Perbandingan Seleksi Fitur Untuk Klasifikasi Sentimen SVM Pada Twitter

Dede Juniawan Suri[1], Ken Kinanti Purnamasari[2]

[1] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
[2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail : dede.juniawan.suri@gmail.com[1], guinev_g@yahoo.co.nz [2]

ABSTRAK

Analisis sentimen adalah bidang studi untuk menganalisis pendapat, sentimen, evaluasi, penilaian sikap dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, peristiwa, dan atribut lainnya. Seleksi fitur memiliki peran penting dalam klasifikasi teks *term*asuk analisis sentimen, penggunaan metode seleksi fitur membantu memahami atribut yang relevan untuk kelas tertentu serta meningkatkan akurasi klasifikasi. Metode seleksi fitur TF-IDF, *Ratio*, dan N-Gram telah diterapkan dalam klasifikasi pesan singkat twitter, metode-metode tersebut menghasilkan akurasi yang baik, akan tetapi belum diketahui metode yang lebih baik karena dilakukan dalam penelitian yang berbeda. Berdasarkan hasil penelitian dengan dataset 1100 *tweet* yang dibagi menjadi 500 untuk *training* filter *tweet*, 500 untuk *training* klasifikasi sentimen, 50 untuk *testing* filter *tweet*, dan 50 untuk *testing* klasifikasi sentimen. Metode TF-IDF menghasilkan akurasi tertinggi yaitu 74% pada klasifikasi filter *tweet* dan 80% pada klasfikasi sentimen, *Ratio* menghasilkan akurasi 72% pada klasifikasi filter *tweet* dan 74% pada klasfikasi sentimen, dan N-Gram menghasilkan akurasi 64% pada klasifikasi filter *tweet* dan 78% pada klasfikasi sentimen.

Kata kunci : Analisis Sentimen, Seleksi Fitur, TF-IDF, *Ratio*, N-Gram, *Support vector machine*, *Twitter*.

1. PENDAHULUAN

Analisis sentimen adalah bidang studi untuk menganalisis pendapat, sentimen, evaluasi, penilaian sikap dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, peristiwa dan atribut lainnya. Analisis sentimen berfokus kepada opini yang mengekspresikan sentimen positif atau negatif. Seleksi fitur memiliki peran penting dalam klasifikasi teks *term*asuk analisis sentimen, penggunaan metode seleksi fitur membantu memahami atribut yang relevan untuk kelas tertentu serta meningkatkan akurasi klasifikasi.

Metode seleksi fitur TF-IDF, *Ratio*, dan N-Gram telah diterapkan dalam klasifikasi pesan singkat twitter dan menghasilkan akurasi yang baik. Namun terdapat permasalahan untuk menentukan metode seleksi fitur yang lebih baik karena dilakukan dalam penelitian yang berbeda dengan penggunaan jumlah dan sumber data yang berbeda, baik pada proses *training* maupun *testing*, maka dari itu untuk mengetahui metode seleksi fitur yang lebih baik untuk analisis sentimen maka perlu dilakukan perbandingan setiap metode seleksi fitur yang pernah diterapkan dengan sumber maupun jumlah data sama.

1.1 Analisis Sentimen

Analisis sentimen atau disebut juga opinion mining adalah bidang studi untuk menganalisis pendapat, sentimen, evaluasi, penilaian sikap dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, peristiwa dan atribut lainnya. Analisis sentimen berfokus kepada opini yang mengekspresikan sentimen positif atau negatif.

Terdapat tiga tahapan umum dalam proses analisis sentimen, yaitu pengambilan data dari sumber data atau yang biasa disebut *crawling*, selanjutnya tahap *training* dan *testing* dengan menggunakan algoritma klasifikasi tertentu. Salah satu algoritma untuk melakukan klasifikasi adalah *Support vector machine*. Sumber data yang dapat digunakan untuk proses analisis sentimen salah satunya adalah media sosial Twitter. Twitter sebagai salah satu jejaring sosial yang interaktif memungkinkan penggunanya untuk memberikan suatu pendapat terhadap suatu hal secara real time. Data yang bisa dianalisis di Twitter yaitu *tweet* yang dibuat pengguna, dimana *tweet-tweet* tersebut nantinya akan diklasifikasikan ke dalam dua kelas sentimen yaitu sentimen positif atau sentimen negatif.

1.2 Seleksi Fitur

Seleksi fitur sebagai tahapan dalam klasifikasi terbagi menjadi 2 pendekatan, yaitu pendekatan *wrapper* dan pendekatan filter. Seleksi fitur dengan pendekatan *filter* dilakukan secara terpisah dengan mesin klasifikasi, atau dengan kata lain seleksi fitur dijadikan tahapan pra-proses sebelum data dimasukkan ke dalam mesin klasifikasi. Sebuah data yang telah dilakukan seleksi fitur dengan pendekatan ini, dapat digunakan atau dikombinasikan dengan beberapa mesin klasifikasi. Berbeda dengan pendekatan *filter*, pendekatan *wrapper* menjadikan mesin klasifikasi sebagai bagian

proses dalam menentukan penting atau tidaknya suatu fitur. Mesin klasifikasi akan dilatih dan diuji dengan kombinasi fitur yang berubah, jika penambahan suatu fitur menurunkan akurasi mesin klasifikasi maka fitur tersebut tidak digunakan. Mesin akan memproses kombinasi fitur yang lain dan tidak menggunakan fitur yang akan memperlemah akurasi sistem.
Pada penelitian ini akan menerapkan seleksi fitur dengan pendekatan *filter* karena lebih sederhana dibandingkan pendekatan wrapper. Selain itu, waktu yang dibutuhkan untuk pendekatan *filter* lebih cepat dibandingkan dengan pendekatan *wrapper*. Adapun metode seleksi fitur yang dipakai yaitu TF-IDF, *Ratio*, dan N-Gram.

1.1.1 TF-IDF

TF-IDF (*Term Frequency–Invers Document Frequency*) merupakan salah satu metode *term weighting. Term weighting* ialah proses memberikan bobot terhadap semua kata pada dokumen.
Term Frequency adalah salah satu metode pembobotan yang paling sederhana. Pada metode ini, setiap *term* diasumsikan memiliki proporsi kepentingan sesuai dengan jumlah terjadinya (munculnya) *term* tersebut dalam dokumen. Persamaan TF adalah sebagai berikut.

$$W(d,t) = TF(d,t) \quad (1)$$

Dimana TF(d,t) adalah frekuensi kemunculan *term* t pada dokumen d.
Inverse Document Frequency memperhatikan kemunculan *term* pada kumpulan dokumen. Pada metode ini, *term* yang dianggap bernilai/berharga adalah *term* yang jarang muncul pada koleksi/ kumpulan dokumen. Persamaan IDF adalah sebagai berikut.

$$idf(t) = log\frac{N}{df(t)} \quad (2)$$

Dimana df(t) adalah banyak dokumen yang mengandung *term* t.
TF-IDF merupakan kombinasi metode TF dengan metode IDF. Sehingga persamaan TF-IDF adalah sebagai berikut.

$$TF - IDF(d,t) = TF(d,t).IDF(t) \quad (3)$$

1.1.2 *Ratio*

Metode *Ratio* didasarkan pada Teori Informasi, pada metode ini terjadi peningkatan nilai atau bobot fitur ketika kata sering terjadi pada suatu kelas dan akan menurun ketika kata tersebut terjadi pada semua kelas. Adapun langkah-langkah dalam metode ini adalah sebagai berikut.

1. Hitung nilai tf_i

$$tf_i = \sum_{j=1}^{n} F_{(i,j)} \quad (4)$$

2. Hitung nilai FR (*Frequency Ratio),* ini akan memberikan gambaran tentang betapa pentingnya kata i untuk kelompok j terpilih

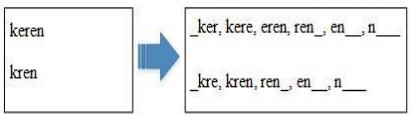
$$FR_{(i,j)} = \frac{F_{(i,j)}}{tf_i} \quad (5)$$

3. Hitung MRF (*Maximum Frequency Ratio*)

$$MRF_i = max\{FR_{(i,j)}\} \quad (6)$$

1.1.3 N-Gram

Salah satu metode yang dapat dimanfaatkan untuk mengambil fitur pada *tweet* yaitu N-Gram. N-Gram adalah potongan n-karakter yang diambil dari suatu string. Metode N-Gram digunakan untuk mengambil potongan-potongan karakter huruf sejumlah n dalam sebuah kata yang secara kontinu, dibaca dari kata sumber hingga akhir dari dokumen. Jika n=1 maka disebut bigram, n=2 disebut bigram dan n=3 disebut trigram.
Keuntungan N-Gram adalah berdasarkan karakteristik N-Gram sebagai bagian dari suatu string, sehingga kesalahan pada sebagian string hanya akan berakibat perbedaan pada sebagian N-Gram. Sebagai contoh, jika N-Gram dari dua string dibandingkan, kemudian dihitung N-Gram yang sama dari dua string tersebut maka akan didapatkan nilai similaritas atau kemiripan dua string tersebut yang bersifat resistan terhadap kesalahan tekstual.
Sebagai contoh kemiripan antara kata baku “keren” dan kata tidak bakunya “kren” dapat diukur kesamaannya dengan cara menghitung berapa buah n-gram yang diambil dari dua kata tersebut yang memiliki kesamaan, yaitu:



Gambar 1 Contoh N-Gram

Dari n-gram yang dibangkitkan pada kedua kata tersebut, terdapat 3 kesamaan n-gram dari kedua kata tersebut. Terlihat hanya terdapat 2 bagian n-gram pada kata “kren” yang tidak terdapat pada kata “keren”, sehingga kata “kren” bisa diasumsikan sebagai “keren”.
Pada penelitian Johannes penggunaan fitur 2-gram dan 3-gram bermanfaat untuk peningkatan performa kategorisasi. Di lain pihak, penggunaan n-gram degan n>3 justru mengurangi performa kategorisasi, maka pada penelitian ini akan menerapkan N-Gram dengan n=3 atau trigram.

1.3 *Support vector machine*

Support vector machine (SVM) merupakan salah satu metode machine learning yang mengubah teks menjadi data vector. Vector dalam penelitian ini

memiliki dua komponen yaitu dimensi (word id) dan bobot. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari Hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space.

Dalam linear *Support vector machine* pemisah merupakan fungsi linear. Data latih dinyatakan oleh (x_i, y_i) dan $x_i = \{ x_1, x_2, \ldots, x_{iq}\}$ merupakan atribut (fitur) set untuk data latih kelas ke-i. Untuk y_i) € {-1,1} menyatakan label kelas. pendefinisian persamaan suatu *hyperplane* pemisah yang dituliskan dengan:

$$wx_i + \mathrm{b} = 0 \tag{7}$$

Data x_i yang terbagi ke dalam dua kelas, yang *term*asuk kelas -1 (sampel negatif) didefinisikan sebagai vektor yang memenuhi pertidaksamaan 8 berikut ini.

$$wx_i + \mathrm{b} < 0 \; untuk \; y_i = 1 \tag{8}$$

Sedangkan yang *term*asuk kelas +1 (sampel positif) memenuhi pertidaksamaan 9 berikut.

$$wx_i + \mathrm{b} > 0 \; untuk \; y_i = +1 \tag{9}$$

Dimana:

x_i = data input

y_i = label yang diberikan

w = nilai dari bidang normal

b = posisi bidang relatif terhadap pusat koordinat

Parameter w dan b adalah parameter yang akan dicari nilainya. Bila label data y_i = -1, maka pembatas menjadi persamaan 10 berikut ini.

$$wx_i + \mathrm{b} \leq -1 \tag{10}$$

Bila label data y_i = +1, maka pembatas menjadi persamaan 11 berikut.

$$wx_i + \mathrm{b} \geq +1 \tag{11}$$

Margin terbesar dapat dicari dengan cara memaksimalkan jarak antar bidang pembatas kedua kelas dan titik terdekatnya, yaitu $\frac{2}{|w|}$. Hal ini dirumuskan sebagai permasalahan *quadratic programming* (QP) problem yaitu mencari titik minimal persamaan (12) dengan memperhatikan persamaan (13) berikut.

$$\min \tau(w) = \frac{1}{2}||w||^2 \tag{12}$$

$$y_i(wx_i + b) - 1 \geq 0, (i = 1, \ldots, n) \tag{13}$$

Permasalahan ini ini dapat dipecahkan dengan berbagi teknik komputasi. Lebih mudah diselesaikan dengan mengubah persamaan 12 ke dalam fungsi *Lagrangian* pada persamaan 14, dan menyederhanakannya menjadi persamaan 15 berikut.

$$L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i(y_i((w^T x_i + b) - 1)) \tag{14}$$

$$L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) + \sum_{i=1}^{n} a_i \tag{15}$$

Dimana a_i adalah *lagrange multiplier* yang bernilai nol atau positif ($a_i \geq 0$). Nilai optimal dari persamaan 15 dapat dihitung dengan meminimalkan L terhadap w, b, dan a. Dapat dilihat pada persamaan 16 sampai 18 berikut ini.

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} a_i \, y_i x_i = 0 \tag{16}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} a_i \, y_i = 0 \tag{17}$$

$$\frac{\partial L}{\partial a} = w - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) - \sum_{i=1}^{n} a_i = 0 \tag{18}$$

Maka masalah *Lagrange* untuk klasifikasi dapat dinyatakan pada persamaan 19 berikut.

$$Min \; L(w, a, b) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} a_i \, y_i(w^T x_i + b) + \sum_{i=1}^{n} a_i \tag{19}$$

Dengan memperhatikan persamaan 20 dan 21 berikut ini.

$$w - \sum_{i=1}^{n} a_i \, y_i x_i = 0 \tag{20}$$

$$\sum_{i=1}^{n} a_i \, y_i = 0 \tag{21}$$

Model persamaan 19 diatas merupakan model primal *Lagrange*. Sedangkan dengan memaksimalkan L terhadap a_i, persamannya menjadi persamaan 22 berikut ini.

$$Max \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1,j=1}^{n} a_i y_i {y_j}^T \, x_i {x_j}^T \tag{22}$$

Dengan memperhatikan persamaan 23 berikut ini.

$$\sum_{i=1}^{n} a_i y_i = 0, a_i > 0 \; (i, j = 1, \ldots, n) \tag{23}$$

Untuk mencari nilai x_i dan y_i dapat dilakukan ketika sudah didapatkan nilai tiap kata (*term*) dari pembobotan dari masing-masing metode seleksi fitur dan inisialisasi kelas . Hasil dari pembobotan diubah

ke dalam bentuk format data svm (x), sedangkan data kelas menjadi label data svm (y).

Untuk mendapatkan nilai a_i, langkah pertama adalah mengubah setiap *tweet* menjadi nilai vektor (*support vector*) = $\binom{x}{y}$. Kemudian nilai vektor dari setiap *tweet* dimasukkan ke persamaan 24 kernel *trick phi* (φ) berikut ini.

$$\varphi\begin{bmatrix}x\\y\end{bmatrix} = \begin{cases}\sqrt{x_n^2+y_n^2} > 2, maka \begin{bmatrix}\sqrt{x_n^2+y_n^2}-x+|x-y|\\ \sqrt{x_n^2+y_n^2}-x+|x-y|\end{bmatrix} \\ \sqrt{x_n^2+y_n^2} < 2, maka \begin{bmatrix}x\\y\end{bmatrix}\end{cases} \quad (24)$$

Nilai x didapatkan dari persamaan 25 kernel linear untuk x berikut.

$$\sum_{i=1,j=1}^{n} x_i^T x_j \ (i,j=1,\dots,n) \quad (25)$$

Nilai y didapatkan dari persamaan 26 kernel linear untuk y berikut.

$$\sum_{i=1,j=1}^{n} y_i^T y_j \ (i,j=1,\dots,n) \quad (26)$$

Untuk mendapatkan jarak tegak lurus yang optimal dengan mempertimbangkan vektor positif, maka hasil perhitungan dari substitusi nilai x dan nilai y ke persamaan 2.24 diberi nilai bias 1. Kemudian cari parameter a_i, dengan terlebih dahulu mencari nilai fungsi setiap *tweet* menggunakan persamaan 26, lalu mencari nilai a_i pada persamaan linear menggunakan persamaan 27 dengan memperhatikan $i,j=1,\dots,n$ berikut.

$$\sum_{i=1,j=1}^{n} a_i {T_i}^T T_j \quad (27)$$

$$\sum_{i=1,j=1}^{n} a_i {T_i}^T T_j = y_i \quad (28)$$

Setelah parameter a_i didapatkan, kemudian masukkan ke persamaan 29 berikut.

$$W = \sum_{i=1}^{n} a_i T_i \quad (29)$$

Hasil yang didapatkan menggunakan persamaan 29, selanjutnya digunakan persamaan 30 untuk mendapatkan nilai w dan b yaitu sebagai berikut.

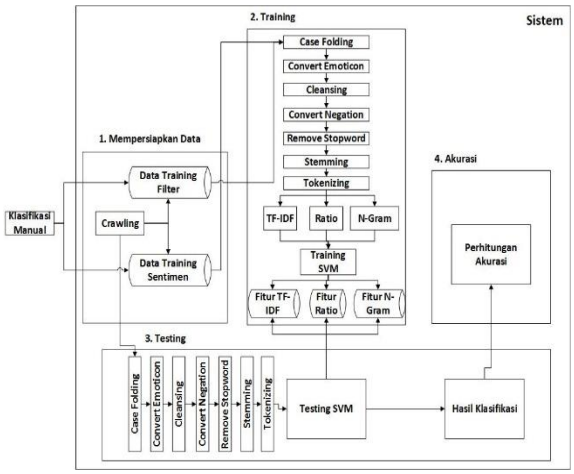
$$y = wx + b \quad (30)$$

Sedemikian sehingga didapatkanlah nilai w dan nilai b atau nilai *hyperplane* untuk mengklasifikasikan kedua kelas.

2. ISI PENELITIAN

2.1 Analisis Sistem

Sistem yang dibangun memiliki beberapa tahapan. Tahapan-tahapan tersebut yaitu proses pengambilan *tweet*, proses *training*, proses *testing* dan proses penghitungan akurasi. Gambaran umum dari aplikasi yang akan dibangun ditampilkan pada gambar berikut.



Gambar 2 Gambaran Sistem

Berikut ini penjelasan dari gambar 2.

1. Bagian pertama adalah proses pengambilan *tweet* (*crawling*) kemudian disimpan di tabel data *training* filter dan data *training* sentimen, kemudian dilakukan proses klasifikasi secara manual.
2. Bagian kedua adalah proses *training* SVM, pada penelitian ini *training* dibagi menjadi dua bagian yaitu *training* klasifikasi *tweet* yang mengandung sentimen dan tidak mengandung sentimen (sentimen/bukan sentimen) kemudian *training* klasifikasi sentimen. Namun sebelumnya data *training* melalui tahap *preprocessing* terlebih dahulu, tahap tersebut meliputi case folding, convert emoticon, cleansing, convert negation, remove stopword, stemming, tokenizing, dan selanjutnya seleksi fitur dengan metode TF-IDF, *Ratio*, dan N-Gram.
3. Bagian ketiga adalah proses *testing* SVM. *Tweet* atau data *testing* terlebih dahulu melewati tahap *preprocessing* (seperti tahap *preprocessing* pada bagian kedua), setelah itu dilakukan proses pencocokan token hasil tahap *preprocessing* dengan fitur yang ada pada database dari hasil *training* (bagian 2), hasilnya adalah *tweet* sentimen atau bukan sentimen. *Tweet* yang terklasifikasi sebagai sentimen akan menjadi input *testing* klasifikasi sentimen.
4. Bagian keempat adalah proses perhitungan akurasi, tahap ini dilakukan untuk melihat pengaruh metode seleksi fitur terhadap akurasi klasifikasi.

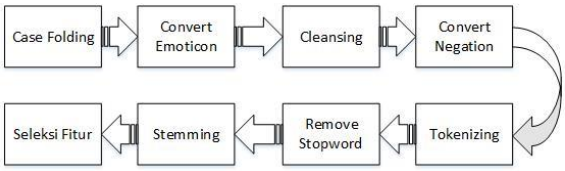
2.2 Analisis Data Masukan

Data masukan yang digunakan adalah data *tweet* dari akun twitter Net TV. Data *tweet* tersebut didapat dengan memanfaatkan fitur API (Application Programming Interface) yang telah disediakan oleh Twitter. Data yang diambil adalah data *tweet* yang ada pada akun Net TV dan SCTV atau yang mengandung keyword @netmediatama, #nettv,

#netmediatama, dan @SCTV_. Data yang didapat berupa teks dengan panjang maksimal 140 karakter dan masih berupa data mentah dan terdapat noise yang tidak memiliki pengaruh pada klasifikasi sentimen seperti URL, hashtag (#), username (@), dan imbuhan sehingga diperlukan tahap praproses.

2.3 Analisis Tahap *Preprocessing*

Berdasarkan dari analisis data masukan yang telah dijelaskan, maka perlu untuk dilakukan tahap *preprocessing*. Tahap *preprocessing* yang akan dilakukan yaitu case folding, convert emoticon, cleansing, convert negation, tokenizing, remove stopword, stemming, dan seleksi fitur. Gambaran dari tahap *preprocessing* dapat dilihat pada gambar berikut.



Gambar 3 Tahap *Preprocessing*

1. Case Folding

Case Folding adalah proses menyeragamkan bentuk kata-kata pada *tweet* menjadi huruf kecil (*lowercase*) atau huruf kapital (*uppercase*). Pada penelitian ini diseragamkan menjadi bentuk *lowercase*.

2. Convert Emoticon

Emoticon merupakan salah satu cara pengungkapan ekspresi perasaan secara tekstual. Hal tersebut akan membantu dalam menentukan sentimen dalam suatu kalimat atau *tweet*. Setiap emoticon akan dikonversikan ke dalam string yang bersesuaian, proses ini dinamakan convert emoticon.

3. Cleansing

Cleansing merupakan proses membersihkan kata-kata yang tidak diperlukan untuk mengurangi noise. Kata-kata yang dihilangkan yaitu URL, hashtag (#) dan username (@). Selain itu juga tanda baca seperti titik (.), koma (,) dan tanda baca lainnya akan dihilangkan karena tidak memiliki pengaruh apapun terhadap perhitungan nilai sentimen. Jika terdapat angka, maka akan dihilangkan. Selain itu, jika terdapat spasi lebih dari satu, maka akan diganti menjadi satu spasi.

4. Convert Negation

Pada sebuah *tweet* yang dibuat pengguna, memungkinkan terdapat negasi yang dapat membalikan arti dari suatu kata. Maka perlu dilakukan tahap convert negation yaitu dengan cara menggabungkan kata negasi tersebut dengan kata setelahnya.

5. Tokenizing

Tokenizing adalah memisahkan kata pada *tweet* menjadi kata-kata tunggal. Pemisahan tersebut ditandai dengan tanda pemisah yaitu spasi.

6. Remove Stopword

Remove stopword yaitu proses membuang kata-kata yang terdapat pada daftar stopword. Kata-kata yang berada dalam daftar stopword berupa kata ganti orang, kata ganti penghubung dan penunjuk.

7. Stemming

Stemming merupakan suatu proses yang mentransformasi kata-kata yang berimbuhan ke kata-kata dasarnya (*root word*). Sebagai contoh, kata menyenangkan, menyenangi, disebut dapat dikatakan serupa atau satu kelompok dan dapat diwakili oleh satu kata umum senang. Proses stemming pada teks berbahasa Indonesia berbeda dengan stemming pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia, selain sufiks, prefiks, dan konfiks juga dihilangkan. Pada penelitian ini, algoritma yang digunakan dalam proses stemming yaitu nazief dan andriani

8. Seleksi Fitur

Pada penelitian ini menggunakan tiga metode seleksi fitur yang bertujuan untuk dibandingkan sehingga mengetahui metode seleksi fitur terbaik untuk analisis sentimen pada twitter, adapan ketiga metode seleksi fitur tersebut yaitu TF-IDF, *Ratio*, dan N-Gram dengan jumlah n=3.

2.4 Analisis Tahap Klasifikasi

Tahap klasifikasi *tweet* dibagi menjadi dua tahap, yaitu filter *tweet* (sentimen dan bukan sentimen) dan klasifikasi sentimen (positif dan negatif). *Tweet* yang tidak mengandung unsur sentimen tersebut perlu dilakukan filter terlebih dahulu supaya menghasilkan akurasi yang baik pada tahap klasifikasi sentimen. Adapun metode klasifikasi yang digunakan pada penelitian ini adalah *Support vector machine*.

2.5 Pengujian

Pengujian akurasi dalam penelitian ini adalah mengukur tingkat keakuratan setiap metode seleksi terhadap klasifikasi dengan menggunakan metode *Support Vector Machine*.

Jumlah data yang digunakan sebanyak 1100 *tweet* yang diambil dari akun twitter Net TV dan SCTV. Data tersebut dibagi menjadi 500 data untuk training *filter tweet*, 500 data untuk training klasifikasi sentimen, 50 data untuk testing *filter tweet*, dan 50 data untuk testing klasifikasi sentimen dengan persebaran data sebagai berikut.

Tabel 1 Dataset *Filter Tweet*

	Bukan Sentimen	**Sentimen**	**Jumlah**
Data *Training*	250 *tweet*	250 *tweet*	500 *tweet*
Data *Testing*	22 *tweet*	28 *tweet*	50 *tweet*

Tabel 2 Dataset Klasifikasi Sentimen

	Negatif	**Positif**	**Jumlah**
Data *Training*	232 *tweet*	268 *tweet*	500 *tweet*
Data *Testing*	20 *tweet*	30 *tweet*	50 *tweet*

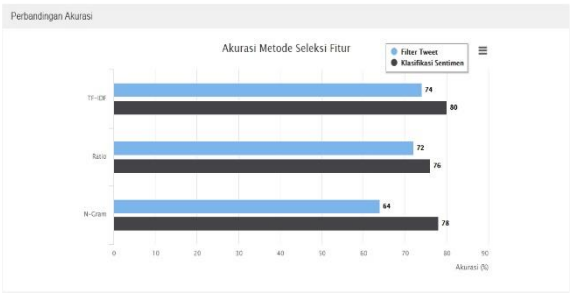
Tabel 3 Hasil Pengujian

	Klasifikasi *Filter Tweet*	**Klasifikasi Sentimen**
TF-IDF	74%	80%
Ratio	72%	76%
N-Gram	64%	78%

3. PENUTUP

3.1 Kesimpulan

Berdasarkan hasil penelitian dengan dataset 1100 *tweet* yang dibagi menjadi 500 untuk *training* filter *tweet*, 500 untuk *training* klasifikasi sentimen, 50 untuk *testing* filter *tweet*, dan 50 untuk *testing* klasifikasi sentimen. Metode TF-IDF menghasilkan akurasi tertinggi yaitu 74% pada klasifikasi filter *tweet* dan 80% pada klasfikasi sentimen, metode *Ratio* menghasilkan akurasi 72% pada klasifikasi filter *tweet* dan 76% pada klasfikasi sentimen, dan N-Gram menghasilkan akurasi 64% pada klasifikasi filter *tweet* dan 78% pada klasfikasi sentimen. Jadi, untuk kasus analisis sentimen pada twitter lebih baik menggunakan metode seleksi fitur TF-IDF yang menghasilkan akurasi yang lebih baik pada tahap klasifikasi.



Gambar 4 Grafik Perbandingan Akurasi

3.2 Saran

Berdasarkan hasil pengujian yang telah dilakukan, maka saran yang diharapkan setelah dilakukannya perbandingan seleksi fitur pada kasus analisis sentimen pada twitter.

1. Penggunaan kernel yang berbeda pada metode SVM sebagai bahan perbandingan seperti kernel polynomial, gaussian, atau RBF.
2. Menambah dan memperbaharui daftar emoticon agar bisa mendeteksi lebih banyak emoticon *term*asuk emoticon terbaru.
3. Penggunaan dataset yang lebih banyak.

DAFTAR PUSTAKA

[1] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publisher, 2012.

[2] P. Koncz and J. Paralic, "An approach to feature selection for sentiment analysis," *IEEE International Conference*, p. 357–362, 2011.

[3] T. H. Susilo and S. Rochimah, "Pengklasifikasian Topik dan Analisis Sentimen Dalam Media Sosial," *SNASTI*, 2013.

[4] A. R. Sentiaji dan A. M. Bachtiar, "Analisis Sentimen Terhadap Acara Televisi Berdasarkan Opini Publik," *Jurnal Ilmiah Komputer dan Informatika*, 2014.

[5] P. Pitria, "Analisis Sentimen Pengguna Twitter pada Akun Resmi Samsung Indonesia dengan menggunakan Naive Bayes," Perpustakaan Unikom, 2014.

[6] F. Z Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Universiteit van Amsterdam, Amsterdam.

[7] Octaviani. P. A., Wilandari. Y., & Ispriyanti. D., "Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang," *Jurnal Gaussian*, vol. 3, pp. 811-820.

[8] I. Dilrukshi and K. d. Zoysa, "A Feature Selection Method for Twitter News Classification," *International Journal of Machine Learning and Computing*, vol. 4, no. 4, 2014.

[9] J. Fürnkranz, "A Study Using n -gram Features for Text Categorization," *Austrian Research Institute for Artificial Intelligence*, p. 8, 1998.

[10] D. S. Pamungkas, N. A. Setiyanto and E. Dolphina, "Analisis Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci "Kurikulum 2013"," *Techno.COM*, vol. 14, no. 4, 2015.

[11] F. T. Anggraeny and M. Widyasri, "Klasifikasi Voting ANN PSO Biclass dengan Seleksi Fitur Gian Ratio," 2012.