ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

KLASIFIKASI POSTING TWITTER CUACA PROVINSI DIY MENGGUNAKAN ALGORITMA C4.5 UNTUK INFORMASI PADA WEB PARIWISATA

Yuli Astuti
Manajemen Informatika STMIK AMIKOM Yogyakarta
Jl Ring Road Utara, Condongcatur, Sleman, Yogyakarta 55281
yuli@amikom.ac.id

ABSTRAK

Setiap hari server Twitter menerima data tweet dengan jumlah yang sangat besar. Dengan demikian, perlu dilakukan pemanfaatan data dengan dikembangkan sistem untuk melakukan data mining dari tumpukan data tersebut yang akan digunakan untuk kepentingan tertentu, salah satunya adalah untuk visualisasi kondisi daerah tertentu berdasarkan cuaca. Teknik klasifikasi akan diterapkan untuk mengklasifikasikan data tweet yang menginformasikan cuaca di kota yogyakarta. Sebelum dilakukan klasifikasi, data tweet melalui preprocessing dan pembobotan term frequency dan tf-idf. Kemudian dari pembobotan ini menghasilkan data training dan data testing yang akan dilakukan klasifikasi menggunakan pohon keputusan dengan algoritma C4.5 untuk memprediksi cuaca di daerah kota yogyakarta untuk kemudian diinformasikan pada web pariwisata dengan pemanfaatan web service. Hasilpengujian dengan perangkat lunak Rapid Miner 5.3 diperoleh nilai akurasi terkecil 71% dengan sampel sebanyak 100 dan nilai akurasi tertinggi 95,58% dengansampel 15106 dengan algoritma C4.5.

Kata Kunci : Twitter, cuaca, klasifikasi, algoritma C4.5

ABSTRACT

Every day the Twitter server receives data tweet with a very large number. Thus, it is necessary to use the data to develop a system to perform data mining on a pile of data that will be used for specific purposes, one of which is for the visualization of certain areas based on the weather conditions.Classification techniques will be applied to classify the data tweet informing the weather in the city of Yogyakarta. Before the classification, the data tweets through preprocessing and weighting term frequency and tf-idf. Then from this weighting generates training data and data testing to be performed classification using C4.5 decision tree algorithm to predict the weather in the city of Yogyakarta to then be informed on the tourism web with the use of the web service.Hasilpengujian with Rapid Miner 5.3 software obtained value of the smallest 71% accuracy with a sample of 100 and the highest value of 95.58% accuracy dengansampel 15106 with C4.5 algorithm.

Keywords*: Twitter, weather, classification, algorithm C4.5*

1. Pendahuluan

1.1 Latar Belakang

Saat ini dunia internet sedang berada pada pada fase *user generated content,* yang berarti seluruh konten yang berada di internet adalah buatan pengguna secara umum. Dengan demikian, internet seperti gudang super besar yang diisi oleh pengguna dan pengguna juga dapat menggunakan isi gudang tersebut. Salah satu aplikasi internet yang mendukung *user generated content* adalah *microblogging*. Saat ini, *microblogging* menjadi populer sebagai alat komunikasi antara pengguna internet. Jutaan pesan setiap hari dikirim oleh pengguna melalui situs microblogging ini. Pengguna biasanya menulis tentang hal-hal yang terjadi dalam kehidupan sehari-harinya, berdiskusi tentang berbagai topik yang sedang banyak diperbincangkan dan lain-lain. Salah satu aplikasi internet yang memungkinkan pengguna dapat berbagi konten secara bebas adalah aplikasi Twitter. Twitter adalah salah satu situs *microblogging* paling populer yang dapat membagikan update status atau pesan yang

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

tidak lebih dari 140 karakter kepada pengguna dalam satu jaringan dalam minat yang sama dikenal dengan istilah *follower*, atau pengguna yang mengikuti karena memiliki kesamaan minat,ketertarikan pada isu tertentu.
[1] menyatakan bahwa pada pertengahan tahun 2010 Twitter memiliki pengguna lebih dari 106 juta pengguna diseluruh dunia dan terus meningkat setiap harinya sebanyak 300.000 pengguna dan Twitter setiap harinya mendapatkan lebih dari 3 juta request. Dari angka tersebut Indonesia menjadi negara yang menduduki peringkat 8 dalam mengakses situs Twitter. Twitter menerima tweet dari pengguna sebanyak 55 juta pesan setiap harinya.
Berdasarkan data tersebut, Twitter memiliki sumber data yang besar. Data dalam hal ini adalah tweet dari pengguna yang berjumlah sangat banyak. Hal ini merupakan sebuah sumber daya yang bisa kita manfaatkan untuk kepentingantertentu misalkan untuk mengetahui cuaca di suatu kota berdasarkan tweet yang dikirim oleh pengguna yang berisi informasi cuaca.

Saat ini pengguna Twitter, khususnya yang berada di kawasan kota yogyakarta sering menginformasikan melalui Twitter mengenai informasi yang berhubungan dengan kota yogyakarta. Informasi tersebut antara lain, informasi kegiatan (event), iklan (advertising), lowongan kerja di kota yogyakarta dan sekitarnya, informasi cuaca dan lain-lain. Kumpulan informasi di Twitter yang berhubungan dengan kota yogyakarta biasanya ditandai dengan hastag #eventyk untuk informasi kegiatan, #infoyk untuk informasi umum, #cuacayk untuk informasi cuaca, dan #diskonyk untuk informasi diskon.

Kota yogykarta merupakan kota yang sering mengalami cuaca tidak menentu disetiap daerahnya, misalnya daerah condong catur cerah namun daerah jalan kaliurang hujan. Terkait dengan ini, sebagian besar pengguna Twitter di kota Yogyakarta sering menginformasikan cuaca disekitarnya melalui media Twitter dengan hastag #cuacayk. Dengan demikian, apabila data Twitter tersebut dikumpulkan dan kemudian dilakukan knowledge discovery terhadap data tersebut, diharapkan dapat memberikan informasi yang akurat mengenai cuaca di kota yogyakarta yang akan dijadikan sebagai bahan pertimbangan dalam pengambilan keputusan oleh wisatawan. Hal ini, seperti yang diungkapkan [2] bahwa sekarang ini melalui data twitter, dapat menentukan gaya hidup, sikap dan perilaku seseorang serta keinginan-keinginannya.

Data mining merupakan sebuah proses dari knowledge discovery (penemuan pengetahuan) dari data yang sangat besar [3]. Sementara itu [4] berpendapat bahwa data mining adalah proses secara otomatis untuk menemukan informasi yang berharga dari repositori data yang sangat besar. Dengan demikian, dari tumpukan data tersebut akan didapat beragam informasi yang berharga dan penting yang sebelumnya tidak diketahui.

Salah satu cabang data mining yang mengkhususkan pada penggalian informasi dari data yang berupa data teks adalah text mining. Text mining merupakan bidang data mining yang bertujuan untuk mengumpulkan informasi yang berguna dari data teks dalam bahasa alami atau proses analisis data teks kemudian mengekstrak informasi yang berguna untuk tujuan tertentu [5]. Dengan demikian, text mining merupakan teknik yang cocok untuk melakukan ekstraksi informasi dari data tweet yang banyak tersebut sehingga dapat menghasilkan informasi yang akurat yang akan digunakan dalam proses pengambilan keputusan.

Ada banyak teknik yang bisa dilakukan untuk melakukan klasifikasi data diantaranya adalah decision tree, bayesian classifiers, bayesian belief network, *Nearest Neighbor* dan rule based classifiers [3]. Berdasarkan penelitian-penelitian sebelumnya, teknik-teknik ini mempunyai kelemahan dan kekurangan. Berdasarkan penelitian yang dilakukan oleh [6], kinerja algoritma decision tree lebih baik jika dibandingkan dengan Multiple Discriminant Analysis (MDA) dalam memprediksi kebangkrutan perusahaan. Algoritma decision tree juga merupakan algoritma paling popular dalam teknik klasifikasi.

Dalam Penelitian ini diusulkan untuk memanfaatkan data twitter untuk memprediksi cuaca kota Yogyakarta secara *realtime* untuk pemanfaatannya pada web pariwisata. Hasil penelitian ini akan menjadi sebuah model bagi daerah lain yang akan dan sedang mengembangkan sistem informasi prakiraan cuaca secara akurat dan *realtime.*

1.2 Rumusan Masalah

Dari latar belakang yang dijelaskan di atas, maka rumusan masalah yang diangkat

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

adalah Bagaimana mengklasifikasikan data *tweet* yang mengandung *hastag* #cuacayk yang merupakan informasi cuaca di kota Yogyakarta?

1.3 Batasan Masalah

Proses klasifikasi data twitter cuaca Yogyakarta, dengan batasan masalah sebagai berikut :

1. Data tweet yang digunakan adalah tweet yang diposting dari tanggal 1 Juni 2015 sampai tanggal 24 September 2015.
2. Data tweet yang di download adalah data tweet yang mengandung hastag #cuacayk
3. Tidak ada keterkaitan satu kata dengan kata yang lain dalam sebuah tweet.
4. Sesuai dengan konsep user generated content sistem ini menganggappengguna Twitter yang menginformasikan cuaca sebagai ageninformasi yang kemudian informasi ini dijadikan sumber data untuk sistem ini.
5. Cuaca yang akan dianalisis dalam sistem ini adalah cuaca dibeberapa daerah kabupaten sleman, kota Yogyakarta dan Bantul

1.4 Metode Penelitian

Penelitian ini dilakukan dengan tahap-tahap sebagai berikut :

1. Pengumpulan data

Melakukan pengumpulan data tweet secara realtime yang mengandunghastag #cuacayk. Pengumpulan data ini dilakukan denganmemanfaatkan API search yang disediakan oleh Twitter. Kemudian hasil daripencarian ini disimpan di database. Data tweet yang dikumpulkan antara lainusername, isi tweet, URI tweet, tanggal dan waktu tweet dan URI profil pengguna.

2. Preprocessing

Preprocessing dilakukan dengan langkah-langkah sebagai berikut :

a) Setelah data terkumpul, kemudian dilakukan preprocessing data sehingga data menjadi lebih mudah untuk dilakukan proses selanjutnya. Pada tahapan ini dilakukan :
   1. Penghapusan kata tertentu yang tidak dipakai dalam proses klasifikasi.
   2. Konversi menjadi huruf kecil.
   3. Menghapus url (http://bit.ly/mHibqV).
   4. Melakukan perbaikan data apabila ada kata yang diperlukan tetapi data tidak sesuai, misalkan ada kata “panaaaasssss” kemudian di ubah menjadi “panas”.
   5. Menghapus mention (@xxx).
   6. Menghapus karakter selain a-z.
   7. Mengganti sinonim (yg=yang, gerah=panas, gerimis=hujan,sumuk=panas, terik=panas).
   8. Menghilangkan stopword.
   9. Menghapus kata dengan satu karakter.
b) Menghitung bobot tiap *term* dari data tweet yang terkumpul dengan menggunakan teknik *tf-idf*. Ditunjukkan pada Persamaan 1.

$$tfidf\,(d, w) = tf\,(d, w) \times \log N/dfw \quad \ldots.(1)$$

Dimana : tf(d,w) adalah frekuensi kemunculan term w pada dokumen d, n adalah jumlah keseluruhan dokumen dan dfw adalah jumlah dokumen yang mengandung term w.

3. Klasifikasi

Kalsifikasi dilakukan dengan langkah-langkah sebagai berikut :

a) Pada proses *learned* model digambarkan dalam bentuk classification rule atau formula matematika yang biasa dikenal dengan algoritma. Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Menurut [7] langkah-langkahnya sebagai berikut :
   1. Pilih variabel sebagai akar
   2. Buat cabang untuk tiap-tiap nilai
   3. Bagi kasus dalam cabang
   4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Untuk memilih variabel sebgai akar, didasrkan pada nilai *gain* tertinggi dari variabel-variabel yang ada. Untuk menghitung *gain* digunakan rumus pada Persamaan 2.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} x Entropy\,(S_i) \ldots..(2)$$

Keterangan :

S : Himpunan kasus
A : Variabel
n : jumlah partisi variabel A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

sementara itu untuk menghitung nilai entropi dapat dilihat pada Persamaan 3.

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

$$Entropy(S,A) = \sum_{i=1}^{n} -pixlog_2pi \quad \ldots\ldots (3)$$

Keterangan :
S : Himpunan Kasus
A : Fitur
n : jumlah partisi S
p_i : proporsi dari S_i terhadap S

b) Kelas dalam klasifikasi ini adalah kelas panas untuk tweet yang mengandung informasi cuaca panas (cerah) dan kelas hujan untuk tweet yang mengandung informasi cuaca hujan di suatu jalan atau tempat tertentu
c) Memisahkan data untuk digunakan sebagai data training dan data testing.
d) Dilakukan pengujian akurasi menggunakan metode *K-Fold Cross Validation* pada model klasifikasi algoritma C4.5 dengan menggunakan Rapid Miner 5.3.

4. Pengujian hasil akurasi

Analisa hasil akurasi model klasifikasialgoritma C4.5 denganmenggunakan metode 10 fold cross validation. Cara kerja metode ini yaitu Pada pengujian ini, sebanyak 10%dari jumlah posting secara bergantian dijadikan data uji sebanyak 10 kali terhadap90% posting lainnya yang dijadikan data training. Nilai akurasi diperoleh dariarata-rata nilai akurasi dari 10 kali pengujian tersebut. Dengan demikian, setiapposting tweet akan menjadi data training dan data testing secara bergantian. Halini bertujuan untukmeminimalkan nilai akurasi yang dihasilkan oleh faktor kebetulan. Untuk menghitung nilai akurasinya digunakan persamaan (4).

$$\text{akurasi} = \frac{jumlah\ klasifikasi\ benar}{jumlah\ data\ uji}\ x\ 100\% \quad (4)$$

2. Pembahasan

2.1 Persiapan Data

Data yang digunakan dalam penelitian ini adalah data tweet yang diambilsecara realtime. Setiap satu menit sekali sistem memberikan request ke serverTwitter untuk mengambil dokumen XML yang berisi tweet yang kemudiandisimpan di database. Sesuai dengan batasan masalah, tweet yang dikumpulkan
adalah tweet yang mengandung hastag #cuacayk.Pengambilan data tweet yang akan digunakan dalam penelitian ini dimulai sejaktanggal 1 Juni 2015 sampai tanggal 24 September 2015yang menghasilkan data
tweet sebanyak 15401 record.
Adapun struktur data tweet yang dikumpulan dalam sistem adalah sebagai berikut :

1. User adalah data pengguna twitter yang memposting tweet.
2. Tweet adalah data konten tweet yang berisi informasi cuaca di kota Yogyakarta.
3. uriTweet adalah link yang menuju halaman tweet.
4. Date adalah tanggal tweet tersebut diposting.
5. Time adalah waktu tweet tersebut diposting.
6. Uri adalah link yang menuju halaman profil pengguna Twitter yang memposting tweet.

Dalam sistem ini, data diklasifikasi menjadi dua kelas/label yaitu kelas ‘cerah’ untuk tweet yang menginformasikan keadaan cuaca yang cerah atau panas di suatu jalan atau tempat, kelas ‘hujan’ untuk tweet yang menginformasikan keadaan cuaca yang hujan di suatu jalan atau tempat.

2.2Analisis dan Pengembangan Sistem

Berdasarkan permasalahan yang telah dirumuskan danpemahaman terhadap teori-teori dan konsep-konsep sebagai dasar pemikiran, maka pada bagian ini merupakan konsep penyelesaian masalah dengan melakukan analisis dan perancangan sistem. Sistem baru yang dirancang merupakan sistem yang dapat melakukan proses pengambilan data tweet dari API search Twitter secara otomatis dan realtime. Kemudian dari tumpukan data teks tweet tersebut dilakukan pembobotan setiapkata (term) yang kemudian hasil pembobotan tersebut diolah dengan algoritma C4.5 untuk menentukan model yang akan digunakan untuk menentukan prediksi kelas pada data tweet yang baru. Inti dari permasalahan yangdiangkat adalah bagaimana melakukan ekstraksi data tweet untuk digunakan untuk training dan testing model klasifikasi algoritma C4.5 yang digunakan untuk prediksi kelas untuk data tweet yang baru. Kemudian hasil klasifikasi tersebut divisualisasikan ke Web Pariwisata yang

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

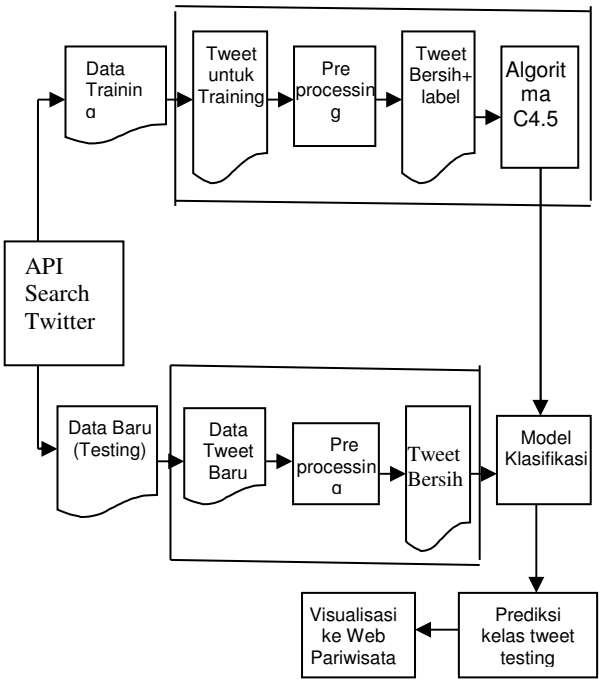
terkoneksi dengan web service dengan menggunakan Google Map.
Secara umum sistem ini terdiri dari empat bagian diantaranya adalah download (pengambilan data), preprocessing tweet, klasifikasi tweet dan visualisasi hasil klasifikasi.
Adapun desain arsitektur sistem dapat terlihat pada Gambar 1.

Gambar 1 Rancangan Arsitektur Sistem

2.2.1Perancangan Sistem

Penelitian ini dimulai dengan mengambil data tweet yang tersimpan diserver Twitter dengan menggunakan bantuan API Search Twitter. Pada proses pengambilan data ini juga dilakukan pe-label-an data. Kemudian data tersebutdiproses untuk tahap persiapan (preprocessing) agar data siap digunakan untukproses klasifikasi. Hal ini dilakukan karena tidak semua data tweet tersebut dapatdigunakan. Pada data preprocessing ini dilakukan pembersihan data tweet daritweet dan kata-kata yang tidak digunakan, penggantian kata-kata tertentu dengandaftar sinonim yang sudah ada, melakukan perbaikan data penting tetapi tidaksesuai (tidak lengkap). Kemudian hasil pembersihan data ini disimpan di tempatyang berbeda dari “data kotor”-nya.

Kemudian data tweet bersih yang sudah diberi label/kelas atau data trainingdiolah oleh algoritma C4.5 untuk menghasilkan model pohon keputusan. Modelkeputusan tersebut terbagi menjadi dua bagian, yaitu :probabilitas kelas danprobabilitas semua kata (term) pada suatu kelas yang dalam sistem ini disimpan didatabase. Setelah itu, model probabilitas ini digunakan untuk memprediksi datatweet yang baru yang sudah dibersihkan pada tahap preprocessing. Kemudianhasil prediksi data tweet yang baru di visualisasikan ke Web Pariwisata yang terkoneksi dengan Web Service dengan menggunakan GoogleMap. Gambaran umum sistem dapat dilihat pada Gambar 2.



Gambar 2 Gambaran Umum Sistem

2.2.2 PerancanganDownload Tweet

Proses download tweet diawali denganmengakses API Search Twitterdengan query yang diberikan adalah hastag cuacayk. Hasilquery yang diberikan API Search Twitter adalah file XML yang kemudian isi fileXML tersebut di parsing ke variabel tertentu yang digunakan untuk prosespenyimpanan data.
Data yang dihasilkan oleh XML tersebut tidaksepenuhnya langsungdisimpan, melainkan di cek terlebih dahulu apakah tweet yang akan dimasukkanke database tersebut sudah ada sebelumnya dengan mencek kesamaan antara userdan tweet dengan yang sudah tersimpan di database. Apabila sudah ada, makadata tidak akan disimpan tetapi apabila data belum ada maka selanjutnya datadisimpan di database.

2.2.3 Perancangan Preprocessing

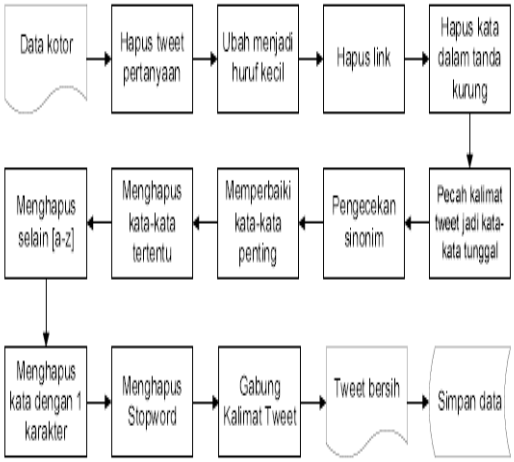
Secara umum preprocessing ini melewati beberapa tahap antara lain :

1. Pembersihan tweet pertanyaan. Dalam sistem ini tweet pertanyaan diasumsikan dengan tweet tersebut mengandung karakter “?”. Tweet pertanyaan dihapus diasumsikan bahwa tweet pertanyaan tidak mengandung informasi cuaca.

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

2. Pembersihan tweet kata per kata.
3. Penyimpanan data tweet bersih ke database.

Pada tahap preprocessing dilakukan pembersihan semua kata yang terkandung pada satu buah tweet. Proses yang dilewati pada preprocessing inidapat dilihat pada gambar 3. Adapun tahap pembersihan kata dalam sistem iniantara lain adalah mengkonversi menjadi huruf kecil untuk menstandarkan data,melakukan perbaikan kata-kata yang diperlukan dalam dalam klasifikasi,menghapus url/link yang merupakan data yang tidak diperlukan, menghapus mention (@userTwitter) karena data ini tidak diperlukan, menghapus karakterselain a-z dan 0-9, mengganti sinonim, menghilangkan stopword dan menghapus kata yang hanya satu karakter.



Gambar 3 Flow Chart Preprocessing

2.2.4 Perancangan Bobot Term Teknik TF-IDF

Sebelum tweet diklasifikasi menggunakan algoritma C4.5,Pada bagian ini akan dijelaskan terlebih dahulu mengenai pembentukan bobot tiap term dari data tweet yang terkumpuldengan menggunakan teknik tf-idf. Misalkan, terdapat 4 buah tweet yang sudah bersih yang terdapat pada Gambar 4.

1. imogiri hujan angin kencang
2. panas malioboro
3. concat panas tenanan
4. panas imogiri

Gambar 4 Contoh Data Tweet

Dari data tweet pada Gambar 4 tersebut kemudian dipilih kata-kata unik dari semua tweet tersebut sehingga menjadi matrik yang tertuang pada Tabel 1.

Tabel 1 Matriks Term Data Tweet Gambar 4

Tweet	Imogiri	hujan	Angin	kencang	Panas	Malioboro	concat	tenanan
1	1	1	1	1				
2					1	1		
3					1		1	1
4	1				1			

Dari matrik tersebut kemudian dihitung dengan menggunakan persamaan (1) sehingga setiap kemunculan kata dalam tweet tersebut diubah menjadi nilaitf-idf. Berikut adalah contoh perhitungan tf-idf untuk kata "imogiri" pada tweet1 :

Jumlah dokumen = 4, tf (imogiri pada tweet1) = 1, df (imogiri) = 2

$tf\text{-}idf$ (Tweet 1,"imogiri") $= 1 \times log\frac{4}{2} = 0{,}301$

Apabila proses perhitungan tersebut dilakukan untuk semua dokumen dan semuakata maka akan dihasilkan matrik perhitungan tf-idf yang terdapat pada tabel 2.

Tabel 2 Matriks *tf-idf* Data Tweet Gambar 4

Tweet	Imogiri	hujan	Angin	kencang	Panas	Malioboro	concat	tenanan
1	0,301	0,602	0,602	0,602	0	0	0	0
2	0	0	0	0	0,124	1	0	0
3	0	0	0	0	0,124	0	0,602	0,602
4	0,301	0	0	0	0,124	0	0	0

2.2.5. Perancangan Klasifikasi Algoritma C4.5

Setelah menemukan hasil perhitungan *tf-idf* tweet, selanjutnyahasil perhitungan *tf-idf* ini akan diolah dengan menggunakanalgoritma C4.5 sehingga menghasilkan model pohon keputusan. Misalkan tweet 1 merupakan tweet dengan kelas hujan, sedangkan tweet 2,tweet 3 dan tweet 4 merupakan tweet dengan kelas panas. Sehingga matrik tf-idf tersebut akan berubah menjadi seperti pada Tabel 3.

Tabel 3 Matriks *tf-idf*Berdasarkan Kelas

Tweet	Imogiri	hujan	Angin	kencang	Panas	Malioboro	concat	tenanan	Kelas
1	0,301	0,602	0,602	0,602	0	0	0	0	Hujan
2	0	0	0	0	0,124	1	0	0	Panas
3	0	0	0	0	0,124	0	0,602	0,602	Panas
4	0,301	0	0	0	0,124	0	0	0	Panas

Node		Jumlah kasus (S)	Panas (S_1)	Hujan (S_2)	Entropy	Gain
1	Total	14	10	4	0,8631	
Daerah	imogiri	4	4	0		0,2585
	malioboro	5	4	1	0,7219	
	concat	5	2	3	0,9709	

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

Baris **Total** kolom Entropy pada Tabel 4
Entropy (Total) =
$$\left(-\frac{4}{14}\times log_2\left(\frac{4}{14}\right)\right)+(-\frac{10}{14}\times log_2\left(\frac{10}{14}\right))$$
Entropy (Total) = 0,8631

Perhitungan tersebut juga dilakukan untuk mencari entropy pada tiap daerah.Sementara itu, nilai Gain pada baris imogiri dihitung dengan menggunakan persamaan 2, sebagai berikut :

Gain (Total,Daerah) $= Entropy(Total) -$
$$\sum_{i=1}^{n}\frac{|Daerah_i|}{|Total|}\ x\ Entropy\ (Daerah_i)$$
Gain(Total,Daerah) =
$0{,}8631 - ((\frac{4}{14}$x 0)+($\frac{5}{14}x0{,}7219$) +
($\frac{5}{14}x0{,}9709$))
Gain(Total,Daerah) = 0,2585

2.3 HasilPengujianAkurasiKlasifikasi

Sesuai dengan metode penelitian yang disampaikan, maka pengujian yang akan dilakukan pada sistem ini adalah dengan menggunakan metode k fold cross validation. Nilai k dalam penelitian ini adalah 10. Dengan demikian, akan dilakukan 10 kali pengujian akurasi klasifikasi dari data bersih yang dijadikan sampel. Adapun aspek yang diuji adalah akurasi model terhadap data penelitian yang dikumpulkan dengan berbagai porsi jumlah data penelitian yang dijadikan sampel. Secara lebih rinci, berikut merupakan mekanisme pengujian sistem yang dilakukan pada sistem ini :

1. Membagi sampel menjadi 10 bagian yang sama rata.
2. Sebanyak 10% dari jumlah sampel tersebut secara bergantian dijadikan sebagai data testing dan 90% lainnya dijadikan sebagai data training.
3. Dari 10% yang dijadikan data testing tersebut kemudian dibandingkan hasil klasifikasi oleh sistem dengan kelas yang sudah ditentukan sebelumnya.
4. Dihitung nilai akurasinya menggunakan persamaan (4) untuk masing k.
5. Dihitung nilai rata-rata seluruh nilai akurasi untuk semua k untuk memperolehnilai akurasi keseluruhan.

Sistem yang sudah dibangun diuji dengan menggunakan data penelitianyang sudah terkumpul sebelumnya. Pada proses pengujian ini, dilakukanpengklasifikasian data dengan berbagai porsi jumlah sampel yang dijadikan sebagai data input dari sistem ini. Adapun jumlah porsi sampel adalah 100, 1000,5000, 10000 dan 15106 sampel data yang digunakan.

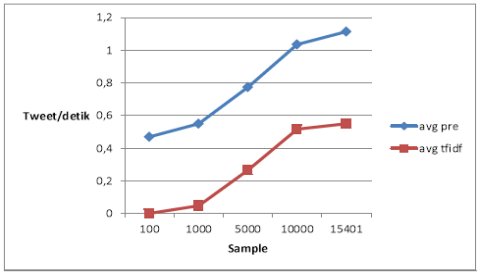
2.3.1 Pengujian *Running Time Preprocessing*

Pada pengujian ini dilakukan perhitungan running time preprocessing dari berbagai porsi jumlah sampel yang telah ditentukan sebelumnya. Tabel 5 tertuang hasil pengujian running time preprocessing.

Tabel 5 Running Time Preprocessing

No	Jumlah Sampel	Preprocessing (detik)	Rata-rata Preprocessing	*Tf-Idf* (detik)	Rata-rata *Tf-Idf*
1	100	41	0,410 tweet/detik	0	0,000 tweet/detik
2	1000	530	0,530 tweet/detik	42	0,042 tweet/detik
3	5000	3810	0,762 tweet/detik	1319	0,263 tweet/detik
4	10000	10337	1,033 tweet/detik	5110	0,511 tweet/detik
5	15106	16585	1,097 tweet/detik	8385	0,555 tweet/detik

Dari Table 5 dapat dilihat bahwa semakin besar jumlah sampel yang digunakan, semakin besar juga rata-rata tweet yang dapat diolah pada proses preprocessing dan tf-idf dalam satu detik. Gambar 5 merupakan grafik perbandingan running time dengan berbagai porsijumlah sampel.



Gambar 5 grafik perbandingan running time

Dengan memperhatikan gambar 5 dapat dilihat bahwa semakin besarjumlah sampel yang diolah, maka semakin besar pula rata-rata tweet yang melalui preprocessing dalam satu

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

detik. Sementara itu, tidak terjadi kenaikan yangsignifikan pada nilai rata-rata tweet yang yang dapat diproses tf-idf dalam satu detik.

2.3.2 Pengujian Akurasi Model

Pada pengujian ini dilakukan pengujian akurasi model pohon keputusan terhadap data bersih yang terbentuk dengan berbagai porsi jumlah. Seperti pada preprocessing, porsi jumlah data pada pengujian akurasi model pohon keputusan terdiridari 100, 1000, 5000, 10000 dan 15106 sampel data yang digunakan yangdihasilkan dari preprocessing. Berikut merupakan rincian hasil pengujian denganberbagai porsi jumlah data yang ditunjukkan pada Tabel 6.

Tabel 6. Rata-rata Nilai Akurasi Model

Jumlah Sampel	Rata-rata Akurasi
100	71 %
1000	83,40 %
5000	88,52 %
10000	91,58 %
15106	*95,58 %*

Tabel 6. merupakan tabel rata-rata nilai akurasi model untuk masing-masing pengujian dengan berbagai porsi data. Dari tabel tersebut dapat diketahui bahwa nilai akurasi tertinggi terdapat pada pengujian dengan menggunakan sampel data sebanyak 15106.

3. KESIMPULAN

Dari penelitian yang telah dilakukan terdapat beberapa kesimpulan yaitu :

1. Metode algoritma C4.5 mengasumsikan bahwa didalam setiap kelas dikelompokkan dengan pasti dan model pohon keputusan lebih mengarah pada perhitungan probabilitas tiap-tiap record,dalah hal ini kata. setiap kata unik bebas bersyarat satu sama lain. Asumsi ini digunakan agar proses training untuk membuat model klasifikasi dapat dijalankan.
2. Dari hasil pengujian preprocessing yang dilakukan, dapat disimpulkan bahwa semakin banyak jumlah data yang di-preprocessing maka semakin lama waktu yang diperlukan untuk preprocessing. Hal ini terlihat dari semakin besar rata-rata tweet yang dapat di-preprocessing dalam satu detik.
3. Dari hasil pengujian akurasi model dari sistem yang dikembangkan, menghasilkan nilai akurasi terkecil sebesar 71% pada proses pengujian dengan menggunakan sampel sebanyak 100 dan menghasilkan nilai akurasi tertinggi sebesar 95,58% pada proses pengujian dengan menggunakan sampel sebanyak 15106.

SARAN

penelitian ini masih memiliki keterbatasan yang dapat dijadikan acuan untuk pengembangan dimasa yang akan datang, sehingga dapat disarankan beberapa hal sebagai berikut :

1. Perlunya parameter untuk mengidentifikasi twitter yang di-input-kan adalah user sebenarnya (manusia) atau mesin
2. Perlu dilakukan analisis hubungan antar kata pada setiap klasifikasi
3. Perlu dilakukan perbandingan dengan teknik lain pada proses *preprocessing*
4. Perlu dilakukan validasi apakah hasil klasifikasi ini sesuai dengan keadaandi temapat sebenarnya.

DAFTAR PUSTAKA

Hepburn, A, 2010, Infographic: Twitter Statistics, Facts & Figures, http://www.digitalbuzzblog.com/infographic-twitterstatistics-factsfigures/Di akses tanggal 9 Mei 2011.

McCormick, H. T., Lee, H., Cesare, N., Shojaie, A., 2013, UsingTwitter for Demographic and Social Science Research:Tools for Data Collection, International Joint Conference for Statistics and the Social Sciences University of Washington,Paper no.127.

Han, J., & Kamber, M., *Data Mining Concept and Technique*, San Fransisco: Morgan Kaufman Publisher, 2006

Tan, P. N., Steinbach, M., & Kumar, V., 2006, Introduction toData Mining, Pearson Education, Boston.

Witten, I. H., 2005, Text mining, Dalam Practical Handbook ofInternetComputing Florida: Chapman & Hall/CRC Press,Boca Raton, 14-22.

Lonneke Mous. Predicting bankruptcy with discriminant analysis and decision tree using financialratios, 2005.Faculty of Economics at Erasmus University Rotterdam.

ISSN : 2407 – 1846
e-ISSN : 2460 – 8416
Website : jurnal.ftumj.ac.id/index.php/semnastek

Kusrini & Luthfi, T. E., 2009, Algoritma Data Mining, Andi Offset,Yogyakarta.