Analisis Sentimen dengan Klasifikasi Rocchio pada Data Twitter Bahasa Indonesia

Jeanette Claudya Weya Pantouw(G64144029)*, Julio Adisantoso

Abstrak/*Abstract*

Berbagai opini masyarakat yang muncul di Twitter khususnya di bidang pemerintahan dan pendidikan dapat dijadikan bahan informasi untuk melihat nilai sentimen mengenai pemerintahan di masyarakat. Penelitian ini menganalisis sentimen masyarakat terhadap kementrian dan pendidikan di Indonesia. Penelitian ini melakukan klasifikasi orientasi sentimen dalam 3 jenis yaitu positif, negatif dan netral menggunakan metode klasifikasi Rocchio. Metode Rocchio akan digunakan untuk mengklasifikasikan data tweet, dengan menggunakan pendekatan berdasarkan kedekatan (*similarity*).

Kata Kunci

Rocchio; Analisis Sentimen; Twitter

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

PENDAHULUAN

Latar Belakang

Analisis sentimen adalah bidang ilmu yang menganalisis opini, penilaian serta sentimen terhadap suatu isu tertentu Liu (2012). Analisis sentimen juga dapat digunakan sebagai penentuan keputusan terhadap suatu isu atau masalah. Opini – opini yang selanjutnya akan digunakan sebagai data untuk penentuan keputusan tehadap isu yang ada. Analisis sentimen juga memegang peranan pada pengolahan opini yang mengandung polaritas, yaitu memiliki nilai sentimen yang positif ataupun negatif (Novantirani, 2014). Sosial media merupakan tempat yang memungkinkan semua orang untuk mengekspresikan opini mereka ke publik Liu (2012). Menurut Semiocast, lembaga riset media sosial yang berpusat di Paris, Prancis, jumlah pemilik akun Twitter di Indonesia merupakan yang terbesar kelima di dunia, dan berada pada posisi ketiga negara yang paling aktif mengirim pesan Twitter (tweet) perhari (Tempo 2012). Banyaknya pengguna Twitter dan adanya kemudahan dalam penyampaian opini melalui media ini, maka data opini berupa tweet tersebut yang kemudian dapat menjadi peluang dan dapat dimanfaatkan sebagai bahan penilaian, tingkat kepuasan dan evaluasi (Novantirani, 2014). Hal ini mendorong beberapa instansi atau kelompok tertentu untuk mendapatkan suatu informasi terkait isu yang akan dianalisis.

Opini masyarakat dari twitter inilah yang akan digunakan selanjutnya menjadi data penelitian ini. Data tweet yang kemudian akan diolah menjadi data yang mengandung sentimen. Penelitian ini melakukan analisis terhadap data tweet terkait isu mengenai kementrian dan pendidikan. Data tweet inilah yang akan digunakan sebagai bahan penelitian untuk penilaian, tingkat kepuasan serta evaluasi kinerja pemerintahan khususnya di bidang kementrian dan pendidikan. Isu inilah yang akan menjadi kata kunci pengambilan data yang akan diolah selanjutnya. 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 kementrian dan pendidikan inilah yang mendorong dilakukannya penelitian ini dengan menyertakan kata tersebut tersebut sebagai kata kunci dalam pengumpulan data.

Untuk dapat mengetahui informasi, data tweet perlu diolah yang selanjutnya dilakukan klasifikasi untuk mengetahui apakah isu tersebut masuk ke dalam sentimen positif, negatif, atau netral. Proses klasifikasi ini dapat dilakukan dengan beberapa metode. Metode klasifikasi yang umumnya digunakan yaitu berbasis peluang dan berbasis vektor. Untuk klasifikasi berbasis peluang metode yang dapat digunakan diantaranya naïve bayes dengan pemodelan bernauli dan multivariant. Sedangkan untuk klasifikasi berbasis vektor, metode yang dapat digunakan diantaranya KNN dan Rocchio.

Penelitian analisis sentimen sebelumnya juga dilakukan

oleh Adityawan 2014 (2014) mengenai klasifikasi Naïve Bayes pada pesan twitter menggunakan data seimbang belum menunjukan akurasi yang cukup baik yaitu 66.42% untuk model Multinomial dan 71.09% untuk model Bernoulli. Untuk itu pada penelitian ini, akan dilakukan penelitian menggunakan metode yang berbeda, yaitu metode berbasis vektor untuk menganalisis apakah hasil klasifikasi metode berbasis vektor memiliki akurasi yang lebih baik dari berbasis peluang.

Penelitian ini menggunakan metode klasifikasi Rocchio dengan pendekatan kesamaan (similarity) dan menggunakan tiga kategori sentimen yaitu netral, postif, dan negatif. Metode ini yang kemudian akan digunakan apakah data tweet masuk ke dalam sentimen positif, negatif ,atau netral. Rocchio dengan pendekatan similarity yang digunakan karena similarity menghitung berdasarkan kedekatan dokumen. Dari sinilah selanjutnya dapat diperoleh kesimpulan data tweet yang ada terkait isu mengenai kementrian dan pendidikan mendapatkan sentimen seperti apa di kalangan masyarakat.

Rumusan Masalah

Berdasarkan latar belakang, perumusan masalah dalam penelitian ini adalah:

1. Apakah Rocchio dapat meningkatkan komputasi tanpa mengurangi akurasi?
2. Bagaimana metode Rocchio diimplementasikan pada data twitter berbahasa Indonesia?
3. Bagaimana perbandingan akurasi metode Rocchio dengan Multinomial naïve bayes ?

Tujuan

Tujuan penelitian ini adalah:

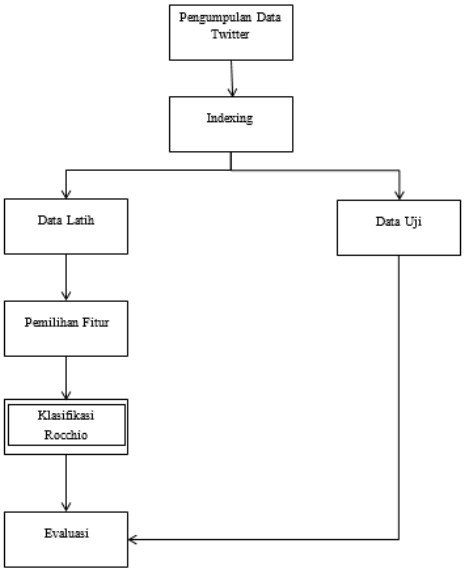
1. Apakah Rocchio dapat meningkatkan komputasi pada proses klasifikasi tanpa mengurangi akurasi?
2. Bagaimana metode Rocchio diimplementasikan pada analisis sentimen twitter berbahasa Indonesia?
3. Bagaimana perbandingan akurasi hasil klasifikasi metode Rocchio jika dibandingkan dengan nenggunakan Multinomial naïve bayes ?

Manfaat

Penelitian ini diharapkan dapat membantu entitas yang ingin mengetahui isu tertentu dari data Twitter. Penelitian ini juga diharapkan dapat memberikan informasi apakah isu tersebut mengandung sentiment positif, negative, atau netral.

METODE PENELITIAN

Penelitian ini diawal dengan pengumpulan data selanjutnya masuk tahap ke indexing, kemudian membagi data tersebut menjadi data latih dan data uji. Data latih kemudian akan di klasifikasikan menggunakan metode klasifikasi Rocchio. Tahapan terakhir yaitu mengevaluasi hasil klasifikasi metode rochhio dengan data uji. Skema tahapan analisis sentimen dapat dilihat pada Gambar 1



Gambar 1. Tahapan Penelitian Metode Klasifikasi Rocchio pada Sentimen Analisis Data Twitter

Pengumpulan Data

Tahapan dalam penelitian analisis sentimen ini yaitu diawali dengan pengumpulan data twitter. Data yang digunakan merupakan data *post user*. *Post* atau pesan dalam twitter dikenal dengan sebutan tweet (Zhang et al. 2011). Data yang akan diambil dari twitter adalah data dengan kata kunci “Kementrian” ,”mentri”, “Pendidikan”, “Sekolah”, dan “Indonesia”. Kata kunci tersebut digunakan untuk mengambil data terkait opini – opini masyarakat di bidang kementrian dan pendidikan. Pada tahap akuisisi data tweet, data diperoleh dari tags.hawksey.info. Data yang didapatkan berupa data excel dengan atribut seperti yang terlihat pada Tabel 1.

Tabel 1 merupakan informasi struktur data Twitter yang diperoleh dari tags.hawksey.info. Data yang diperoleh dari sistem masih berupa data mentah post user yang belum ada sentimennya. Data dengan atribut text yang akan diambil untuk diproses selanjutnya. Data atribut text ini yang selanjutnya akan diolah sentimennya. Proses pengolahan data mentah menjadi sentimen

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

dilakukan secara manual. Jumlah data yang digunakan pada penelitian ini sebanyak 6000 data yang sudah diberi sentimen. Dari data tersebut akan dibagi menjadi dua, yaitu data latih sebanyak 70% dan data uji sebanyak 30%.

Indexing

Setelah data didapatkan dan memiliki sentimen tahap selanjutnya yaitu *indexing*. *Indexing* merupakan proses persiapan yang dilakukan terhadap dokumen sehingga dokumen siap untuk diproses. Proses *indexing* dibagi menjadi dua proses, yaitu *document indexing* dan *term indexing*. Dari *term indexing* akan dihasilkan koleksi kata yang akan digunakan untuk meningkatkan performansi pencarian pada tahap selanjutnya. Selain itu, teknik *indexing* ini juga dilakukan agar hasil yang diperoleh lebih baik. Karena kebanyakan *tweet* hanya berisi tautan dan tidak menunjukkan sentimen tertentu, dan penulisannya ditulis dalam bahasa asing (Parikh dan Movassate 2014). Bahasa asing yang dimaksud dalam penelitian adalah kata – kata yang dalam penulisannya menggunakan pengabungan antara alfanumerik dan simbol – simbol lainnya. Ada beberapa tahapan yang dilakukan didalam proses *indexing* diantaranya *tokenizing*, pengahapusan *stopwords*, normalisasi kata, *stemming*, dan pembuatan *document term matrix*.

Tokenizing

Tahap awal dalam proses *indexing* adalah *tokenizing*. Pada tahap ini setiap data *post twitter* yang berupa kata – kata akan diubah menjadi kumpulan term, dengan tidak menyertakan mention, URL, tanda baca, dan angka pada *tweet*. Selain itu, semua data pada tweet juga akan diubah menjadi huruf kecil. 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. Sebagai contoh, jika ada masukkan teks “Pendidikan di Indonesia sangat buruk dan sungguh memprihatinkan”, maka hasil keluaran dari proses *tokenizing* adalah seperti yang disajikan pada Tabel 2.

Tabel 2. Tokenizing

Input	Data Twitter	
Output	Data	Twitter

Tabel 2 merupakan contoh hasil proses *tokenizing*, setiap kalimat yang ada akan dipilah menjadi potongan – potongan kata. Pada penelitian ini *tokenizing* dilakukan dengan menggunakan kode dari Nette yang didapat dari https://github.com/nette/tokenizer.

Penghapusan Stopwords

Stopwords merupakan kata – kata atau term yang tidak berhubungan dan tidak memiliki makna atau informasi yang berhubungan dengan dokumen, walaupun kata tersebut sering muncul pada dokumen. *Stopwords* adalah sebuah kata-kata dalam bahasa tertentu yang sangat umum digunakan dan memiliki nilai informasi nol (Meyer et.al . 2008). Penghapusan kata tersebut tidak akan mengubah makna dan isi dari informasi tweet, beberapa contoh stopwords dalam bahasa Indonesia diantaranya: yang, juga, dari, dia, kami, kamu, aku, saya, ini, dan itu. Pada penelitian ini digunakan dataset daftar *stopword* yang didapatkan dari penelitian Tala (2003) sebanyak 759 kata. Dataset penelitian Tala inilah yang nantinya akan digunakan untuk menghapus *stopword* pada data *tweet*.

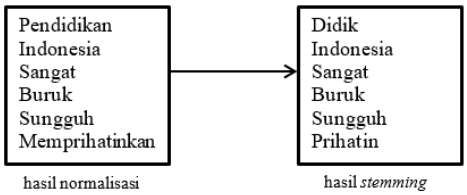
Normalisasi Kata

Normalisasi kata merupakan proses penggantian kata yang tidak baku menjadi kata baku. Normalisasi ini dilakukan untuk mempermudah proses penelitian pada tahapan selanjutnya. Kata baku akan cenderung lebih kecil ambiguitas dibandingkan dengan kata yang tidak baku, untuk itu normalisasi dilakukan untuk mentranslasi kata tidak baku menjadi baku (Aziz 2013). Untuk itu perlu dilakukan normalisasi kata dengan cara mengganti kata yang tidak baku (Sproat et al. 2001). Pada penelitian ini proses pengantian kata tidak baku menjadi baku menggunakan dataset yang sudah ada. Dataset yang digunakan adalah sebuah kamus yang berisi kumpulan

data tidak baku dengan kata bakunya. Hal ini dilakukan untuk memudahkan proses penggantian kata. Dataset kata tidak baku dan kata baku yang digunakan sebanyak 3719 baris data.

Stemming

Stemming merupakan proses transformasi kata – kata menjadi kata dasarnya dalam sebuah teks dokumen, hal ini juga digunakan untuk meningkatkan performa IR (Agusta, 2009). *Stemming* adalah proses konversi term ke bentuk umumnya. Tidak hanya ditransformasi menjadi kata dasar, tetapi kata – kata juga dapat ditransformasikan ke dalam bentuk sinonim kata tersebut. Sinonim adalah kata yang memiliki kesamaan makna tetapi berbeda dari sudut pandang morfologis. Contoh stemming dapat dilihat pada gambar 2



Gambar 2. Stemming

Tahap *stemming* bertujuan untuk mengurangi jumlah kata dan mendapatkan kata dasar yang benar-benar sesuai. Untuk itu penghapusan kata dan berbagai variasi lainnya seperti *prefix* dan *suffix* perlu dilakukan. Penelitian ini menggunakan algoritma Nazief dan Adriani (1996) dan kamus kata dasar yang digunakan sebanyak 28.526 kata.

Pembuatan Document Term Matrix (DTM)

Term Document Matrix (TDM) dilakukan untuk menghitung jumlah kemunculan kata pada dokumen (Nadilah 2016). Cara yang paling umum untuk merepresentasikan teks ke dalam bentuk matrik adalah melalui pembuatan DTM. 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. Setiap elemen matrik yang ada merupakan representasi dari frekuensi kemunculan kata. Sebagai contoh ada dua dokumen dengan id 1 dan 2 mempunyai kata yang sama yaitu “Saya suka makan nasi dan saya suka ayam goreng” dan “ayam goreng”. Tabel 3 menunjukkan contoh DTM yang terbentuk. .

Sedangkan pada penelitian ini, kolom matriks menunjukkan kata yang ada pada data *tweet*, dan baris matriks menunjukkan indeks dari dokumen pada kumpulan

Tabel 3. Document Term Matrix

ID	saya	suka	makan	nasi	dan	ayam
1	2	2	1	1	1	1
2	0	0	0	0	1	1

korpus. Pada penelitian ini satu tweet menandakan satu dokumen.

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 untuk tahapan selanjutnya sementara data uji digunakan untuk melakukan pengujian terhadap sistem klasifikasi yang telah dibuat dalam penelitian ini.

Pemilihan Fitur

Pemilihan fitur merupakan proses pemilihan term yang mewakili informasi penting dari suatu dokumen atau teks. Adanya pemilihan fitur ini dapat meningkatkan akurasi karena adanya seleksi pada *term* yang bukan merupakan penciri (Manning et all. 2008). Menurut Garnes (2009) pemilihan fitur secara umum dibagi menjadi dua metode, yaitu *unsupervised feature selection* dan *supervised feature selection*. Metode *Unsupervised feature selection* tidak menggunakan informasi kelas dalam data latih ketika memilih fitur untuk *classifier*, contohnya adalah *Inverse Document Frequency* (IDF). Sedangkan *Supervised feature selection* adalah metode seleksi fitur yang menggunakan informasi kelas dalam data latih, sehingga *set pre-classied* harus tersedia agar seleksi fitur dapat dilakukan. Pada penelitian ini pemilihan fitur yang akan digunaka yaitu IDF. IDF dipilih karena metode ini efisien, mudah dan memiliki hasil yang akurat (Robertson 2005).

Inverse document frequency (IDF)

Inverse Document Frequency (IDF) merupakan salah satu metode yang digunakan dalam pemilihan fiitur. Sebelum masuk ke tahapan klasifikasi, pembobotan dilakukan pada setiap data yang ada untuk dapat menghitung kesamaan dan jarak antar data traning dengan data uji. Pada Penelitian ini untuk menghitung bobot setiap kata dalam dokumen digunakan skema pembobotan IDF. *Inverse document frequency* (IDF) adalah *inverse* atau kebalikan dari nilai DF. Hal ini dikarenakan *term* yang sering muncul di dokumen dianggap sebagai term umum,

sehingga tidak penting nilainya, sehingga ukuran kepentingan suatu term dari dokumen yang akan digunakan penciri yang memiliki nilai kecil dengan rentang yang tidak begitu jauh. Sebaliknya *term* yang jarang muncul pada dokumen perlu diperhatikan dalam pembobotan. Menurut Witten (1999) kata yang jarang atau paling sedikit muncul justru harus diperhatikan sebagai kata yang lebih penting dari pada kata yang paling sering muncul dalam dokumen. 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. Nilai IDF dapat diperoleh dari persamaan [**??**]

$$idf_{\mathrm{t}} = log(\frac{N}{df_{\mathrm{t}}}) \tag{1}$$

Pada persamaan **??** variabel N adalah banyaknya dokumen dan sedangkan df adalah banyaknya dokumen didalam koleksi yang mengandung *term* tertentu, sehingga dapat dikatakan bahwa IDF merupakan frekuensi *term* atau data yang jarang muncul dalam suatu dokumen.

Mutual information (MI)

Mutual information (MI) merupakan seleksi fitur yang melibatkan kontribusi term. MI mengukur seberapa besar kontribusi keberadaan suatu *term* t,dalam pembuatan keputusan klasifikasi yang benar. MI juga menunjukan keputusan klasifikasi secara benar atau salah melalui kontribusi keberadaan term tersebut. Nilai dari MI disimbolkan dengan notasi I, dimana [2]

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

Pada persamaan (2) U adalah variabel acak dengan nilai-nilai et = 1 , yang menunjukkan bahwa dokumen berisi term t, sedangkan untuk et = 0 menunjukkan bahwa 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 persamaan berikut : [3]

$$\frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_1 N_1} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_0 N_1} + \frac{N_{10}}{N} log_2 \frac{NN_{11}}{N_1 N_0} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_0 N_0} \tag{3}$$

Pada persamaan (3) nilai 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 $\in$ 0, 1). N adalah jumlah total dokumen atau N= N00 + N01 + N10 + N11.

Chi-Square (X^2)

Chi-square (x^2)adalah suatu ukuran yang menyatakan perbedaan antara frekuensi observasi (O) dan frekuensi harapan (E) untuk setiap term (i) yang dirumuskan dengan persamaan: [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 4.

Tabel 4. Nilai kritis untuk taraf nyata α

α	Nilai Kritis
0.050	3.840
0.010	6.630
0.005	7.880

Pada penelitian ini akan menggunakan IDF sebagai seleksi fiturnya. IDF dipilih karena metode ini efisien, mudah dan memiliki hasil yang akurat Robertson (2005).

Klasifikasi

Pada analisis sentimen klasifikasi digunakan untuk mengkategorikan setiap data yang ada ke kelas pencirinya. Salah satu tujuan dari klasifikasi teks atau dokumen adalah penggolongkan atau mengelompokkan suatu dokumen ke dalam suatu kategori tertentu (Manning et all. 2008). Klasifikasi juga bertujuan untuk memprediksi karakteristik dari suatu objek. Klasifikasi juga dapat digunakan untuk mendeteksi sentimen terhadap suatu isu. Data hasil indexing akan diklasifikasikan terhadap analisis sentiment. Pada penelitian ini terdiri 3 kelas sentimen yang digunakan, yaitu 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. Fungsi klasifikasi terbagi menjadi dua metode yaitu, berbasis vektor dan berbasis peluang (Manning et all. 2009). Secara garis besar pada pendekataan berbasis peluang, penentuan kelas pada sebuah dokumen atau data adalah dengan cara menghitung peluang keberadaan data tersebut dalam suatu kelas. Metode yang sering digunakan adalah metode *Naïve Bayes*. Sedangkan pada pendekatan berbasis vektor, penentuan kelas pada sebuah data dilakukan dengan cara menghitung jarak data tersebut ke centroid suatu kelas. Metode yang sering digunakan pada pendekatan ini adalah metode Rocchio dan k Nearest Neighbor (KNN).

Metode Naive Bayes

Model klasifikasi Multinomial dan Bernoulli merupakan metode klasifikasi berbasis peluang yang paling sering digunakan. Model klasifikasi ini banyak digunakan karena mudah diaplikasikan dan prosesnya sederhana (Manning et al. 2008). Pada model Multinomial Bernoulli setiap dokumen memiliki atribut yang menunjukkan ada atau tidaknya kata- kata atau *term* dalam dokumen tersebut, tetapi jumlah kemunculan term dalam dokumen tidak ikut diperhitungkan. Pada model Multinomial *Naïve Bayes*, jumlah kemunculan term pada dokumen ikut diperhitungkan, setiap dokumen diwakili oleh kemunculan *term* dari dokumen. Pada model ini, dapat diasumsikan jika kemunculan masing-masing term t bersifat independen antara satu term dengan yang lainnya. Dengan menggunakan nilai dari P(c/d) peluang suatu dokumen d di dalam kelas c dapat ditulis sebagai (Manning et al. 2009) [6]

$$P(C|d)\alpha P(c)\Xi_{1\leq k\leq nd}P(t_k|c) \tag{6}$$

dengan P(tk—c) adalah peluang dari suatu term tk muncul pada dokumen d yang diketahui memiliki kelas c. Pendugaan parameter P (tk—c) dihitung dengan cara: [7]

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

dengan Tct 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. 2009). *Term* tidak selalu muncul pada salah satu kelas saat dilakukan klasifikasi sehinggga nilai p (tk —c) yang dihasilkan adalah nol. Untuk mengatasi permasalahan tersebut, digunakan laplace smoothing, yaitu menambahkan frekuensi term sebanyak 1 sehingga perhitungan dari p (tk —c) menjadi (Manning et al. 2009) [8]

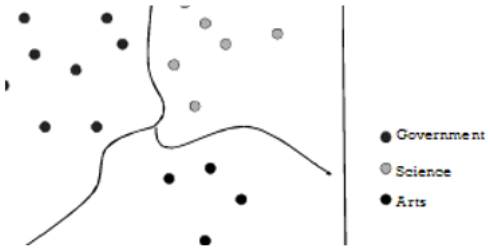
$$P(t_k|C) = \frac{T_{ct}}{\sum_{t'\in V} T_{ct'} + B} \tag{8}$$

Metode Rocchio

Klasifikasi ini merupakan salah satu metode pembelajaran *supervised document classification*. Metode *Rocchio relevance feedback* adalah strategi reformulasi query paling populer karena sering digunakan untuk membantu user pemula suatu *information retrieval systems* (Joachims 2013). Dalam siklus *relevance feedback*, kepada user disajikan hasil pencarian dokumen, setelah itu user dapat memeriksa dan menandai dokumen yang benar-benar relevan Klasifikasi yang digunakan pada penelitian ini adalah fungsi klasifikasi dengan basis vektor, yaitu metode klasifikasi Rocchio. Klasifikasi rocchio merepresentasikan data ke dalam sebuah vektor. Kedekatan kesamaan isi dihitung dari kedekatan sudut yang terbentuk antara bobot data training dan bobot data test menggunakan aturan sodinus. Untuk menghitung bobot setiap kata dalam dokumen digunakan skema pembobotan TFIDF (Term Frequency / Invers Document Frequency) karena komponen heuristic utama adalah dalam

klasifikasi rocchio yaitu skema pembobtan tfidf, untuk itu metode pembelajaran rocchio disebut juga dengan TFIDF Classifiers (Joachihms 1997). Pendekatan ini menggunakan perhitungan jarak atau kemiripan suatu data dengan pusat sebuah kelas. Metode ini membagi ruang vektor menjadi beberapa bagian berdasarkan centroid yang ada.

Pada penelitian ini setiap dokumen training direpresentasikan sebagai vektor. Setiap titik atau vektor dokumen traning akan diberikan label sesuai dengan kategori kelasnya. Contohnya saja seperti pada gambar 3



Gambar 3. Pelabelan pada kategori kelas

Teknik Rocchio menerapkan batas - batas tersebut dalam bentuk *centroid* untuk memberi batasan tersebut. Nilai *centroid* ini didapat dengan menghitung rata- rata jarak pada setiap data atau dokumen. *Centroid* dapat diperoleh dari persamaan berikut [9]

$$\mu(c) = \frac{1}{D_c} \sum_{d \in D_c} v(d) \tag{9}$$

Kemudian selanjutnya dari masing masing vektor dokumen akan dicari nilai centroidnya dari setiap kelas yang ada menggunakan persamaan (10). Persamaan diatas digunakan untuk menghitung centroid dari kelas C, dimana Dc merupakan kumpulan dari dokumen di dalam korpus c, sedangkan v(d)merupakan vektor dokumen yang telah dinormalisasi. Terdapat dua cara untuk menentukan kemiripan dua vektor space model yaitu dengan mengukur jarak atau mengukur kemiripan. Untuk menentukan jarak kedekatan data uji ke dalam suatu kelas adalah dengan menghitung jarak antara kedua vektor menggunakan persamaan *Euclidean*, yang didefinisikan sebagai berikut [10]

$$|x-y| = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{10}$$

Sedangkan untuk menghitung kemiripan antara dua vektor dokumen, didefinisikan dengan persamaan berikut [11]

$$sim(d_1, d_2) = \frac{v(d_1).v(d_2)}{|v(d_1)||v(d_2)|} \tag{11}$$

Pada penelitian ini pendekatan yang akan digunakan untuk mencari kemiripan antara dua vector adalah pendekatan *similarity*, sedangkan untuk pembobotannya digunakan IDF. Pendekatan ini digunakan karena pendekatan *similarity* mencari kemiripan berdasarkan kesamaan, bukan berdasarkan kedekatan. Sedangkan pada pendekatan jarak, pengukuran dilakukan berdasarkan kedekatan. Kedekatan belum tentu menunjukkan kesamaan antar *term*. Untuk itulah pendekatan menggunakan *similarity* yang akan digunakan, yang selanjutnya diharapkan resiko kesalahan dalam pengambilan dokumen akan lebih sedikit terjadi.

Evaluasi

Tahapan evaluasi adalah tahapan untuk mengetahui tingkat akurasi dan kinerja dari hasil klasifikasi menggunakan metode Rocchio. Kinerja klasifikasi dievaluasi dengan cara menghitung nilai akurasi, *recall*, *precision*, dan *F-measure* dengan bantuan tabel *confusion matrix*. *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*. Menurut Manning (2008) terdapat dua parameter yang umum digunakan untuk mengukur kinerja sebuah sistem temu kembali informasi, yaitu *precision* dan *recall* Pengukuran efektivitas dilakukan untuk mengevaluasi system IR. Perlu adanya suatu tolak ukur yang digunakan untuk mengukur kualitas hasil klasifikasi. Pengukuran selanjutnya akan mengunakan nilai precision, recall, akurasi, dan F1.

Precision

Precision adalah jumlah kelompok dokumen relevan dari total jumlah dokumen yang ditemukan oleh sistem. *Precision* direpresentasikan sebagai presentase dokumen yang di-retrieve yang benar-benar relevan. [12]

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

Recall

Recall adalah rasio jumlah dokumen relevan yang ditemukan kembali dengan total jumlah dokumen dalam kumpulan dokumen yang dianggap relevan. [13]

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

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

Accuracy

Setelah nilai precision dan recall didapatkan keakuratan hasil klasifikasi juga dinilai dari akurasinya, kemudian membandingkannya dengan kelas aktual dari data uji dan kelas hasil prediksi dengan menggunakan confusion matrix untuk untuk menghitung akurasi digunakan rumus seperti berikut dan mengacu tabel confusion matrix pada Table 5.

Tabel 5. Tabel Kontigensi

	Positif	Netral	Negatif
Positif	TP	FNt1	FNg1
Netral	FP1	TNt	FNg2
Negatif	FP2	FNt2	TNg

Penelitian ini membagi tiga sentimen yaitu positif , netral , dan negatif. Untuk itu table *confusion matrix* akan memiliki kolom dan baris yang direpresentasikan seperti Tabel 5. Pada Tabel 5 TP menunjukkan semua prediksi yang benar untuk data aktual positif, FP1 dan FP2 adalah jumlah prediksiyang salah untuk data aktual positif, TNt adalah jumlah prediksi yang benar untuk data aktual netral, FNt1 dan FNt2 adalah jumlah prediksi yang salah untuk data aktual netral, TNg menunjukkan jumlah prediksi yang benar untuk data aktual negatif, sedangkan FNg1 dan FNg1 menunjukkan jumlah prediksi yang salah untuk data aktual negatif. Dari Tabel 5 selanjutnya nilai akurasi dapat diperoleh dengan menggunakan persamaan [15]

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

Persamaan 15 akan menghasilkan nilai akurasi hasil klasifikasi yang berupa pembagian dari penjumlahaan nilai benar actual positif, negatif , dan netral dengan semua penjumlahan nilai prediksi yang didapat.

F-measure

Pengukuran selanjutnya adalah dengan menggunakan F-measure yang merupakan *weighted harmonic mean* dari *precision* dan *recall*.Dimana $\alpha \in [0,1]$ dan $\beta^2 \in [0, \Psi]$ [16]

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

Balanced F-measure menyamakan bobot dari precision dan recall, yang berarti membuat α = 1/2 atau β = 1. Ketika menggunakan β = 1, formula dapat disederhanakan sebagai berikut [17]

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

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".

Anwar Hridoy, Syed Akib *et al.* 2015. "Localized twitter opinion mining using sentiment analysis" dalam: *Decision Analytics* 2 (1), pp. 1–19. ISSN: 2193-8636. DOI: 10.1186/s40165-015-0016-4.

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.

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan Claypool Publishers. [Internet]. [Diunduh tanggal 11/08/2016]. Dapat diunduh dari: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

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

Mudinas, Andrius, Dell Zhang, dan Mark Levene. 2012. "Combining Lexicon and Learning Based Approaches for Concept-level Sentiment Analysis". *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. WISDOM '12. Beijing, China: ACM, 5:1–5:8. ISBN: 978-1-4503-1543-2. DOI: 10.1145/2346676.2346681.

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.

Pak, Alexander dan Patrick Paroubek. 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari (Conference Chair) *et al.* Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

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