Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen

Joang Ipmawati[1)], Kusrini[2)], Emha Taufiq Luthfi[3)]
[1), 2),3)] *Magister Teknik Informatika STMIK Amikom Yogyakarta*
joangipmawati@gmail.com[1)], kusrini@amikom.ac.id[2)], emhataufiqluthfi@amikom.ac.id[3)]

Abstract - Opininion mining also called sentiment analysis is a computational research of opinions, sentiments and emotions that are textually to see opinion on an issue, or to identify the tendency of things in the market. This time public opinion be an important resource in making decisions for a product. Classification algorithm to perform text mining including Support Vector Machine (SVM), Naïve Bayessian classification (NBC) and K-Nearest Neighbor (K-NN). These of algorithms will compired to find out a good performance in terms of accuracy for two different datasets that imdb movie reviews and twitter sentiment. The results of the comparison showed SVM obtain good results in accuracy in the data imdb movie reviews 78.55% and on twitter dataset 72%. Similarly, NBC obtained the data accuracy at 78.55% twitter but different data twitter 67.33%. The results of F-Measure SVM movie review show and NBC showed the same results, namely 0.785 and also for the AUC, the results surpass NBC 0.869, SVM get results 0.786 and while KNN obtain the results 0.572. F-Measure to twitter SVM is superior obtaining results of 0.720 and 0.673 NBC obtained results while K-NN 0.545. and for the results of the AUC, as dataset imdb, on twitter this dataset NBC also outperformed SVM and K-NN. AUC to obtain results NBC 0.735, SVM obtain results K-NN 0.658 and 0.618 get results.
Keywords : Text Mining, Sentiment Analysis, SVM, Naïve Bayessian, K-NN, compare, comparation

Abstrak - Opininion mining juga disebut analisis sentimen adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual dilakukan untuk melihat pendapat terhadap sebuah masalah, atau untuk identifikasi kecenderungan hal di pasar. Saat ini pendapat masyarakat menjadi sumber yang penting dalam pengambilan keputusan akan suatu produk. Algoritma klasifikasi yang dapat melakukan teks mining diantaranya Support Vector Machine (SVM), Naïve Bayessian classification (NBC) dan K-Nearest Neighbor (K-NN). Ketiga algoritma ini akan dikomparasi untuk mengetahui performa yang baik dalam hal akurasi untuk dua dataset yang berbeda yaitu imdb review film dan sentimen twitter. Hasil dari komparasi menunjukkan SVM memperoleh hasil yang baik dalam akurasi pada data imdb review film 78,55% dan pada dataset twitter 72%. Sama halnya dengan NBC yang memperoleh akurasi pada data twitter 78.55% tetapi berbeda pada data twitter 67,33%. Hasil F-Measure review film menunjukan SVM dan NBC memperoleh hasil yang sama yaitu 0,785 dan untuk hasil AUC, NBC mengungguli hasil 0,869, SVM memperoleh hasil 0,786 sedangkan KNN memperoleh hasil 0,572. F-Measure untuk twitter SVM lebih unggul memperoleh hasil 0,720 dan NBC memperoleh hasil 0,673 sedangkan K-NN 0,545. Dan untuk hasil AUC, sama seperti dataset imdb, pada dataset twitter ini NBC juga mengungguli SVM dan K-NN. AUC untuk NBC memperoleh hasil 0,735, SVM memperoleh hasil 0,658 dan K-NN memperoleh hasil 0,618.

Kata kunci: teks mining, sentimen analisis, SVM, Naïve Bayessian, K-NN, komparasi

1. PENDAHULUAN

Lebih dari satu decade terakhir ini, online teks data seperti kicauan di twitter, update status di facebook, review produk di Amazon ataupun refiew film di IMDB dapat menjadikannya sumber informasi yang berharga. Twitter menjadi salah satu social media dalam bentuk teks yang banyak diakses. Jutaan tweet yang berisikan pikiran, pertanyaan, komentar dan kritik dipost setiap harinya. Demikian juga dengan web. Banyak situs yang menyediakan review tentang suatu produk yang dapat mencerminkan pendapat pengguna. Contohnya adalah *Internet Movie Database* (IMDb). IMDb adalah situs yang berhubungan dengan film dan produksi film. Informasi yang diberikan IMDb sangat lengkap. Termasuk siapa saja aktor/aktris yang main di film itu, sinopsis singkat dari film, link untuk trailer film, tanggal rilis untuk beberapa negara dan review dari user-user yang lain. Ketika seseorang ingin membeli atau menonton suatu film, komentar-komentar orang lain dan peringkat film biasanya mempengaruhi perilaku pembelian mereka. Hal ini yang mengarahkan orang melakukan penelitian analisis sentimen. Sumber informasi ini dapat digali sebagai opinion mining.

Opininion mining juga disebut analisis sentimen adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Dilakukan untuk melihat pendapat terhadap sebuah masalah, atau untuk identifikasi kecenderungan hal di pasar. Teks mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi. Teks mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar.

Teks mining merupakan ilmu bidang baru yang sedang berkembang, mengacu pada proses pengambilan informasi berkualitas tinggi dari teks. Informasi ini dapat diperoleh dari peramalan pola dan kecenderungan melalui sarana seperti pembelajaran pola statistik. Tujuan utama dari teks mining adalah untuk memproses *unstructured* data (tekstual) guna dicari pola makna serta ditindaklanjuti dengan pengambilan keputusan yang terbaik.

Secara umum, analisis sentimen menjadi lebih sulit ketika volume teks meningkat, karena hubungan yang kompleks antara kata-kata dan frase. Salah satu masalah pada klasifikasi sentimen teks adalah banyaknya atribut yang digunakan pada sebuah dataset. Pada umumnya, atribut dari klasifikasi sentimen teks sangat besar, dan jika semua atribut tersebut digunakan, dapat mengurangi kinerja dari *classifier.* Atribut yang banyak membuat akurasi menjadi rendah. Untuk mendapatkan akurasi yang lebih baik, atribut yang ada harus dipilih dengan algoritma yang tepat (Chandani, 2015). Seleksi fitur dapat membuat pengklasifikasian lebih efektif dengan mengurangi jumlah data yang dianalisa, maupun mengidentifikasi fitur yang sesuai untuk dipertimbangkan dalam proses pembelajaran.

Metode untuk melakukan seleksi fitur diantaranya Information Gain. Information Gain mengukur berapa banyak informasi kehadiran dan ketidakhadiran dari suatu kata yang berperan untuk membuat keputusan klasifikasi yang benar dalam kelas apapun. Informastion gain adalah salah satu pendekatan filter yang sukses dalam pengklasifikasian teks.

Beberapa algoritma klasifikasi yang dapat digunakan untuk teks mining diataranya adalah Suppot Vector Machine (SVM), Naïve Bayes Classification (NBC) serta K-Nearest Neighbor (KNN). Dari ketiga algoritma tersebut belum ditemukan model yang paling tepat untuk melakukan analisis sentimen (Chandani, 2015). Untuk itu, penelitian ini akan melakukan komparasi terhadap ketiga algoritma tersebut serta pengaruhnya menggunakan seleksi fitur Information Gain terhadap analisis sentimen pada dataset twitter dan IMDB review film.

Penelitian sebelumnya, [1] mengklasifikasikan opini menggunakan naïve bayes dengan seleksi fitur Chi Square, memperoleh hasil klasifikasi Naïve Bayes pada data uji negative dengan ketepatan 72% data uji positif 96%. Secara keseluruhan data uji memperoleh akurasi 83% dan *Fmeasure* sebesar 90,713%. Penelitian yang dilakukan oleh [2] studi alur proses penggunaan Support Vector Machine sebagai klasifikasi teks mining yang handal menyimpulkan Support Vector Machine termasuk algoritma klasifikasi teks mining, teks yang digunakan diolah untuk dapat diklasifikasikan. Dan hasil dari klasifikasi dievaluasi menggunakan hitungan precicion dan recall. Kehandalan ini akan peneliti bandingkan dengan algoritma klasifikasi yang lain. Pada penelitian [3] bertujuan melakukan penelitian menggunakan KNN untuk mengklasifikasikan dokumen teks serta mengimplementasikan penggunakan KNN, Term Graph algorithm, and Naïve Bayes untuk mengklasifikasi dokumen kedalam 5 kategori dengan menggunakan dataset Reuters dalam bentuk file sgml memperoleh hasil akurasinya diantaranya KNN mencapai 98.00 dibanding Term Graph 97.41 dan Naïve 74.68 pada salah satu kategori yaitu Exchange.

Landasan teori yang digunakan dalam penelitian ini antara lain :

2.1. Teks Mining

Text mining adalah proses ekstraksi pola berupa informasi dan pengetahuan yang sebelumnya tidak diketahui pada sejumlah besar sumber data yang berupa teks (chiwara, dkk, 2016). Jenis masukan untuk penambangan teks ini disebut data tak terstruktur dan merupakan pembeda utama dengan penambangan data yang menggunakan data terstruktur atau basis data sebagai masukan. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan penambangan data. Proses yang umum dilakukan oleh penambangan teks di antaranya adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dll (Turban, dkk, 2015).

2.2. Analisis Sentimen

Analisis sentimen merupakan salah satu cabang penelitian text mining (Falahah, 2015). Sentiment analysis atau opinion mining mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan text mining yang bertujuan menganlisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik , produk, layanan, organisasi, individu, ataupun kegiatan tertentu (Liu, 2011).

Analisis sentimen adalah proses memahami, mengekstrak dan mengolah data tekstual

untuk mendapatkan informasi. Hal ini dilakukan untuk mengetahui pendapat terhadap sebuah masalah, atau dapat juga digunakan untuk identifikasi kecenderungan hal di pasar. Banyak sekali manfaat dari analisis sentimen dalam berbagai sudut pandang diantaranya dapat digunakan untuk memperoleh gambaran umum persepsi masyarakat terhadap kualitas layanan, melakukan pemantauan terhadap sebuah produk, prediksi penjualan, politik dan pengambilan keputusan para investor.

2.3. Support Vector Machine

Support Vector Machines (SVM) adalah metode pembelajaran terbimbing yang menganalisis data dan mengenali pola, digunakan untuk klasifikasi dan analisis regresi. Metode ini dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory.

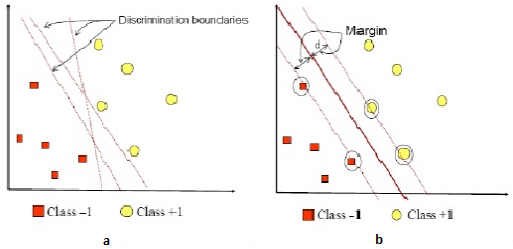
SVM termasuk klasifikasi yang memberi aturan untuk mengubah dokumen teks menjadi vector sebelum diklasifikasi. Biasanya teks dokumen diubah menjadi vector multidimensional tf-idf. Vector dalam penelitian ini memiliki dua komponen yaitu dimensi (word id) dan bobot.

Bobot ditujukan untuk memberikan skor terhadap frekuensi kemunculan sebuah kata. Salah satu metode populer untuk melakukan pembobotan kata adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF merupakan metode pembobotan yang menggabungkan dua konsep, yaitu *Term Frequency* dan *Document Frequency.*

Term Frequency adalah konsep pembobotan dengan mencari seberapa sering (frekuensi) munculnya sebuah term dalam satu dokumen. Karena setiap dokumen mempunyai panjang yang berbeda, kemungkinan terjadi sebuah kata muncul lebih banyak di dokumen yang panjang dibandingkan dengan dokumen yang pendek. Dengan demikian, *term frequency* sering dibagi dengan panjangnya dokumen (total kata yang ada di dokumen tersebut). Sedangkan *Document Frequency* adalah banyaknya jumlah dokumen di mana sebuah term itu muncul. Semakin kecil frekuensi kemunculannya, maka semakin kecil pula nilai bobotnya. Ketika proses perhitungan *term frequency*, semua kata di dalamnya dianggap sama pentingnya. Tetapi terdapat kata yang sebenarnya kurang penting untuk diperhitungkan seperti “di-”, “oleh”, “pada” dan sebagainya. Oleh sebab itu, kata-kata yang kurang penting tersebut perlu dikurangi bobotnya dan menambah bobot kata penting lainnya.

SVM memperkenalkan strategi baru dengan menemukan hyperplane terbaik pada ruang input (input space).

Gambar 1.a memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class : +1 dan –1. Pattern yang tergabung pada class –1 disimbolkan dengan warna merah (kotak), sedangkan pattern pada class +1, disimbolkan dengan warna kuning(lingkaran). Problem klasifikasi dapat diterjemahkan sebagai usaha menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut.



Gambar 1. *SVM Mencari Hyperplane Terbaik yang Memisahkan Kedua Class –1 dan +1*

2.4. Naive Bayesian Classification

Naïve Bayes Classifier (NBC) adalah sebuah pengklasifikasi sederhana berdasarkan penerapan teorema Bayes (dari statistik Bayesian) dengan asumsi independen (naif) yang kuat. Sebuah istilah yang lebih deskriptif untuk model probabilitas yang digaris bawahi adalah "model fitur independen". Dalam terminologi sederhana, sebuah NBC mengasumsikan bahwa kehadiran (atau ketiadaan) fitur tertentu dari suatu kelas tidak berhubungan dengan kehadiran (atau ketiadaan) fitur lainnya.

NBC hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan varian dari variabel) yang diperlukan untuk klasifikasi. Karena variabel diasumsikan independen, hanya varian dari variabel-variabel untuk setiap kelas yang perlu ditentukan dan bukan keseluruhan covariance matrix.

Dalam algoritma naïve bayes classifier setiap dokumen direpresentasikan dengan pasangan atribut “X1, X2, X3…Xn” dimana X1 adalah kata pertama, X2 adalah kata kedua dan seterusnya. Sedangkan **V** adalah himpunan kategori data. Pada saat klasifikasi, algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (Vmap), dimana persamaannya adalah sebagai berikut:

$$V_{map} = \arg\max_{v_j \in V} \frac{P(x_1, x_2, x_3, \ldots, x_n \mid v_j) P(v_j)}{P(x_1, x_2, x_3, \ldots, x_n)} \quad (1)$$

Untuk P(x1, x2, x3, …, xn) nilainya konstan untuk semua kategori (Vj) sehingga persamaan dapat ditulis sebagai berikut :

$$Vmap = \arg\max_{Vj \in V} P(x1, x2, x3, \ldots, xn \mid Vj) P(Vj) \quad (2)$$

Atau dapat disederhanakan menjadi persamaan berikut :

$$Vmap = \arg\max_{Vj \in V} \prod_{i=1}^{n} P(xi \mid Vj) P(Vj) \quad (3)$$

Vj = Kategori data dimana j1 = kategori data sentimen negative,j2 = kategori data sentimen positif, $P(xi \mid Vj)$ = Probabilitas xi pada kategori Vj, $P(Vj)$ = Probabilitas dari Vj

Untuk $P(Vj)$ dan $P(xi \mid Vj)$ dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut :

$$P(Vj) = \frac{|doc\ j|}{|contoh|} \quad (4)$$

$$P(xi \mid Vj) = \frac{n_k + 1}{n + |vocabulary|} \quad (5)$$

Dimana $|doc\ j|$ adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan sedangkan $|contoh|$ adalah banyaknya dokumen dalam contoh yang digunakan untuk pelatihan. n_k adalah jumlah frekuensi kemunculan kata xi dalam dokumen Vj dan $|vocabulary|$ adalah banyaknya kata dalam contoh pelatihan.

2.5. K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah salah satu metode sederhana pada algoritma *machine learning.* Tujuan algoritma ini adalah untuk mengklasifikasikan objek ke dalam salah satu kelas yang telah ditetapkan dari kelompok sampel yang telah dibuat oleh *machine learning*. Algoritma KNN tidak memerlukan penggunaan pelatihan data untuk melakukan klasifikasi. Data pelatihan dapat digunakan selama fase pengujian. KNN didasarkan pada menemukan objek yang paling mirip (dokumen) dari kelompok sampel antara jarak Euclidean (Trstenjak, 2013).
Preprocessing dan persiapan dokumen diikuti oleh prase latih. Algoritma menentukan dokumen-dokumen dasar yang akan dibandingkan dengan masing-masing dokumen baru. Algoritma memeriksa di mana dokumen dikategorikan dengan melihat dokumen-dokumen latih yang paling mirip. Algoritma ini mengasumsikan bahwa mengklasifikasikan dokumen diruang Euclidean sebagai titik. Euclidean jarak adalah jarak antara dua titik dalam ruang Euclidean. Jarak antara dua titik dalam Koordinat p = (x, y) dan q = (a, b) dapat dihitung

$$d(p,q) = d(q,p) = \sqrt{(x-a)^2 + (y-b)^2} \quad (6)$$

3.1. METODE

Metode penelitian yang digunakan yaitu metode eksperimen dengan melakukan observasi terhadap variable-variabel objek yang diteliti. Metode eksperimen merupakan suatu penelitian yang kondisi-kondisi tertentu dikendalikan sehingga satu atau beberapa variabel dapat dikontrol untuk menguji hipotesis.

3.2. Pengambilan Data

Data yang digunakan merupakan data review film dalam teks berbahasa Inggris diambil dari situs http://www.cs.cornell.edu/people/pabo/movie-review-data/. Ada beberapa data dari situs tersebut. Yang dipakai dalam penelitian ini adalah data dengan 1000 sentimen positif dan 1000 sentimen negatif. Dan data yang lainnya adalah data sentimen analisis dari twitter teks berbahasa Inggris. Data ini diambil dari situs http://help.sentiment140.com/for-students/ yang akan menghubungkan ke linkhttp://cs.stanford.edu/people/ alecmgo/trainingandtestdata. zip. Data twitter yang digunakan ada 1200 twitt sentimen analisis yang terdiri dari 600 kelas “0 = negatif” dan “4 = positif”. Data tersebut masih berupa twitt teks. Data kelas negatif disatukan dalam satu folder dan diberi nama 0, sedangkan data kelas positif disatukan dalam satu folder dan diberi nama “4”.

3.3. Preprocessing

Sebelum dikomparasi, dataset dilakukan tekt preprocessing terlebih dahulu. Teks preprocessing itu sendiri bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya.

Tahapan text processing yang dilakukan, diantaranya adalah:

a. Tokenizing

Tokenizing adalah tahap pemotongan dokumen teks bedasarkan tiap kata yang menyusunnya. Potongan kata tersebut disebut dengan token atau term. Pada tahap ini akan dilakukan pengecekan dataset dari karakter pertama sampai karakter terakhir.

b. Filter stopwords

Filter stopword adalah proses menghilangkan kata-kata yang sering muncul namun tidak memiliki pengaruh apapun dalam ekstraksi sentimen suatu review. Kata yang termasuk seperti kata penunjuk waktu, kata tanya, dan sebagainya.

c. Stemming

Stemming yaitu proses pengubahan bentuk kata menjadi kata dasar. Metode pengubahan bentuk kata menjadi kata dasar ini menyesuaikan struktur bahasa yang digunakan dalam proses *stemming.*

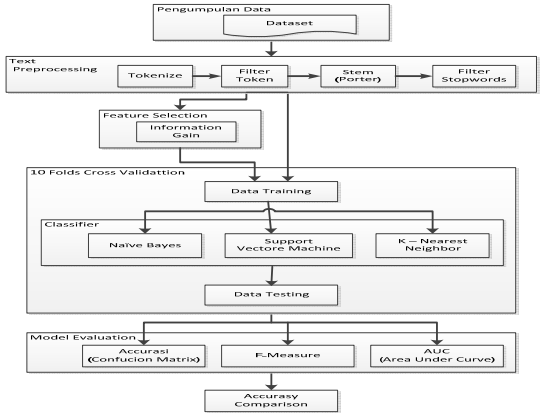
Proses selanjutnya adalah pemilihan fitur (*Feature Selection*). Pada penelitian ini feature selection yang penulis pilih menggunakan *Information Gain* (IG) karena *Information Gain* lebih mudah diterima secara luas sebagai metode penyeleksi fitur terbaik untuk klasifikasi sentimen atau kategorisasi teks.

Setelah dilakukan pemilihan fitur proses selanjutnya yaitu melakukan cross-validation. Cross-validation disini dilakukan untuk menghidari overlapping pada data testing. Di penelitian ini penulis menggunakan validasi standart yaitu 10 folds *cross-validation* dimana proses ini membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya.

Proses yang dilakukan setelah melakukan pengujian yaitu mengukur *performance* dari algoritma klasifikasri teks mining yang dipakai. Dalam penelitian ini *performance* diukur menggunakan 3 cara yaitu *Accuracy*, F-*Measure* dan AUC.

3.4. Alur Penelitian

Adapun alur dari penelitian yang dilakukan dapat dilihat dalam gambar 2 berikut ini.



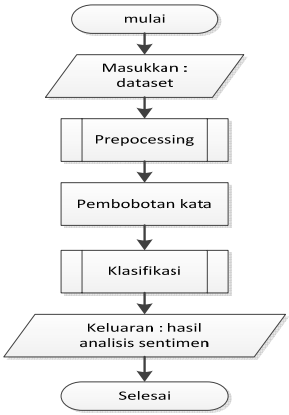
Gambar 2. *Alur Penelitian*

4.0. HASIL DAN PEMBAHASAN

4.1. Manajemen Model

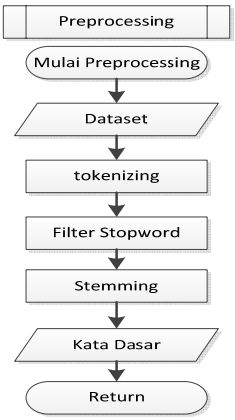
Penelitian ini bertujuan untuk mengkomparasi tiga algoritma klasfiikasi yaitu *Support Vector Machine (SMV), Naïve Bayessian Classification (NBC) , K-Nearest Neighbor (K-NN)* dan pengaruh *Feature Selection Information Gain* pada klasifikasi tekt mining dalam pengaruhnya terhadap sentimen analisis.

Sebelum dikomparasi, dataset dipersiapkan pada *machine learning* yang digunakan yaitu weka dan dilakukan persiapan data preprocessing, yang dilanjutkan dengan pembobotan term / kata. Kemudian dilakukan proses klasifikasi dimana klasifikasi ini mengklasifikasikan dengan tiga algoritma SVM, NB dan K-NN. Gambar 3 menunjukkan diagram alur proses klasifikasi.



Gambar 3. *Diagram Alur Proses Klasifikasi*

Text preprocessing bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya. Gambar 4 menunjukkan potongan diagram alur proses preprossesing.



Gambar 4. *Diagram Alur Preprosesing*

Dari langkah preprocessing ini menghasilkan kata (term). Term-term yang telah melalui proses stemming kemudian dihitung bobotnya dengan menggunakan TF-IDF. Hasil dari perhitungan bobot kemudian disimpan untuk proses selanjutnya yaitu klasifikasi dengan menggunakan algoritma SVM, NBC dan K-NN.

4.2. Teks Preprocessing

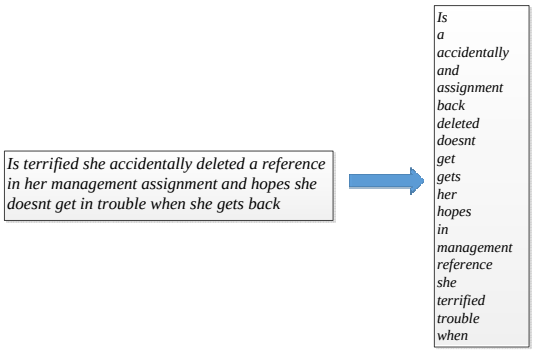
Teks preprocessing bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya. Adapun tahapan dari preprocessing pada penelitian ini meliputi *tokenizing, filtered stopword dan stemming*.

Contoh kalimat dari dataset twitter yang telah diambil peneiti, akan dilakukan preprocessing

"*Is terrified she accidentally deleted a reference in her management assignment and hopes she doesnt get in trouble when she gets back*"

a. *Tokenizing*

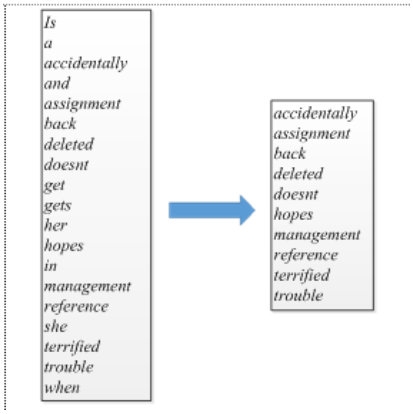
Tokenizing pada contoh kalimat dataset twitter yang teks aslinya berupa kalimat mengalami pemotongan menjadi term. Di gambar 5 berikut ini adalah hasil tokenizing menggunakan tokenizers.WordTokenizer yang telah tersedia pada *machine learning* weka.



Gambar 5. *Proses Tokenizing Opini Twitter*

b. *Filter Stopword*

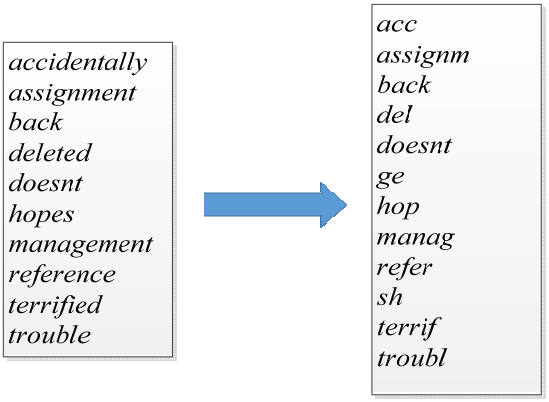
Proses selanjutnya melakukan filter stopword yang hasilnya dapat dilihat pada gambar 6. Hasil ini menggunakan *stopwords.Rainbow* bawaan dari *machine learning* weka.



Gambar 6. *Proses Stopword Opini Twitter*

c. *Stemming*

Hasil dari stemming pada contoh kalimat opini twitter setelah mengalami tokenizing dan filtering dapat dilihat pada gambar 7. Stemmer yang digunakan bawaan *machine learning* weka *stemmers.IteratedLovinsStemmer.*



Gambar 7. **Proses Stemming Opini Twitter**

d. *Pembobotan kata*

Term-term yang telah melalui proses stemming kemudian dihitung bobotnya dengan menggunakan TF-IDF. Bobot ditujukan untuk memberikan skor terhadap frekuensi kemunculan sebuah kata. *Term Frequency* adalah konsep pembobotan dengan mencari seberapa sering (frekuensi) munculnya sebuah term dalam satu dokumen. Karena setiap dokumen mempunyai panjang yang berbeda, kemungkinan terjadi sebuah kata muncul lebih banyak di dokumen yang panjang dibandingkan dengan dokumen yang pendek. Dengan demikian, *term frequency* sering dibagi dengan panjangnya dokumen (total kata yang ada di dokumen tersebut). Sedangkan *Document Frequency* adalah banyaknya jumlah dokumen di mana sebuah term itu muncul. Semakin kecil

frekuensi kemunculannya, maka semakin kecil pula nilai bobotnya. Ketika proses perhitungan *term frequency*, semua kata di dalamnya dianggap sama pentingnya.

Setelah tahapan pada teks preprocessing dilalui, maka melakukan klasifikasi menggunakan algoritma Support Vector Machine, Naïve Bayesian Classification dan K-Nearest neighbor.

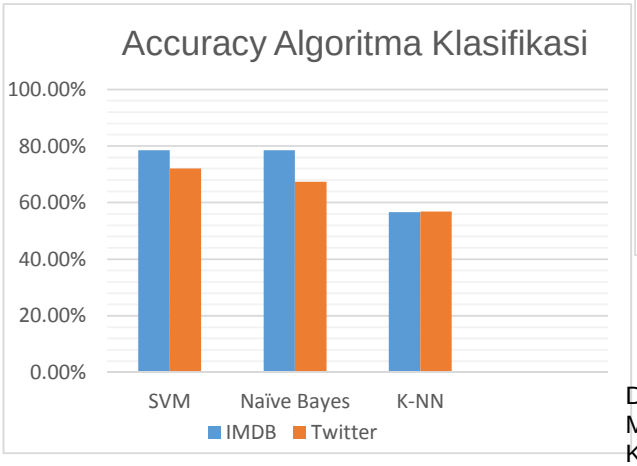
4.3. Analisis Hasil

Hasil Percobaan klasifikasi dengan Metode SVM, NBC dan K-NN

Percobaan dilakukan dengan dua dataset yang berbeda yang mengalami teks preprocessing yang sama. Hasil accuracy eksperimen klasifikasi dengan metode Support Vector Machine, Naïve Bayessian Classification dan K-Nearest Neighbor untuk data sentimen berbahasa Inggris pada dua dataset yang berbeda yaitu IMDB Review Film dan opini twitter ditunjukkan pada tabel 1 dan gambar 8.

Tabel 1. *Komparasi Accuracy Algoritma Klasifikasi Terhadap Dataset IMDB Dan Twitter*

	IMDB Review Film	Twitter
SVM	78,55%	72%
Naïve Bayes	78,55%	67,33%
K-NN	56,7%	56,83%



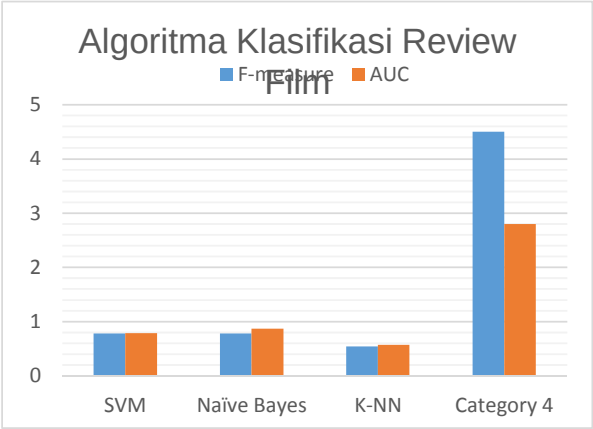
Gambar 8. *Komparasi Akurasi Algoritma Klasifikasi Terhadap Dataset IMDB Dan Twitter*

Dari tabel 1 dan gambar 8 didapatkan perbedaan hasil akurasi yang terjadi apabila dataset menggunakan data review film atau twitter. Hal ini menunjukkan bahwa jenis dataset mempengaruhi hasil akurasi algoritma klasifikasi. Tetapi secara umum jika dilihat dari hasilnya, jenis dataset tidak berpengaruh banyak pada algoritmanya. Untuk data review film, baik metode SVM maupun Naïve Bayes memperoleh hasil akurasi yang sama yaitu 78,55% mengungguli metode K-NN yang hanya memperoleh hasil akurasi 56,7%. Untuk dataset opini twitter, SVM mengungguli metode NBC dan KNN dengan hasil berturut-turut 72% , 67,33% dan 56,83%.

Hasil komparasi SVM, NBC serta K-NN untuk F-Measure dan Area Under Curve (AUC) pada dataset imdb review film dapat dilihat pada tabel 2 dan gambar 8. Sedangkan hasil komparasi SVM, NBC serta K-NN untuk F-Measure dan Area Under Curve (AUC) pada dataset opini twitter dapat dilihat pada tabel 3 dan gambar 10.

Tabel 2. *Komparasi F-Measure Dan AUC Algoritma Klasifikasi Dataset Review Film*

	F-measure	AUC
SVM	0,785	0,786
Naïve Bayes	0,785	0,869
K-NN	0,543	0,572



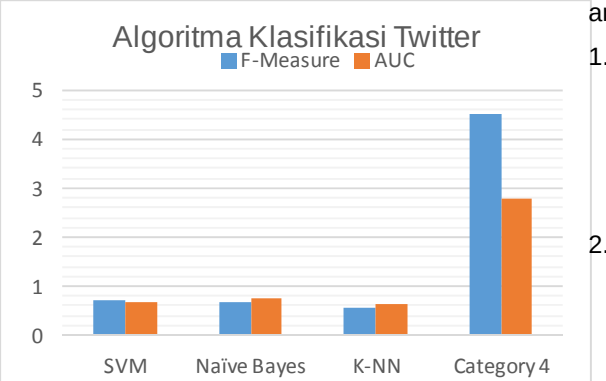
Gambar 9. *Komparasi F-Measure Dan AUC Algoritma Klasifikasi Dataset Review Film*

Dari tabel 2 dan gambar 9 didapat hasil F-Measure komparasi metode SVM, NBC serta KNN untuk dataset review film menunjukakn baik SVM dan NBC memperoleh hasil yang sama yaitu 0,785 sedangkan K-NN hanya 0,543. Dan untuk hasil AUC, NBC mengungguli SVM dan K-NN. AUC untuk NBC memperoleh hasil 0,869 , SVM memperoleh hasil 0,786 sedangkan KNN memperoleh hasil 0,572.

Tabel 3. *Komparasi F-Meassure Dan AUC Algoritma Klasifikasi Terhadap Dataset Twitter*

	F-Measure	AUC
SVM	0,720	0,658

Naïve Bayes	0,673	0,735
K-NN	0,545	0,618



Gambar 10. *Komparasi F-Measure Dan AUC Algoritma Klasifikasi Dataset Twitter*

Dari tabel 3 dan gambar 10 didapat hasil F-Measure komparasi metode SVM, NBC serta KNN untuk dataset opini twitter SVM lebih unggul memperoleh hasil 0,720 dan NBC memperoleh hasil 0,673 sedangkan K-NN hanya 0,545. Dan untuk hasil AUC, sama seperti dataset imdb, pada dataset twitter ini NBC juga mengungguli SVM dan K-NN. AUC untuk NBC memperoleh hasil 0,735 , SVM memperoleh hasil 0,658 dan K-NN memperoleh hasil 0,618.

5.1. SIMPULAN

Berdasarkan hasil dari komparasi algoritma klasifikasi antara Support Vector Machine (SVM), Naïve Bayes Classification (NBC) dan K-Nearest Neighbor (KNN) didapatkan kesimpulan :

1. Pada dataset review film, hasil accuracy sebanding antara Support Vector Machines dengan Naïve Bayes Classification = 78,55% sedangkan pada dataset twitter SVM memperoleh hasil 72%, lebih unggul dibanding NBC dan KNN sebesar 67,33% dan 56,83%.
2. Perbedaan hasil accuracy pada dataset diatas dikarenakan bentuk format dataset berbeda, dimana pada twitter mempunyai keterbatasan karakter dibanding review film.
3. Secara rata-rata dari hasil pengukuran untuk teks mining, baik menggunakan dataset review film maupun twitter, metode Support Vector Machine (SVM) lebih unggul dari Naïve Bayesian Classification (NBC), sedangkan untuk K-Nearest neighbor (KNN) memperoleh hasil yang terpaut jauh dibawah SVM ataupun NBC.

Beberapa saran yang diharapkan dapat dikembangkan pada penelitian selanjutnya antara lain sebagai berikut :

1. Dibutuhkannya penelitian lebih lanjut atau pengembangan untuk penelitian analisis sentimen menggunakan metode pengklasifikasian untuk meningkatkan hasil akurasi, terutama pengolahan di preprocessingnya.
2. Dataset yang digunakan dapat dikembangkan dengan dataset yang lain seperti dataset berbahasa Indonesia atau dataset yang berupa tulisan selain latin.

DAFTAR RUJUKAN

[1] Ling, J., Kencana, I.P.K., Oka, B.O., Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square, E-Jurnal Matematika, ISSN: 2303-1751, Vol. 3 (3), Agustus 2014, pp. 92-99 ISSN: 2303-1751

[2] Patil, G., Galande, V., Kekan, V., Dange, K., 2014, *Sentiment Analysis Using Support Vector Machine, International Journal of Innovative Research in Computer and Communication Engineering*, ISSN(Online): 2320-9801,ISSN (Print): 2320-9798, Vol. 2, Issue 1, January 2014

[3] Bijalwan, V., Kumar, V., Kumari,P.,Pascual, J., *KNN based Machine Learning Approach for Text and Document Mining*, International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.61-70, 2014

[4] Aulianita, R., Komparasi Metode K-Nearest Neighbors dan Support Vector Machine Pada Sentiment Analysis Review Kamera, Journal Speed Sentra Penelitian Engineering dan Edukasi Volume 8 No 3, 2016

[5] Chandani, V., Wahono, R.S., Purwanto, Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film, Journal of Intelligent Systems,ISSN 2356-3982, Vol. 1, No. 1, February 2015

[6] Dan, L., Lihua, L., Zhaoxin, Z., Research Of Text Categorization On WEKA, 2013 Third International Conference on Intelligent System Design and Engineering Applications, 2013

[7] Falahah, Nur, D.D.A., Pengembangan Aplikasi Sentiment Analysis Menggunakan

Metode Naïve Bayes (Studi Kasus Sentiment Analysis dari media Twitter), Seminar Nasional Sistem Informasi Indonesia, 2-3 November 2015
[8] Hamzah, A., Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis, Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST), ISSN:1979-911X, Nopember 2012
[9] Lidya, S.K., Sitompul, O.S., Efendi, S., Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) Dan K-Nearest Neighbor (K-NN), Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015) ISSN: 2089-9815, 28 Maret 2015
[10] Muthia, D.A., Opinion Mining Pada Review Buku Menggunakan Algoritma Naïve Bayes, Jurnal Teknik Komputer Amik Bsi Vol.II No.1 Februari 2016
[11] Rozi, I.F., Pramono, S.H., Dahlan, E.A., Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi, Jurnal EECCIS Vol. 6, No. 1, Juni 2012
[12] Saputra, N., Adji, T.B., Permanasari, A.E., Analisis Sentimen Data Presiden Jokowi Dengan Preprocessing Normalisasi Dan Stemming Menggunakan Metode Naive Bayes Dan Svm, Jurnal Dinamika Informatika, Volume 5, Nomor 1, November 2015
[13] Saraswati, N.W.S., Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis, Seminar Nasional Sistem Informasi Indonesia, 2 - 4 Desember 2013
[14] Trstenjak, B., Mikac,S., Donko, D., KNN with TF-IDF Based Framework for Text Categorization, 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013
[15] Utami, L.D., Wahono, R.S., Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes, Journal of Intelligent Systems,ISSN 2356-3982, Vol. 1, No. 2, December 2015
[16] Purwanti, E., Klasifikafi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour, e-ISSN 2442-5168 Volume 1, Nomor 2, Juli-Desember 2015
[17] Waila, P., Marisha, Singh, V.K., Singh, M.K., Evaluating Machine Learning and Unsupervised Semantic Orientation Approaches for Sentiment Analysis of Textual Reviews, IEEE International Conference on Computational Intelligence and Computing Research, 2012
[18] Wahyuni, E.S., Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara, ISSN: 2252-4983, Jurnal SIMETRIS, Vol 7 No 1 April 2016
[19] Younis, E.M.G., Sentment Analysis and Text Mining For Social Media Microblog using Open Source Tools: An Empirical Study, International Journal of Computer Application (0975-8887), Volume 112 – No. 5, February 2015
[20] Chiwara, M., Al-Ayyoub, M., Sajjad, M., Gupta, H.R., Wasilewska, A., 24 mei 2016, Text Mining, http://www3.cs.stonybrook.edu/~cse634/presentations/ TextMining.pdf
[21] Go, A., Bhayani,R., Huang, L., Agustus 2016, Twitter Sentiment Classification using Distant Supervision, http://cs.stanford.edu/people/alecmgo/papers/ TwitterDistant Supervision09.pdf
[22] Liu, Bing, 31 Mei 2016, Sentiment Analysis And Opinion Mining. Chicago: Morgan & Claypool Publisher, http://www.dcc.ufrj.br/~valeriab/ DTMSentiment-AnalysisAndOpinionMining-BingLiu.pdf.
[23] Mouthami,K., Devi, K.N., Bhaskaran, M., Agustus 2016, Sentiment Analysis and Classification BasedOn Textual Reviews, http://ieeexplore.ieee.org /
[24] Pang, B., & Lee, L., September 2016, A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Association for Computational Linguistics, http://dl.acm.org/citation.cfm?id=1218990
[25] Waila, P., Marisha, Singh, M.K., and Singh, M.K., Agustus 2016, Evaluating Machine Learning and Unsupervised Semantic Orientation Approaches for Sentiment Analysis of Textual Reviews, http:// ieeexplore.ieee.org/
[26] http://help.sentiment140.com/for-students/ diunduh bulan September 2016