Klasifikasi Analisis Sentimen Tweet pada Twitter Menggunakan Metode Multinomial Naive Bayes dengan Seleksi Fitur Mutual Information

Riky Sutriadi Putra(G64144058)*, Julio Adisantoso

Abstrak/*Abstract*

Penelitian ini bertujuan untuk mengklasifikasi data *tweet* pada Twiiter menjadi 3 sentimen yaitu positif, negatif, dan netral. Data yang digunakan terdiri atas 6000 data yang diperoleh dari tags.hawksey.info. Sebelum tahap klasifikasi, dilakukan beberapa tahap yaitu *indexing* seperti tokenisasi, normalisasi kata, pembuangan *stopwords* dan *stemming* serta pemilihan fitur menggunakan *Inverse Document Frequency* dan *Mutual Information*. Data yang dihasilkan setelah proses *indexing* dibagi menjadi dua *subset* data yang terdiri dari 70 persen data latih dan 30 persen data uji. Data latih akan digunakan dalam tahapan pemilihan fitur sementara data uji digunakan untuk melakukan pengujian terhadap sistem klasifikasi yang telah dibuat dengan mengggunakan metode *Multinomial Naïve Bayes*. Adapun manfaat dari penlitian ini adalah mengetahui hasil akurasi terbaik antara metode *Multinomial Naïve Bayes* menggunakan seleksi fitur *Mutual Information* dengan *Inverse Document Frequency* dalam melakukan analisis sentimen.

Kata Kunci

analisis sentimen; klasifikasi; Multinomial Naive Bayes; Mutual Information; tweet; Twitter

***Alamat Email**: rikysutriadiputra@gmail.com

PENDAHULUAN

Latar Belakang

Pertumbuhan pengguna internet memberikan dampak langsung terhadap penggunaan media sosial. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) mengungkapkan jumlah pengguna internet di Indonesia mencapai 88.1 juta orang hingga akhir tahun 2014. Data survei ini menyatakan bahwa ada tiga alasan utama orang Indonesia menggunakan internet. Tiga alasan itu adalah untuk mengakses sarana sosial/komunikasi (72%), sumber informasi harian (65%), dan mengikuti perkembangan jaman (51%). Tiga alasan utama mengakses internet itu dipraktikan melalui empat kegiatan utama, yaitu menggunakan jejaring sosial (87%), mencari informasi (69%), instant messaging (60%) dan mencari berita terbaru (60%).

Twitter merupakan salah satu jejaring sosial yang sering digunakan sebagai alat komunikasi , Twitter juga dimanfaatkan sebagai media untuk promosi, dan kampanye politik. Twitter merupakan jaringan sosial berupa *microlog* serta memiliki karakteristik dan format yang unik dengan simbol ataupun aturan khusus. Disebut *microblog* karena penggunanya hanya dapat mengirim dan membaca pesan blog seperti pada umumnya dengan batas maksimal sejumlah 140 karakter, pesan tersebut dikenal dengan *tweet* (Zhang *et al.* 2011). Setiap *tweet* yang di posting pengguna beraneka ragam sesuai dengan keinginan pengguna. Postingan itu bisa berupa pendapat, saran, ataupun kritikan tentang topik-topik tertentu. Keanekagaraman postingan tersebut serta banyaknya penggunaan bahasa yang tidak baku pada *tweet* menjadi alasan diperlukan analisis sentimen.

Analisis sentimen dapat disebut juga *opinion mining* yang bertujuan untuk menganalisis, memahami, mengolah, dan mengestrak data tekstual yang berupa opini terhadap entitas seperti organisasi dan topik tertentu agar mendapatkan suatu informasi (Liu 2010). Analisis sentimen juga berfokus pada pengolahan opini yang mengandung polaritas, yaitu memiliki nilai sentimen yang positif ataupun negatif (Novantirani, 2014).Masalah yang ada dalam analisis sentimen biasanya sulit dalam mendefinisikan, menentukan konsep masalah, sub masalah, dan tujuan yang berfungsi sebagai kerangka kinerja dalam berbagai penelitian (Liu 2010).

Metode klasifikasi sentimen dapat digunakan untuk menentukan sentimen dari sebuah tweet. Metode klasifikasi Naïve Bayes classifier dapat dibagi menjadi dua, yaitu *multi-variate Bernoulli* dan *Multinomial Naïve Bayes* (Manning *et al.* 2008). Dalam penelitian ini

metode klasifikasi yang digunakan adalah *Multinomial Naïve Bayes*. *Multinomial Naïve Bayes* digunakan karena proses yang sederhana dan mudah diaplikasikan pada berbagai keadaan sehingga tidak akan mengalami kegagalan secara keseluruhan pada hasilnya (Manning *et al.*, 2008). Membuat model klasifikasi diperlukan sebuah training set atau data latih yang akan menghasilkan *term* atau fitur sebagai penciri. Namun tidak semua *term* dapat dijadikan sebagai penciri, sehingga perlu dilakukan seleksi fitur. Tujuan utama dari seleksi fitur adalah efisiensi *term* yang dihasilkan sebagai penciri serta peningkatan akurasi untuk klasifikasi (Manning *et al.* 2008). Seleksi fitur yang digunakan dalam penelitian ini adalah *Mutual Information* (MI).

Menurut Dimastyo (2014) dalam penelitiannya dalam klasifikasi dokumen *email* dengan menggunakan metode peluang *Multinomial Naïve Bayes* maka seleksi fitur dengan *supervised feature selection* yaitu MI lebih bagus dalam melakukan klasifikasi dokumen *spam* dibandingkan dengan *unsupervised feature selection* yaitu *Inverse Document Frequency* (IDF). Adityawan (2014) telah melakukan analisis sentimen dengan proses klasifikasi pada *tweet* menjadi 3 sentimen yaitu positif, negatif, dan netral menggunakan data seimbang. Data yang digunakan terdiri atas 8 entitas berbeda dengan masing-masing entitas setiap sentimennya terdiri atas 80-90 data. Hasil akurasi dari klasifikasi data tweet dengan menggunakan metode *Multinomial Naïve Bayes* adalah 66.42%. Adapun nilai akurasi tiap sentimennya untuk model Multinomial yaitu 58.62% untuk sentimen positif, 77.42% untuk sentimen negatif, 64.40% untuk sentimen netral.

Penelitian ini menggunakan data *tweet* bahasa Indonesia yang akan diklasifikasikan kedalam tiga kelas sentimen positif, negative, dan netral menggunakan seleksi fitur *Mutual Information* dengan menggunakan metode *Multinomial Naïve Bayes*.

Perumusan Masalah

Rumusan malasah dalam penelitian ini adalah:

1. Bagaimana mengklasifikasikan sentimen pada data Twitter menggunakan seleksi fitur *Mutual Information*(MI) dan *Inverse Document Frequency* dengan metode *Multinomial Naïve Bayes*?
2. Apakah seleksi fitur dengan menggunakan MI mampu meningkatkan akurasi jika dibandingkan dengan IDF?

Tujuan

Tujuan dari penelitian ini adalah:

1. Mengklasifikasikan sentimen pada data tweet menggunakan seleksi fitur MI dan IDF dengan metode *Multinomial Naïve Bayes*.
2. Dapat membandingkan hasil akurasi menggunakan seleksi fitur MI dengan IDF.

Manfaat

Manfaat dari penelitian ini adalah:

1. Mengetahui cara mengklasifikasikan sentimen pada data tweet menggunakan seleksi fitur MI dan IDF dengan metode *Multinomial Naïve Bayes* dalam melakukan analisis sentimen.
2. Mengetahui hasil akurasi klasifikasi menggunakan seleksi fitur MI dan IDF dengan metode *Multinomial Naïve Bayes* dalam melakukan analisis sentimen.

Ruang Lingkup

Ruang lingkup penelitian adalah:

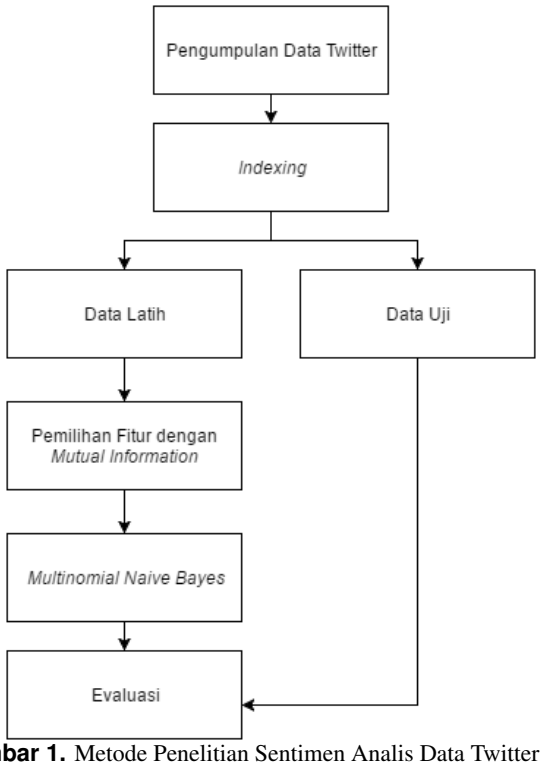
1. Metode klasifikasi yang digunakan adalah *Multinomial Naive Bayes*.
2. Seleksi fitur yang digunakan adalah MI dan IDF.
3. Penelitian ini menggunakan data Twitter dengan kata kunci ”kementerian” dan ”pendidikan”.

METODE PENELITIAN

Metode yang digunakan dalam penelitian ini ada 6 tahap, seperti pada Gambar 1. Tahap pertama adalah pengumpulan data Twitter yang diperoleh dari tags.hawksey.info. Langkah selanjutnya adalah melakukan *indexing* yang terdiri dari tokenisasi, normalisasi kata, pembuangan *stopwords*, dan *stemming*. Setelah proses *indexing* dilakukan, tahap selanjutnya adalah melakukan pembagian data menjadi data latih dan data uji. Pada data latih dilakukan pemilihan fitur dengan menggunakan *Mutual Information* (MI). langkah selanjutnya adalah klasifikasi dengan menggunakan *Multinomial Naïve Bayes*. Tahap terakhir adalah evaluasi hasil klasifikasi.

Pengumpulan Data

Tahapan pertama dalam penelitian ini adalah pengumpulan data. Penelitian ini menggunakan data Twitter bahasa Indonesia. Data yang akan diambil dari Twitter adalah data dengan *term* “Kementrian” dan “Pendidikan”. Pada tahap akuisisi data *tweet*, data diperoleh dari tags.hawksey.info.



Gambar 1. Metode Penelitian Sentimen Analis Data Twitter Bahasa Indonesia

Data yang didapatkan berupa data Excel dengan atribut seperti pada Tabel 1.

Tabel 1. Struktur Data Response Twitter

Atribut	Keterangan
id_str	id dari *post* twitter
from_user	*username* pemakai twitter
text	*post* twitter
created_at	tanggal dan waktu *post* dibuat
geo_coordinates	koordinat tempat *user*
source	tautan profil *user*
profile_image_url	gambar profil dari *user*
user_followers_count	jumlah *follower user*
user_friends_count	jumlah teman *user*
user_location	lokasi dari *user*
status_url	link dari *post* twitter

Dari struktur data yang didapatkan dari sistem tersebut, akan diambil atribut *text* sebagai data untuk diolah sentimennya. Data yang diperoleh dari sistem masih berupa data mentah post user yang belum ada sentimennya. Untuk memberikan sentimen pada data tersebut yang akan digunakan sebagai data latih dilakukan dengan memberikan opini secara manual dengan orang nyata. Jumlah data yang digunakan pada penelitian ini sebanyak 6000 data yang sudah diberi sentimen.

Indexing

Indexing merupakan sebuah proses untuk melakukan pengindeksan terhadap kumpulan dokumen yang akan disediakan sebagai informasi kepada pemakai. Proses pengindeksan bisa secara manual ataupun secara otomatis. Dewasa ini, sistem pengindeksan secara manual mulai digantikan oleh sistem pengindeksan otomatis. Adapun tahapan dari pengindeksan adalah tokenisasi, normalisasi kata, pembuangan *stopwords*, dan *stemming*.

Tokenisasi

Tokenisasi adalah proses pemotongan teks menjadi bagian-bagian yang disebut token, yaitu sebuah *instance* dari urutan karakter dalam beberapa dokumen tertentu yang dikelompokkan bersama sebagai unit semantik yang berguna untuk diproses (Manning *et al.* (2008). Proses memotong dokumen atau kata menjadi bagian-bagian yang lebih kecil disebut token. Token bisa berupa paragraf, kalimat, frasa kata tunggal, sederhana, dan konsep. Teknik yang digunakan dalam proses tokenisasi adalah segmentasi dan memilah. Pada penelitian ini token yang dihasilkan berupa kata tunggal yang nantinya akan menjadi term yang akan digunakan sebagai penciri untuk klasifikasi sentiment pada Twitter. Dalam tahap ini dokumen atau data post Twitter diubah menjadi kumpulan term dengan cara menghilangkan mention, URL, tanda baca dan angka pada tweet. Semua huruf pada tweet diubah menjadi huruf kecil. Contoh dari tokenisasi adalah seperti Tabel 2.

Tabel 2. Tokenizing

Input	Data Twitter	
Output	Data	Twitter

Proses memotong dokumen atau kata menjadi bagian-bagian yang lebih kecil disebut token. token bisa berupa paragraf, kalimat, frasa kata tunggal sederhana dan konsep. teknik yang digunakan dalam proses tokenisasi adalah segmentasi dan memilah. Dalam tahap ini dokumen atau data *post* twitter diubah menjadi kumpulan *term* dengan cara menghilangkan *mention*, URL, tanda baca dan angka pada tweet. Semua huruf pada tweet diubah menjadi huruf kecil. Pada penelitian ini tokenisasi dilakukan dengan menggunakan kode dari Nette yang terdapat di https://github.com/nette/tokenizer.

Penghapusan Stopwords

Stopwords adalah sebuah kata-kata dalam bahasa tertentu yang sangat umum dan memiliki nilai informasi nol. Meyer et.al.(2008). Stopwords didefinisikan sebagai term yang tidak berhubungan (*irrelevant*) dengan dokumen meskipun kata tersebut sering muncul di dalam dokumen. Contoh *stopwords* dalam bahasa Indonesia : yang, juga, dari, dia, kami, kamu, aku, saya, ini, itu, atau, dll. Penghapusan stopwords dilakukan untuk menghilangkan kata dalam daftar kata buang (*stopwords*). Kata tersebut merupakan kata yang jika dihapus tidak mengubah makna dari tweet. Daftar *stopwords* didapatkan dari penelitian Tala (2003) sebanyak 759 kata.

Normalisasi Kata

Menurut Aziz (2013) tahap normalisasi kata dilakukan dengan penggantian kata yang tidak baku menjadi baku, karena kata yang sudah baku akan cenderung lebih kecil ambiguitas dalam pelafalannya dibanding dengan kata yang tidak baku. Misalnya, kata dengan dapat ditulis dengan ”dg” dan ”dgn”. Untuk itu perlu dilakukan normalisasi kata dengan cara mengganti kata yang tidak baku dengan kata yang sesuai konteknya (Sproat et al. 2001). Sebelumnya sudah dibuat terlebih dahulu sebuah kamus yang tidak baku dengan kata bakunya, agar memudahkan dalam fungsi penggantian dan kemudian menggantinya dengan kata baku yang telah ada di dalam kamus tersebut. Dataset kata tidak baku dan kata baku yang digunakan sebanyak 3719 baris data.

Stemming

Stemming adalah proses konversi term ke bentuk umumnya. Dokumen dapat pula diekspansi dengan mencarikan sinonim bagi *term* tertentu di dalamnya. Sinonim adalah kata-kata yang mempunyai pengertian serupa tetapi berbeda dari sudut pandang morfologis. Seperti *stemming*, operasi ini bertujuan menemukan suatu kelompok kata terkait. *Stemming* merupakan salah satu cara yang digunakan untuk meningkatkan performa IR dengan cara mentransformasi kata-kata dalam sebuah dokumen teks ke kata dasarnya (Agusta, 2009). Tahap *stemming* bertujuan untuk mengurangi jumlah kata dan mendapatkan kata dasar yang benar-benar sesuai.Tahap ini menggunakan algoritme Nazief dan Adriani (1996) untuk menghapus berbagai variasi *prefix* (awalan) dan suffix (akhiran). Kamus kata dasar sebanyak 28.526 kata.

Pembuatan Document Term Matrix (DTM)

Menurut Nadilah (2016) tahap pembuatan *term document matrix* (TDM) dilakukan untuk membuat matriks jumlah kemunculan suatu kata pada dokumen. *Document Term Matrix* (DTM) merupakan cara yang paling umum digunakan untuk merepresentasikan text. DTM dapat diekspor dari korpus dan digunakan sebagai mekanisme *bag-of-words*. Pendekatan ini menghasilkan matrik dengan id dokumen sebagai baris dan *term* sebagai kolom. Elemen matrik merupakan frekuensi. Sebagai contoh ada dua dokumen dengan id 1 dan 2 mempunyai kata yang sama yaitu “Nama saya budi dan nama ayah saya budi” dan “nama teman saya budi”.

Pada penelitian ini kolom matriks menunjukkan kata yang ada pada data tweet, sedangkan baris matriks menunjukkan indeks dari dokumen pada kumpulan korpus. Pada penelitian ini satu tweet menandakan satu dokumen.

Pembagian Data

Tahap selanjutnya adalah pembagian data. Data yang dihasilkan setelah proses indexing dibagi menjadi dua subset data yaitu data latih dan data uji dengan perbandingan 70:30. Sebanyak 70 persen data latih dan 30 persen data uji. Data latih ini akan digunakan dalam tahapan pemilihan fitur sementara data uji digunakan untuk melakukan pengujian terhadap sistem klasifikasi yang telah dibuat dalam penelitian ini.

Pemilihan Fitur

Narayanan *et al.* (2013) menyatakan pemilihan fitur merupakan proses menghilangkan fitur yang berlebihan tapi tetap mempertahakan fitur yang memiliki kemampuan disambiguitas. Pemilihan fitur mempunyai dua tujuan utama. Pertama, membuat data latih yang digunakan untuk classifier lebih efisien dengan cara mengurangi ukuran kosakata yang efektif. Kedua, pemilihan fitur biasanya dapat meningkatkan akurasi klasifikasi dengan menghilangkan fitur noise (Narayanan *et al.* 2013). Garnes (2009) menyatakan pemilihan fitur secara umum dibagi menjadi dua metode, yaitu unsupervised feature selection dan supervised feature selection. Unsupervised feature selection adalah sebuah metode pemilihan fitur yang tidak menggunakan informasi kelas dalam data latih ketika memilih fitur untuk classifier. Contoh dari Unsupervised feature selection adalah IDF. Supervised feature selection adalah metode pemilihan fitur yang menggunakan informasi kelas dalam data latih, sehingga untuk menggunakan pemilihan fitur ini harus tersedia se-

buah set pre-classied. Contoh dari supervised feature selection adalah Mutual Information (MI) dan Chi-square.

Inverse document frequency (IDF)

Banyaknya dokumen d yang mengandung term t tertentu disebut DF. Ukuran kepentingan suatu term dari dokumen yang digunakan sebagai penciri adalah nilai DF yang besar, namun nilai dari DF memiliki rentang nilai yang lebar. Inverse document frequency (IDF) adalah inverse dari nilai DF, sehingga ukuran kepentingan suatu term dari dokumen yang akan digunakan penciri yang memiliki nilai kecil dengan rentang yang tidak begitu jauh. Nilai dari IDF disimbolkan dengan idft yang ditulis dengan formula [1]

$$idf_t = log(\frac{N}{df_t}) \tag{1}$$

sedangkan N adalah banyaknya dokumen dan df_t adalah banyaknya dokumen didalam koleksi yang mengandung term tertentu.

Mutual information (MI)

Salah satu seleksi fitur yang sering digunakan untuk menghitung bobot dari fitur adalah *Mutual information*. MI menunjukan seberapa banyak informasi ada atau tidaknya sebuah term memberikan konstribusi dalam membuat keputusan klasifikasi secara benar atau salah. Nilai dari MI disimbolkan dengan notasi I, dimana [2]

$$I(U;C) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1.0\}} P(U=et, C=ec) \; log_2 \frac{P(U=et, C=ec)}{P(U=et)P(C=ec)} \tag{2}$$

sedangkan U adalah variabel acak dengan nilai-nilai et = 1 (dokumen berisi term t) dan et = 0 (dokumen tidak mengandung t), dan C adalah variabel acak dengan nilai-nilai ec = 1 (dokumen di kelas c) dan ec = 0 (dokumen tidak di kelas c). Nilai dari I juga bisa dijabarkan menjadi seperti persamaan [3]

$$\frac{N11}{N} log_2 \frac{NN11}{N1.N.1} + \frac{N01}{N} log_2 \frac{NN01}{N0.N.1} + \frac{N10}{N} log_2 \frac{NN11}{N1.N.0} + \frac{N00}{N} log_2 \frac{NN00}{N0.N.0} \tag{3}$$

dengan N adalah jumlah dokumen yang memiliki nilai-nilai et dan ec yang ditunjukan oleh dua subscript. Sebagai contoh, N10 adalah jumlah dokumen yang mengandung term t (et = 1) dan tidak dalam c (ec = 0). N1. = N10 + N11 adalah jumlah dokumen yang berisi term t (et = 1) dan untuk menghitung dokumen independen keanggotaan kelas (ec = 0, 1). N adalah jumlah total dokumen atau N= N00 + N01 + N10 + N11.

Chi-Square (X^2)

x^2 biasanya digunakan digunakan dalam menguji independensi dari dua variabel yang berbeda. Hipotesis nol jika kedua variabel saling bebas satu sama lain jika nilai dari x^2 tinggi maka hubungan kedua variabel tersebut semakin erat. Dalam seleksi fitur x^2 digunakan untuk mengukur independensi term t dan kelas c. hipotesis yang diuji adalah term dan kelas benar-benar independen, artinya fitur ini tidak berguna untuk mengelompokkan dokumen. Semakin tinggi nilai maka semakin rendah nilai independensinya. Persamaan dari x^2 dapat ditulis sebagai [4]

$$X^2(D,t,c) = \sum_{et\in\{1,0\}} \sum_{et\in\{1,0\}} \frac{(N_{etec} - E_{etec})^2}{E_{etec}} \tag{4}$$

Sedangkan D adalah variabel acak dengan nilai-nilai et = 1 adalah dokumen berisi term t dan et = 0 adalah dokumen yang tidak mengandung t, ec = 1 adalah dokumen di kelas c dan ec = 0 adalah dokumen tidak di kelas c. N adalah frekuensi yang diamati dalam dokumen D dan E adalah frekuensi yang diharapkan. Pengambilan keputusan dilakukan berdasarkan nilai dari masing-masing kata. Kata yang memiliki nilai x2 di atas nilai kritis pada taraf nyata adalah kata yang akan dipilih sebagai penciri dokumen. Kata yang dipilih sebagai penciri merupakan kata yang memiliki pengaruh terhadap kelas c. Nilai kritis x2 untuk taraf nyata α yang digunakan dalam penelitian ini ditunjukkan pada Tabel 3.

Tabel 3. Nilai kritis untuk taraf nyata α

$\propto$	Nilai Kritis
0.050	3.840
0.010	6.630
0.005	7.880

Pada penelitian ini pemilihan fitur yang digunakan adalah MI dan IDF yang nantinya dapat dilihat perbandingan hasil akurasinya.

Klasifikasi

Tahapan klasifikasi ini bertujuan mengklasifikasikan atau mengkategorikan data tweet menjadi tiga kelas sentimen yaitu sentimen positif, negatif dan netral. Fungsi klasifikasi secara umum untuk memetakan suatu dokumen ke dalam kategori tertentu yaitu: [5]

$$\gamma : X -> C \tag{5}$$

Secara umum fungsi ini yang akan dipakai untuk mengelompok data ke dalam himpunan kelas atau kategori yang ada, dengan X adalah kumpulan dokumen dan C merupakan kategori. Ada beberapa algoritme yang dapat dilakukan untuk melakukan klasifikasi data yaitu algoritme berbasis peluang dan algoritme berbasis vektor.

Algoritme berbasis peluang yaitu Bernouli Naïve Bayes dan Multinomial Naïve Bayes.Sedangkan algoritme berbasis vektor yaitu Rocchio dan k nearest neighbor (kNN) (Manning *et al.* 2008). Pada algoritme berbasis peluang, penentuan kelas pada sebuah dokumen atau data adalah dengan cara menghitung peluang keberadaan data tersebut dalam suatu kelas. Sedangkan pada algoritme berbasis vektor, penentuan kelas pada sebuah data dilakukan dengan cara menghitung jarak data tersebut ke centroid suatu kelas.

Algoritme Naïve Bayes dengan model multinomial dan model Bernoulli digunakan karena proses yang sederhana dan mudah diaplikasikan pada berbagai keadaan sehingga tidak akan mengalami kegagalan secara keseluruhan pada hasilnya (Manning *et al.* 2008).Pada model Naïve Bayes Bernouli dokumen diwakili oleh atribut biner yang menunjukkan bahwa ada dan tidak ada term dalam dokumen. Frekuensi kemunculan term dalam dokumen tidak ikut diperhitungkan. Sedangkan Multinomial Naïve Bayes, dokumen diwakili oleh kemunculan term dari dokumen. Pada model ini, sebelumnya dibuat asumsi jika kemunculan masing-masing term t bersifat independen antara satu term dengan yang lainnya. Manning *et al.* (2008) menyatakan dengan menggunakan nilai dari $\hat{P}(c|d)$ peluang suatu dokumen d di dalam kelas c dapat ditulis seperti persamaan [6]

$$\hat{P}(c|d) \propto P(c) \prod_{1 \leq k \leq nd} P(t_k|c) \tag{6}$$

dengan $\hat{P}(t_k|c)$ adalah peluang dari suatu term t_k muncul pada dokumen yang dikeutahui memiliki kelas c. Pendugaan parameter $\hat{P}(t_k|c)$ dihitung dengan cara seperti persamaan [7]

$$\hat{P}(t_k|C) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \tag{7}$$

dengan T_{ct} adalah jumlah kemunculan term t dalam dokumen training yang berada di kelas c. adalah jumlah seluruh term yang muncul berulang kali pada dokumen yang sama.

Manning *et al.* (2008) menyatakan Term tidak selalu muncul pada salah satu kelas saat dilakukan klasifikasi sehinggga nilai $\hat{P}(t_k|c)$ yang dihasilkan adalah nol. Untuk mengatasi permasalahan tersebut, digunakan laplace smoothing, yaitu menambahkan frekuensi term sebanyak 1. sehingga perhitungan dari $\hat{P}(t_k|c)$ menjadi seperti persamaan [8]

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B} \tag{8}$$

Algoritme Rocchio merepresentasikan data ke dalam sebuah vector dengan menggunakan perhitungan jarak atau kemiripan suatu data dengan pusat sebuah kelas. Algoritme Rocchio membagi ruang vektor menjadi beberapa bagian berdasarkan centroid yang ada. Nilai centroid merupakan hasil perhitungan rata- rata jarak pada setiap data atau dokumen. Nilai centroid dapat dihitung dari persamaan [9]

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{9}$$

dimana $|\vec{x} - \vec{y}|$ merupakan jarak data uji x ke kelas y, x_i merupakan data uji ke-i, y_i merupakan kelas ke-i, dan m adalah banyak data. Sedangkan untuk menghitung kemiripan antara dua vektor dokumen, didefinisikan seperti persamaan [10]

$$sim(d_1, d_2) = \frac{\vec{V}(d_1)\vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2|} \tag{10}$$

dimana $sim(d_1, d_2)$ merupakan kemiripan dokumen d_1 dan dokumen d_2 , $\vec{V}(d_1)$ merupakan panjang vektor dokumen d_1 , dan$\vec{V}(d_2)$ merupakan panjang vektor dokumend_2.

Menurut Alkhatib *et al.* (2013) kNN merupakan algoritme klasifikasi berbasis vector yang sederhana.

Algoritme kNN mendefinisikan vektor dokumen yang berdekatan memiliki kelas yang sama. Algoritme ini menggunakan ukuran kemiripan suatu data dengan membandingkan data uji dengan data latih. Perhitungan jarak data uji dengan data latih dapat dihitung dengan menggunakan persamaan Euclidean sebagai berikut

$$|d(x,y)| = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{11}$$

dimana d(x,y) merupakan jarak data uji x ke data latih y, x_i merupakan data uji ke-i, y_i merupakan data latih ke-i, dan m adalah banyak data.

Penelitian ini menggunakan algoritme Multinomial Naïve Bayes dalam mengklasifikasikan data tweet. Penggunaan model klasifikasi ini dikarenakan proses yang sederhana dan mudah diaplikasikan pada berbagai keadaan sehingga tidak akan mengalami kegagalan secara keseluruhan pada hasilnya (Manning *et al.* 2008).

Evaluasi

Langkah selanjutnya adalah melakukan evaluasi. Evaluasi bertujuan untuk mengetahui tingkat akurasi hasil klasifikasi data menggunakan metode Multinomial Naïve Bayes terhadap data uji. Tahap evaluasi dilakukan untuk mengetahui tingkat akurasi dari hasil penggunaan indexing, pemilihan fitur, serta klasifikasi pada data uji. Pengujian penelitian ini dilakukan pada data uji terhadap fungsi klasifikasi yang sudah di training. Token dari hasil seleksi fitur, akan dihitung peluangnya berdasarkan kelas-kelasnya dari dokumen Twitter. Setelah itu, membandingkannya dengan kelas aktual dari data uji dan kelas hasil prediksi dengan menggunakan confusion matrix. Isi dari confusion matrix adalah jumlah kasus-kasus yang telah diklasifikasikan dengan benar dan kasus-kasus yang salah diklasifikasikan.Evaluasi digunakan untuk mengukur efektivitas dari sistem IR. Dua pengukuran yang sering digunakan dalam IR adalah precision dan recall.

Precision

Precision merupakan teknik untuk evaluasi yang didefinisikan sebagai presentase dokumen yang di-retrieve yang benar-benar relevan. [12]

$$\begin{aligned} Precision &= \frac{(relevantitemsretrieved)}{(retrieveditems)} \\ &= P(relevant|retrieved) \end{aligned} \tag{12}$$

Recall

Recall adalah teknik evaluasi untuk menemukan semua item yang relevan dari dalam koleksi dokumen dan idefinisikan sebagai presentase dokumen yang relevan. [13]

$$\begin{aligned} Recall &= \frac{(relevantitemsretrieved)}{(relevantitems)} \\ &= P(retrieved|relevant) \end{aligned} \tag{13}$$

Konsep tersebut dapat diperjelas pada tabel 4.

Tabel 4. Tabel Kontigensi

aasd	Relevant	Not Relevant
Retrieved	True positive (tp)	False positives (fp)
Not Retrieved	False negatives (fn)	True negatives (tn)

Maka didapatkan rumus seperti berikut [14]

$$\begin{aligned} P &= tp/(tp+fp) \\ R &= tp/(tp+fn) \end{aligned} \tag{14}$$

Accuracy

Selain Precision dan Recall, keakuratan sentiment analisis juga dinilai dari akurasinya, untuk menghitung akurasi digunakan rumus seperti berikut dan mengacu tabel 4 [15]

$$accuracy = (tp+tn)/(tp+fp+fn+tn) \tag{15}$$

F-measure

Pengukuran yang lain adalah F-measure yang merupakan *weighted harmonic mean* dari *precision* dan *recall*. [16]

$$\begin{aligned} F &= \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{R}} \\ &= \frac{(\beta^2+1)PR}{\beta^2 P + R} \\ where \beta^2 &= \frac{1-\alpha}{\alpha} \end{aligned} \tag{16}$$

Dimana$\alpha \in [0,1]$ dan$\beta^2 \in [0, \Psi]$

Balanced F-measure menyamakan bobot dari precision dan recall, yang berarti membuat α = 1/2 atau β =

1. Ketika menggunakan $\beta = 1$, formula dapat disederhanakan sebagai berikut [17]

$$F_\beta = \frac{2PR}{P+R} \tag{17}$$

dimana P adalah *precision* , dan R adalah *recall*.

Jadwal Kegiatan

Penelitian ini akan dilakukan selama 4.5 bulan dengan rincian kegiatan seperti tercantum pada Tabel 5.

DAFTAR PUSTAKA

Adityawan, E. 2014. "Analisis Sentimen Dengan Klasifikasi Naïve Bayes Pada Pesan Twitter Menggunakan Data Seimbang". Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Alkhatib, K, Najadat H, Hmeidi I, dan Shatnawi MKA. 2013. "Stock Price Prediction Using k-Nearest Neighbor (kNN) Algorithm".

Aziz, ATA. 2013. "Sistem pengklasifikasian entitas pada pesan twitter menggunakan ekspresi reguler dan naïve Bayes". Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Dimastyo, JG. 2014. "Pengukuran Kinerja Spam Filter dengan Seleksi Fitur yang berbeda menggunakan Fungsi Klasifikasi Multinomial Naïve Bayes". Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Liu, Bing. 2010. "Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing". Chicago (US): University of Illinois.

Manning, Christopher D., Prabhakar Raghavan, dan Hinrich Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press Cambridge, England.

Narayanan, V, I Arora, dan A Bhatia. 2013. "Fast and accurate sentiment classification using an enhanced Naive Bayes model". Department of Electronics Engineering, Indian Institute of Technology (BHU), Varanasi, India.

Zhang *et al.* 2011. "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis".

Tabel 5. Rencana Jadwal Penelitian

Kegiatan	1		2				3				4				5			
	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Penyusunan Proposal Skripsi	■	■	■	■														
Kolokium					■													
Perbaikan proposal						■												
Pengambilan data lapangan						■	■	■										
Pengolahan dan analisis data								■	■	■	■							
Penulisan draft skripsi												■	■	■	■			
Uji petik																■		
Sidang skripsi																	■	
Perbaikan laporan penelitian																		■