

Analisis Sentimen Data Twitter menggunakan Naive Bayes dengan Negation Handling

Aswinda Prima Putra(G64144007)*, Julio Adisantoso

Abstrak/Abstract

Analisis Sentimen dalam penelitian ini akan menggunakan data twitter dengan term yang berkaitan dengan kementerian dan pendidikan. Klasifikasi yang akan digunakan dalam penelitian adalah *multinomial naive bayes*. Masalah utama yang dihadapi dalam klasifikasi sentimen adalah bagaimana cara menangani negasi. Dalam seleksi fitur pada penelitian ini akan ditambahkan metode *Negation Handling*. *Negation Handling* akan diimplementasikan dengan data Twitter berbahasa Indonesia.

Kata Kunci

analisis sentimen; bahasa indonesia; kementerian; negation handling; opini; pendidikan; twitter;

*Alamat Email: aswinda.pp@gmail.com

PENDAHULUAN

Latar Belakang

Microblogging saat ini telah menjadi alat komunikasi yang sangat populer di kalangan pengguna internet. Jumlah penggunaannya membagikan opini mereka tentang berbagai macam aspek dari kehidupan mereka ataupun membahas isu-isu saat ini. Oleh karena itu microblogging merupakan sumber data yang sangat kaya untuk melakukan opinion mining dan sentimen analisis **PAK10.385** (PAK10.385). Berbagai macam perusahaan sering menggunakan survey online atau dengan kertas untuk mengumpulkan opini dari penggunaannya. Namun dengan kemunculan sosial media, orang cenderung lebih memilih mengutarakan pendapatnya melalui facebook, twitter atau sosial media yang lainnya Anwar (2015). Salah satu situs microblogging yang populer adalah Twitter. Twitter merupakan situs web yang dioperasikan oleh Twitter, Inc yang memberikan jaringan sosial berupa microblog serta memiliki karakteristik dan format yang unik dengan simbol ataupun aturan khusus. Disebut microblog karena penggunaannya hanya dapat mengirim dan membaca pesan blog seperti pada umumnya dengan batas maksimal sejumlah 140 karakter, pesan tersebut dikenal dengan tweet (L.Zhang,2011). Twitter digunakan dalam penelitian ini karena sifat dari semua tweetnya bersifat publik. Berbeda dengan sosial media lain seperti facebook yang karakteristik dari *post* nya dapat dirubah menjadi privat atau hanya di lingkaran teman nya saja. Kementerian Komunikasi dan Informatika Indonesia menyampaikan bahwa Indonesia menempati peringkat ke lima terbe-

sar di dunia dari jumlah penggunaannya, dan berdasarkan data dari PT Bakrie yang disampaikan oleh kominfo, Indonesia memiliki 19,5 juta pengguna. Data ini menjadi dasar untuk dilakukannya penelitian pada Twitter berbahasa Indonesia. Pada penelitian Institute for Development of Economics and Finance (Indef) pada tahun 2015, berhasil menjaring 12 juta tweet terkait pemerintahan dan 150 ribu diantaranya memiliki tema pembangunan (Tempo 2015). Banyaknya jumlah tweet terkait pemerintahan khususnya dibidang kementerian dan pendidikan inilah yang mendorong dilakukannya penelitian ini dengan menyertakan kata tersebut tersebut sebagai kata kunci dalam pengumpulan data. Sentimen analisis atau *opinion* mining adalah studi komputasional dari opini-opini orang, sentimen dan emosi melalui entitas dan atribut yang dimiliki yang diekspresikan dalam bentuk teks **LIU2012** (LIU2012). Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang diungkapkan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral (**PANG2002**). Salah satu teknik pembelajaran mesin untuk analisis sentimen adalah Naïve Bayes classifier (NBC). NBC merupakan teknik pembelajaran mesin yang berbasis probabilitas. NBC adalah metode sederhana tetapi memiliki akurasi serta performansi yang tinggi dalam pengklasifikasian teks Routray (2013). Xhemali (2010) melakukan perbandingan antara tiga metode. Metode-metode tersebut adalah Naïve Bayes, Pohon Keputusan, dan Neural Networks. Hasil penelitian secara keseluruhan menunjukkan

bahwa Naïve Bayes classifier adalah pilihan terbaik untuk pelatihan domain. Penggunaan model Naïve Bayes Classifier ini dikarenakan proses yang sederhana dan mudah diaplikasikan pada berbagai keadaan sehingga tidak akan mengalami kegagalan secara keseluruhan pada hasilnya (Manning et al. 2008). Masalah utama yang dihadapi dalam klasifikasi sentimen adalah bagaimana cara menangani negasi. Dalam penelitian Narayanan (2013) Negation Handling Diimplementasikan dalam klasifikasi sentiment menggunakan naive bayes dan hasilnya menunjukkan peningkatan akurasi sebesar 1%. Data yang digunakan dalam penelitian tersebut adalah data review film. Dalam penelitian ini akan berfokus pada metode Negation Handling pada data twitter berbahasa Indonesia.

Rumusan Masalah

Rumusan Masalah dari penelitian adalah:

1. Bagaimana mengimplementasikan Metode Negation Handling pada analisis sentimen Twitter berbahasa Indonesia ?
2. Apakah metode Negation Handling dapat meningkatkan akurasi sentimen analisis dibandingkan dengan tanpa menggunakan Negation Handling ?

Tujuan

Tujuan penelitian adalah:

1. Mengimplementasikan sentimen analisis dengan menggunakan metode Negation Handling pada data Twitter berbahasa Indonesia.
2. Membandingkan akurasi dari sentimen analisis dengan Negation Handling dan tanpa menggunakan Negation Handling.

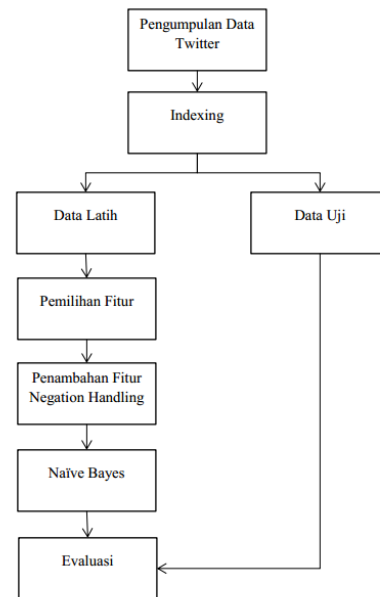
Manfaat

Hasil penelitian diharapkan dapat membantu kementerian pendidikan dalam menangani isu-isu negatif terkait dengan pendidikan dengan memanfaatkan sentimen analisis untuk mencari isu negatif tentang 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.



Gambar 1. Metode Penelitian Sentimen Analisis Data Twitter Bahasa Indonesia

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

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.

Tabel 1. Struktur Data Response Twitter

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

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)). 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	Analisis Sentimen	
Output	Analisis	Sentimen

Pada penelitian ini tokenisasi dilakukan dengan menggunakan kode dari Nette yang terdapat di <https://github.com/nette/tokener>

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 konteksnya (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 mencari 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 sub-set 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 mempertahankan 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-classified. 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 idf_t yang ditulis dengan formula [1]

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (1)$$

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 menunjukkan seberapa banyak informasi ada atau tidaknya sebuah term memberikan kontribusi dalam membuat keputusan klasifikasi secara benar atau salah. Nilai dari MI disimbolkan dengan notasi I , dimana [2]

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

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]

$$\begin{aligned} & \frac{N11}{N} \log_2 \frac{N11}{N1.N.1} + \frac{N01}{N} \log_2 \frac{N01}{N0.N.1} + \\ & \frac{N10}{N} \log_2 \frac{N11}{N1.N.0} + \frac{N00}{N} \log_2 \frac{N00}{N0.N.0} \end{aligned} \quad (3)$$

dengan N adalah jumlah dokumen yang memiliki nilai-nilai et dan ec yang ditunjukkan oleh dua subscript.

Sebagai contoh, N_{10} adalah jumlah dokumen yang mengandung term t ($et = 1$) dan tidak dalam c ($ec = 0$). $N_{11} = N_{10} + N_{11}$ 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 $N = N_{00} + N_{01} + N_{10} + N_{11}$.

Chi-Square (χ^2)

χ^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 χ^2 tinggi maka hubungan kedua variabel tersebut semakin erat. Dalam seleksi fitur χ^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 χ^2 dapat ditulis sebagai [4]

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

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 χ^2 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 χ^2 untuk taraf nyata α yang digunakan dalam penelitian ini ditunjukkan pada Tabel 3.

Tabel 3. Nilai kritis untuk taraf nyata α

α	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]

$$\gamma: X \rightarrow C \quad (5)$$

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]

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

dengan $\hat{P}(t_k|c)$ adalah peluang dari suatu term t_k muncul pada dokumen yang diketahui memiliki kelas

c. Pendugaan parameter $\hat{P}(t_k|c)$ dihitung dengan cara seperti persamaan [7]

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

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 sehingga 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]

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

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]

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

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]

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

dimana $\text{sim}(d_1, d_2)$ 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 dokumen d_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

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

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]

$$\begin{aligned} \text{Precision} &= \frac{(\text{relevant items retrieved})}{(\text{retrieved items})} \\ &= P(\text{relevant} | \text{retrieved}) \end{aligned} \quad (12)$$

Recall

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

$$\begin{aligned} \text{Recall} &= \frac{(\text{relevant items retrieved})}{(\text{relevant items})} \\ &= P(\text{retrieved} | \text{relevant}) \end{aligned} \quad (13)$$

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]

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

Accuracy

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

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

F-measure

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

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

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

Balanced F-measure menyamakan bobot dari precision dan recall, yang berarti membuat $\alpha = 1/2$ atau $\beta =$

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

$$F_\beta = \frac{2PR}{P + R} \quad (17)$$

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

[illegible]