313437369405
happy = [12.0, 215.0, 0.0, 21.0] =
0.0063382313437369405
beauty = [11.0, 216.0, 0.0, 21.0] =
0.005797293706972409
enjoy = [11.0, 216.0, 0.0, 21.0] =
0.005797293706972409
dirty = [15.0, 6.0, 0.0, 227.0] =
0.25627589676399487
rubbish = [4.0, 17.0, 0.0, 227.0] =
0.05963297190928474
bad = [3.0, 18.0, 1.0, 226.0] =
0.03165155302889778
garbage = [2.0, 19.0, 0.0, 227.0] =
0.029250296530508785
sad = [2.0, 19.0, 0.0, 227.0] =
0.029250296530508785
```

Gambar 5. Implementasi Setelah Seleksi Fitur MI

C. Implementasi NBC

Hasil probabilitas kelas NBC ditunjukkan pada Gambar 6 sedangkan hasil dari probabilitas fitur dapat dilihat pada Gambar 7.

```
Jumlah data kelas positif = 227.0
Jumlah semua data = 248.0
Probabilitas kelas positif :
227.0/248.0=0.9153225806451613
====================================
Jumlah data kelas negatif = 21.0
Jumlah semua data = 248.0
Probabilitas kelas negatif :
21.0/248.0=0.0846774193548387
```

Gambar 6. Implementasi NBC Untuk Probabilitas Kelas

```
Probabilitas Fitur Positif :
============================
happy
(14.0+1)/(972.0+610.0)=0.009481668773704172
beauty
(11.0+1)/(972.0+610.0)=0.007585335018963337
sunset
(12.0+1)/(972.0+610.0)=0.008217446270543615
dirty
(0.0+1)/(972.0+610.0)=6.321112515802782E-4
rubbish
(0.0+1)/(972.0+610.0)=6.321112515802782E-4
.
.
etc
```

```
Probabilitas Fitur Negatif :
============================
happy
(0.0+1)/(108.0+610.0)=0.001392757660167131
beauty
(0.0+1)/(108.0+610.0)=0.001392757660167131
sunset
(0.0+1)/(108.0+610.0)=0.001392757660167131
dirty
(14.0+1)/(108.0+610.0)=0.020891364902506964
rubbish
(5.0+1)/(108.0+610.0)=0.008356545961002786
.
.
etc
```

Gambar 7. Implementasi NBC untuk probabilitas fitur

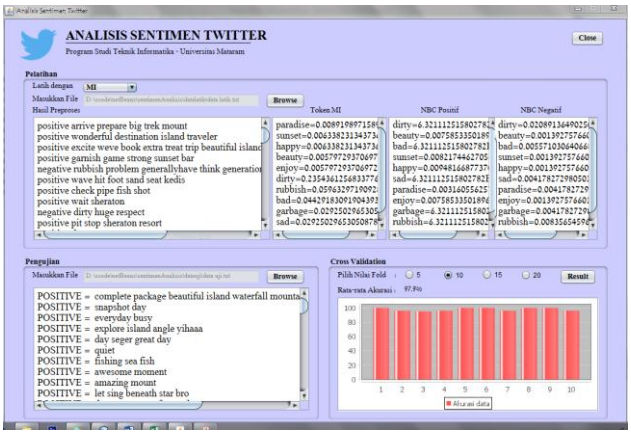
Hasil klasifikasi dengan NBC dapat dilihat pada Gambar 8.

```
monday happy bless god amazing thing week
Kata yang ada di vocab =
happy
Positive = 0.008575789325220124
Negative = 1.1728174425877937E-4
Result = POSITIVE
====================================
loang baloq bad kinda dirty
Kata yang ada di vocab =
bad
dirty
Positive = 7.142027337264312E-7
Negative = 1.0396165696069085E-5
Result = NEGATIVE
====================================
.
.
etc
```

Gambar 8. Implementasi NBC Untuk Hasil Klasifikasi

D. Implementasi Antarmuka

Tampilan antarmuka pada saat dijalankan dapat dilihat pada Gambar 9.



Gambar 9. Implementasi Antarmuka

E. Pengujian Akurasi Metode NBC

Variasi nilai akurasi dengan menggunakan metode MI dapat dilihat pada Tabel 8 dan variasi nilai akurasi tanpa menggunakan metode MI dapat dilihat pada Tabel 9.

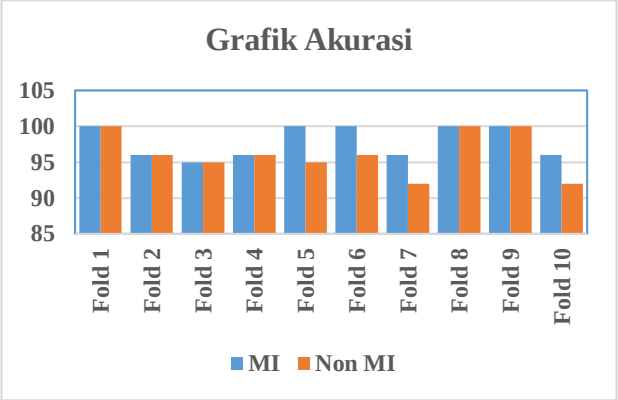
TABEL VIII. Tabel Pengujan Akurasi, *Precision* dan *Recall* MI

Fold	Data Latih	Data Uji	Akurasi (%)	*Precision* (%)	*Recall* (%)
1	223	25	100	100	100
2	223	25	96	97	75
3	224	24	95	97	75
4	223	25	96	97	75
5	224	24	100	100	100
6	223	25	100	100	100
7	223	25	96	97	75
8	223	25	100	100	100
9	223	25	100	100	100
10	223	25	96	97	83

TABEL IX. Tabel Pengujan Akurasi, *Precision* dan *Recall Non* MI

Fold	Data Latih	Data Uji	Akurasi (%)	*Precision* (%)	*Recall* (%)
1	223	25	100	100	100
2	223	25	96	97	75
3	224	24	95	97	75
4	223	25	96	97	75
5	224	24	95	97	75
6	223	25	96	97	75
7	223	25	92	92	50
8	223	25	100	100	100
9	223	25	100	100	100
10	223	25	92	95	66

Pada Tabel 8 dan Tabel 9 dapat dilihat bahwa nilai rata-rata akurasi metode NBC menggunakan seleksi fitur MI yaitu sebesar 97,9% dan rata-rata akurasi tanpa menggunakan seleksi fitur MI yaitu sebesar 96,2%. Oleh karena itu dengan adanya penggunaan seleksi fitur MI dapat meningkatkan akurasi sekitar 1,7%. Berdasarkan pada Tabel 4.2 dan Tabel 4.3 dapat dilihat grafik pengaruh penggunaan seleksi fitur MI dan tanpa seleksi fitur MI pada Gambar 10.



Gambar 10. Grafik Perbedaan Akurasi

F. Waktu Klasifikasi

Variasi waktu klasifikasi pada mesin pengklasifikasi dapat dilihat pada Tabel 10.

TABEL X. Perbedaan Waktu Klasifikasi

Fold	MI (ms)	Non MI (ms)
1	130	332
2	132	222
3	130	202
4	97	203
5	96	451
6	102	157
7	91	179
8	141	192
9	100	199
10	80	130

Pada Tabel 10 didapatkan rincian waktu klasifikasi mesin pengklasifikasi sebagai berikut :

Rata-rata waktu klasifikasi untuk setiap *fold* menggunakan metode MI = 109,9 ms

Rata-rata waktu klasifikasi untuk setiap *fold* tanpa menggunakan metode MI = 226,7 ms

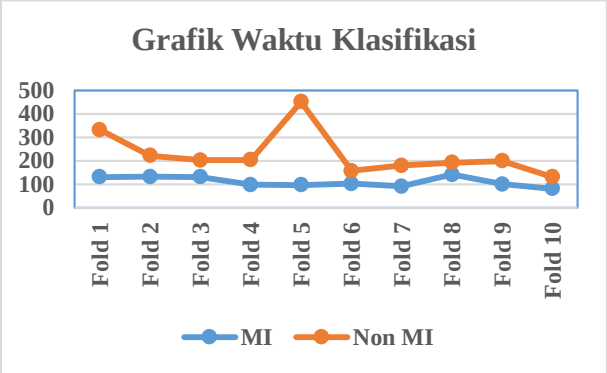
Waktu klasifikasi untuk 1 *tweet* menggunakan metode MI = 4,39 ms

Waktu klasifikasi untuk 1 *tweet* tanpa menggunakan metode MI = 90,68 ms

Selisih waktu klasifikasi untuk 1 *tweet* antara metode MI dan Non MI = 86,28 ms

Data waktu klasifikasi pada Tabel 10 menunjukkan bahwa waktu klasifikasi dengan menggunakan metode seleksi fitur MI lebih cepat dibandingkan dengan klasifikasi tanpa menggunakan metode seleksi fitur MI. Berdasarkan pada Tabel 10 dapat dilihat grafik pengaruh penggunaan seleksi fitur MI dan tanpa seleksi fitur MI

dalam waktu klasifikasi pada mesin pengklasifikasi pada Gambar 11.



Gambar 11. Grafik Perbedaan Waktu Klasifikasi

G. Pengujian Manual Data Uji

Hasil pengujian manual pada 250 data uji dapat dilihat pada Tabel 11:

TABEL XII. Tabel Pengujan Manual

Data	Jumlah Data	Hasil Klasifikasi		
		Mesin	Manual	
			Benar	Salah
Positif	237	228	225	3
Negatif	13	10	10	0
Total	250	238	235	3

Dari 235 data yang diklasifikasikan dengan benar maka dapat dihitung nilai akurasi dari klasifikasi yaitu sebesar 98,7%.

V. KESIMPULAN

A. Kesimpulan

Berdasarkan hasil pembahasan dan pengujian mengenai analisis sentimen twitter maka dapat ditarik kesimpulan sebagai berikut :

1. Akurasi rata-rata klasifikasi dengan menggunakan metode NBC dalam mengklasifikasikan *tweet* ke dalam kelas positif dan kelas negatif dengan menggunakan seleksi fitur MI yaitu sebesar 97,9% dan tanpa menggunakan seleksi fitur MI sebesar 96,2%. Oleh karena itu dengan adanya penggunaan seleksi fitur MI dapat menaikkan akurasi sekitar 1,7%.
2. Penggunaan metode seleksi fitur MI terbukti dapat membuat waktu mesin pengklasifikasi dalam mengklasifikasikan *tweet* menjadi lebih cepat dibandingkan tanpa menggunakan seleksi fitur MI karena jumlah fitur yang berkurang mempengaruhi proses pencarian fitur saat klasifikasi.

B. Saran

Adapun beberapa saran yang dapat digunakan untuk pengembangan kedepannya yaitu sebagai berikut :

1. Pada saat melakukan penarikan data twitter disarankan mencoba menggunakan data twitter yang seimbang dengan topik data yang lebih universal sehingga akan banyak data yang dapat diambil.
2. Dalam penelitian selajutnya disarankan penggunaan bahasa tidak hanya dalam bahasa Inggris tetapi dapat menggunakan bahasa Indonesia mapun bahasa daerah.

DAFTAR PUSTAKA

[1] Liu, B., 2012, *Sentiment Analysis and Opinion Mining*, San Rafael, Morgan and Claypool.

[2] Narayanan, V., Arora, I. & Bhatia, A., 2013, *Fast and Accurate Sentiment Clasification Using Anhanced Naïve Bayes Model*, Indian Institut of Technology, India.

[3] Utami, L.A., 2017, *Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine dan K-Narest Neighbor Berbasis Particle Swarm Optimization*, Jurnal Evolusi ISSN, Vol. 13, No. 1, p. 2527-6514.

[4] Ernawati, S., 2016, *Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen*, Jurnal Evolusi ISSN, Vol. 4, No. 1.

[5] Darma, I.M.B.S., Perdana, R.S., & Indriati 2017, *Analisis Sentimen Televisi Pada Twiter Menggunakan Support Vector Machine dan Algoritma Genetika Sebagai Metode Seleksi Fitur*, Jurnal Evolusi ISSN, Vol. 2, No. 3, p. 998-1007.

[6] Gupta, N. & Parveen, S., 2016, *Efficient Sentiment Analysis Using Optimal Feature and Bayesian Classifier*, International Journal of Computer Applications, Vol. 145, No. 8.

[7] Garcia, V., Sanchez, J. & Mollineda, R., 2007, *An Empirical Study of the Behavior of Classifiers on Imbalanced and Overlapped Data Sets*, Lecture Notes in Computer Science, Vol 4756, p. 397-406.

[8] Akbani, R., Kwek, S. & Japkowicz, N., 2004, *Applying Support Machines to Imbalanced Dataset*, Proceedings of the 15th European Conference on Machine Learning, pp. 39-50.

[9] Sobran, N.M.M., Ahmad, A. & Ibrahim Z., 2013, *Classification of Imbalanced Dataset Using Conventional Naïve Bayes Classifier*, Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT, p. 978-967.

[10] Manning, C. D., Raghavan, P. & Schutze, H., 2009, *An Introduction to Information Retrieval*, Cambridge University Press, England.

[11] Sokolova, M. & Lapalme, G., 2009, *A Systematic Analysis of Performance Measures for Classification Tasks*, International Journal Information Processing and Management, No. 45, p. 427-437.