Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

Analisa Algoritma C4.5 Dalam Menentukan Faktor Yang Mempengaruhi Munculnya Professional Blogger

Robby [a] *, Lavinia* [b]*, Darsono Nababan*[c]

[a,b] *Mahasswa Univesitas Pelita H arapan Medan Campus, Jl. Imam Bonjol No.6, Medan Petisah, Kota Medan, Indonesia*
[b] *Dosen Univesitas Pelita H arapan Medan Campus, Jl. Imam Bonjol No.6, Medan Petisah, Kota Medan, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 23 November 2018
Revisi Akhir: 19 Maret 2019
Diterbitkan *Online*: 22 Maret 2019

KATA KUNCI

Blog
Blogger Profesional
Algoritma C4.5.

KORESPONDENSI

No HP: 085763481568
E-mail: robbytanrotan@gmail.com

A B S T R A C T

The growing of information technology and technological advance has made a lot of changes in media. Blogs are the recent emerging mediaand data information.They are provided for ideas and exchanging opinions. We apply C4.5 algorithm by applying decision tree with 0.25 confidence to determine the ratio from attributes that influence the professional blogger.The dataset uses blogger data from UCI Machine Learning Repository. C4.5 algorithm calculating entropy and information gain to produce a decision tree. Professional Blogger which attributes as destination attributes, while the degree, political caprice, topics, local media turnover, and local political social spaces as a source attribute to obtain the root node and other nodes.

1. PENDAHULUAN

Dengan semakin berkembangnya teknologi informasi dan kemajuan teknologi telah membuat terciptanya beberapa media sosial. Blog sebagai salah satu media dan data informasi yang sedang berkembang. Kemudahan dalam mengakses media melalui internet telah memberikan kesempatan bagi para praktisi media untuk menulis pendapat mereka.Pada era ini, untuk memiliki sebuah blog tidak memerlukan keahlian atau keterampilan tertentu. Telah terdapat banyak software yang dapat digunakan untuk membuat blog dan umumnya dapat diakses tanpa menggunakan biaya. Dikarenakan kemudahan ini, blog telah memiliki per-kembangan yang cukup signifikan di seluruh penjuru dunia.

Blog dapat berisikan informasi maupun pendapat mengenai hal yang sedang terjadi saat ini. Walaupun Blog dapat memiliki tema, dan gaya penulisan yang berbeda, tetapi secara umum memiliki aliran atau *genre* yang jelas. Dengan mengetahui kecenderungan blogger dalam menulisdapat menyediakan data yang penting bagi perencanaan dan juga pemerintahan.
Penelitian ini menggunakan metode decision tree yang termasuk dalam algoritma klasifikasi data mining. Algoritma klasifikasi data mining adalah suatu metode pembelajaran untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. beberapa algoritma klasifikasi yang sering digunakan antara lain adalah naïve bayes, decision tree, neural network, k-nn, random forest dan lain sebagainya. Dengan menggunakan data public dari uci repository yang memiliki 6 atribut dan 100 record. dimana data yang digunakan adalah data blogger, yaitu untuk mengklasifikasi blogger profesional.

Penelitian yang pertama yang dijadikan sebagai referensi atau acuan dalam penelitian ini yaitu penelitian yang dilakukan oleh Soleimanian et al menjelaskan bahwa Blog adalah media baru yang muncul yang bergantung pada teknologi informasi dan kemajuan teknologi. Simulasi dari informasi yang diperoleh dari 100 pengguna dan blogger di Kohkiloye dan Boyer Ahmad Province dan menggunakan alat bantu Weka 3.6 dan algoritma C4.5 dengan menerapkan pohon keputusan dengan lebih dari 82% presisi untuk mengantisipasi kecenderungan pengguna di masa depan untuk blog dan menggunakan di area strategis.
Tujuan dari penelitian ini adalah menentukan urutan dan pentingnnya faktor-faktor dari ke-6 atribut tersebut yang

mempengaruhi seseorang agar dapat menjadi professional blogger.

2. TINJAUAN PUSTAKA

2.1. Blog

Blog merupakan sebuah media sosial yang baru-baru ini berada di ruang cyber adalah salah satu layanan internet dan web yang menyediakan komponen perangkat lunak gratis bagi pengguna untuk membiarkan mereka berpartisipasi sebagai anggota jaringan dan komunitas virtual, sehingga menyebabkan hubungan dinamis dan interaktif yang tidak terbatas, dan opini tentang masalah yang diberikan (Ardiyansyah, Rahayuningsih, & Maulana, 2018).

2.2. Data Mining

Data mining merupakan proses iteratif dan interaktif untuk menemukan pola atau model baru yang dapat digeneralisasi untuk masa yang akan datang, bermanfaat dan dapat dimengerti dalam suatu database yang sangat besar (massive databases). Data mining berisi pencarian trend atau pola yang diinginkan dalam database besar untuk membantu pengambilan keputusan di waktu yang akan datang (Ardiyansyah, Rahayuningsih, & Maulana, 2018).
Inti dari Data mining adalah menggali data untuk mendapatkan informasi penting yangtersembunyi dalam data tersebut. Data mining mendukung task/fungsionalitas yang meliputi:

a. Prediktive
   Menghasilkan model berdasarkan sekumpulan data yang dapat digunakan untuk memperkirakan nilai data yang lain. Metode yang termasuk prediktive data mining :
   1) Klasifikasi : pembagian data ke dalam beberapa kelompok/kelas yang telah ditentukansebelumnya.
   2) Regresi : memetakan data ke suatu prediction variable.
   3) Time Series Analysis : pengamatan perubahan nilai atribut dari waktu ke waktu.

b. Deskriptive
   Mengidentifikasi pola atau hubungan dalam data untuk menghasilkan informasi baru. Metode yang termasuk deskriptive data mining :
   1) Clustering: mengelompokkan beberapa objek yang serupa ke dalam sebuah cluster, dan yang tidak serupa ke cluster yang lain.
   2) Association rules : identifikasi hubungan antara data yang satu dengan yang lainnya.
   3) Summarization : pemetaan data ke dalam subset dengan deskripsi sederhana.
   4) Sequence discovery : identifikasi pola sekuensial dalam data.

2.3. Metode C.45

Pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menrapkan serangkaian aturan keputusan.
Algoritma yang dapat dipakai dalam pembentukan pohon keputusan.

1. ID3
2. CART
3. C4.5

C4.5 Merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh Quinlan Algoritma C4.5 banyak digunakan peneliti untuk melakukan tugas klasifikasi. Output dari algoritma C4.5 adalah sebuah pohon keputusan atau sering dikenal dengan decissin tree. Dalam beberapa penelitian algoritma C4.5 ini menjadi pilihan terbaik dibandingkan dengan beberapa algoritma klasifikasi lain (Ardiyansyah, Rahayuningsih, & Maulana, 2018). Tahapan Algoritma C4.5 adalah, sebagai berikut:

3. METODOLOGI

Tahapan yang dilalui dalam penelitian, pembangunan konsep, atau penyelesaian kasus, dituliskan pada bagian metodologi.

1. Pilih atribut sebagai akar

$$Entropy\ (S) = \sum_{i=1}^{n} - p * log_2\ pi$$

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

2. Buat cabang untuk tiap- tiap nilai
3. Bagu kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

4. HASIL DAN PEMBAHASAN

4.1. Data Sampling

Data Sampling
Dalam penerapan C.45 ini, kita mengambil data dari UCI Machine Learning Repository tentang data para blogger disertai kriteria apa saja yang menjadikan mereka menjadi professional blogger.
Berikut 100 data tersebut :

Table 1. Nilai Entropy dan Information Gain dari Data Sampling

Pendidikan	Kepentingan Politik	Topik	Penggunaan media lokal	Pengaruh Politik Lokal dan Sosial	Professional Blogger
Ph.D.	Reformasi	impression	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Tidak Berkepentingan	tourism	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Tidak Berkepentingan	news	no	yes	yes
M.Sc.	Tidak Berkepentingan	news	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
Ph.D.	Tradisional	political	no	no	yes
Ph.D.	Tradisional	political	no	no	no
M.Sc.	Tradisional	tourism	no	no	yes
Ph.D.	Tradisional	tourism	no	yes	yes
M.Sc.	Reformasi	news	no	no	yes
Ph.D.	Reformasi	political	no	yes	no
Bachelor	Tradisional	news	yes	yes	no
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Reformasi	impression	no	yes	yes
M.Sc.	Reformasi	political	no	yes	yes
Ph.D.	Tradisional	political	no	yes	yes
M.Sc.	Reformasi	impression	no	yes	yes
Ph.D.	Tradisional	tourism	no	yes	no
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Reformasi	news	no	yes	yes
Ph.D.	Tradisional	political	yes	yes	no
Bachelor	Reformasi	tourism	no	no	no
Ph.D.	Reformasi	news	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
Bachelor	Tradisional	impression	yes	no	yes
Ph.D.	Tradisional	political	no	yes	yes
Ph.D.	Reformasi	impression	yes	no	yes
M.Sc.	Reformasi	scientific	no	yes	no
Ph.D.	Tradisional	political	no	yes	yes
Bachelor	Reformasi	scientific	no	yes	no
M.Sc.	Tradisional	tourism	no	yes	no
Bachelor	Tradisional	political	no	yes	yes
Ph.D.	Reformasi	impression	no	no	yes
M.Sc.	Reformasi	tourism	no	no	yes
M.Sc.	Tidak Berkepentingan	scientific	no	no	yes
M.Sc.	Tidak Berkepentingan	impression	yes	yes	no
M.Sc.	Tradisional	scientific	no	yes	no
M.Sc.	Reformasi	impression	yes	no	yes
Ph.D.	Reformasi	political	no	yes	no
M.Sc.	Reformasi	news	yes	yes	yes
Ph.D.	Reformasi	political	no	yes	yes

M.Sc.	Tradisional	news	no	yes	no
M.Sc.	Reformasi	tourism	no	no	yes
M.Sc.	Tidak Berkepentingan	news	no	yes	yes
Bachelor	Tidak Berkepentingan	impression	no	no	no
Bachelor	Tradisional	impression	no	no	no
M.Sc.	Tradisional	news	no	yes	no
M.Sc.	Reformasi	impression	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Tidak Berkepentingan	tourism	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Tidak Berkepentingan	news	no	yes	yes
M.Sc.	Tidak Berkepentingan	news	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
Ph.D.	Tradisional	political	no	no	yes
Ph.D.	Tradisional	political	no	no	no
M.Sc.	Tradisional	tourism	no	no	yes
M.Sc.	Tradisional	tourism	no	yes	yes
M.Sc.	Reformasi	news	no	no	yes
Ph.D.	Reformasi	impression	no	yes	no
Bachelor	Tradisional	news	yes	yes	no
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Reformasi	impression	no	yes	yes
M.Sc.	Reformasi	political	no	yes	yes
Ph.D.	Tradisional	political	no	yes	yes
M.Sc.	Reformasi	political	no	yes	yes
Ph.D.	Tradisional	impression	no	yes	no
M.Sc.	Reformasi	political	no	yes	yes
M.Sc.	Reformasi	news	no	yes	yes
M.Sc.	Tradisional	political	no	yes	no
Bachelor	Reformasi	tourism	no	no	no
Ph.D.	Reformasi	news	no	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
Bachelor	Tradisional	impression	yes	no	yes
Ph.D.	Tradisional	political	no	yes	yes
Ph.D.	Reformasi	impression	yes	no	yes
M.Sc.	Reformasi	scientific	no	yes	no
Ph.D.	Tradisional	political	no	yes	yes
Bachelor	Reformasi	scientific	no	yes	no
M.Sc.	Tradisional	tourism	no	yes	no
Bachelor	Tradisional	political	no	yes	yes
Ph.D.	Reformasi	impression	no	no	yes
M.Sc.	Reformasi	tourism	no	no	yes
M.Sc.	Tidak Berkepentingan	impression	no	no	yes
M.Sc.	Tidak Berkepentingan	impression	yes	yes	no
M.Sc.	Tradisional	scientific	no	yes	no
M.Sc.	Reformasi	impression	yes	no	yes
Ph.D.	Reformasi	political	no	yes	no

M.Sc.	Reformasi	news	yes	yes	yes
Ph.D.	Reformasi	political	no	yes	yes
M.Sc.	Tradisional	news	no	yes	no
M.Sc.	Reformasi	tourism	no	no	yes
M.Sc.	Tidak Berkepentingan	impression	no	yes	yes
Bachelor	Tidak Berkepentingan	impression	no	no	no
Bachelor	Tradisional	impression	no	no	no
M.Sc.	Tradisional	news	no	yes	no
M.Sc.	Reformasi	impression	no	yes	yes

4.2. Entropy and Information Gain

Tahap- tahap untuk menghitung data dengan metode C. 45 adalah :

a. Tentukan jumlah kasus secara keseluruhan dengan Professional Blogger sebagai tumpuan karena dalam hal ini kita ingin mencari presentasi professional blogger dengan atribut yang ada.
a. Lalu tentukan Atribut lain yang mendukung seperti pendidikan, kepentingan politik, topik, penggunaan media lokal, pengaruh politik lokal dan sosial.
b. Tentukan jumlah kasus dari masing- masing atribut dan apakah kasus tersebut tergolong dalam kriteria
c. professional blogger (bernilai Ya di atribut professional blogger) atau sebaliknya.
d. Tentukan Entropi dari seluruh kasus dengan cara membagikan kasus yang tergolong professional blogger (bernilai Ya di atribut professional blogger) atau sebaliknya di bagi dengan jumlah kasus dikali $\log_2$ dan seterusnya sesuai dengan rumus.
e. Tentukan entropi masing- masing atribut dan gain dari tiap - tiap kategori.

Nilai entropy dan information gain dari data sampling dapat dilihat pada tabel dibawah ini:

Table 1. Nilai Entropy dan Information Gain dari Data Sampling

Atribut	Jumlah Kasus	Ya	Tidak	Entropy	Gain
Akar (Professional Blogger)	100	68	32	0.90438146	
Pendidikan					0.07973591
Bachelor	14	4	10	0.86312057	
M.Sc.	47	34	13	0.8507707	
Ph.D.	39	30	9	0.77934984	
Kepentingan Politik					0.07713119
Tradisional	34	16	18	0.99750255	
Reformasi	52	42	10	0.70627409	
Tidak Berkepentingan	14	10	4	0.86312057	
Topik					0.08120284
Impression	24	16	8	0.91829583	
Political	35	28	7	0.72192809	
Tourism	15	10	5	0.91829583	
News	19	13	6	0.89974376	
Scientific	7	1	6	0.59167278	
Pengunaan Media Lokal					0.00201384
Ya	13	8	5	0.9612366	
Tidak	87	60	27	0.8935711	
Pengaruh Politik Lokal dan Sosial					0.0015347
Ya	72	48	24	0.91829583	
Tidak	28	20	8	0.86312057	

Untuk menghitung nilai *entropy* dan *information gain* digunakan rumus sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2 pi$$

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

Cara perhitungan yang di jelaskan dalam tahap sebelumnya adalah sebagai berikut :

Entropi (All) = $\sum_{i=1}^{n} - pi * \log_2 pi$

$$= -\frac{68}{100}\log_2\frac{68}{100} + -\frac{32}{100}\log_2\frac{32}{100}$$

$$= 0.90438146$$

Entropi (Pendidikan, Bachelor) = $\sum_{i=1}^{n} - pi * \log_2 pi$

$$= -\frac{4}{14}\log_2\frac{4}{14} + -\frac{10}{14}\log_2\frac{10}{14}$$

$$= 0.86312057$$

Hitung entropi tiap- tiap atribut seperti perhitungan entropi (pendidikan, bachelor). Setelah semua atribut selesai dalam 1 kategori (pendidikan, kepentingan politik, topik, ,...), hitung gain dari kategori tersebut.

Gain (Pend., All) = $Entropy(S) - \sum_{i=1}^{n}\frac{|S_i|}{|S|} * Entropy\ (S_i)$

$$= 0{,}90438146 - \left(\frac{14}{100}(0.86312057) + \frac{47}{100}(0.8507707) + \frac{39}{100}(0.77934984)\right)$$

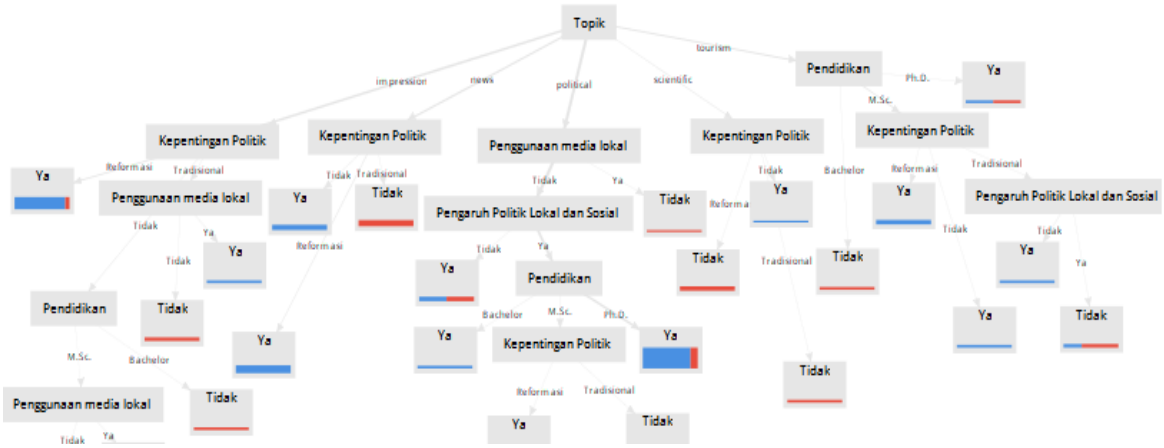
$$= 0.07973591$$

dst...

Hitunglah gain dari kategori lainnya seperti contoh perhitungan di atas.

4.3. Pohon Keputusan

Setelah mendapatkan hasil dari entropy dan gain dari masing-masing atribut yang ada pada data sampling, kemudian dilakukan pengujian nilai yang didapatkan dengan menggunakan rapid miner.
Output yang dihasilkan adalah pohon keputusan yang menunjukkan pengaruh atribut terhadap terciptanya professional blogger.
Setelah melakukan proses pada Gambar 1 maka didapatkan hasil akhir pohon keputusan seperti dibawah ini :



Gambar 1. Pohon Keputusan Akhir

Dengan memperhatikan pohon keputusan pada gambar 2 diketahui bahwa semua kasus sudah masuk dalam kelas dengan demikian pohon keputusan pada gambar merupakan pohon keputusan terakhir yang terbentuk.Setelah pohon terbentuk, dihasilkan sejumlah aturan dalam pohon tersebut. Contoh aturan yang dapat terbentuk dari pohon pada gambar 2 adalah sebagai berikut:
"JIKA topik = impression DAN kepentingan politik = reformasi MAKA professional blogger = YA"
"JIKA topik = impression DAN kepentingan politik = reformasi DAN pendidikan = Bachelor MAKA professional blogger = TIDAK"
"JIKA topik = news DAN kepentingan politik = reformasi MAKA professional blogger = YA"
"JIKA topik = news DAN kepentingan politik = tradisional MAKA professional blogger = TIDAK"

5. KESIMPULAN DAN SARAN

Berdasarkan analisa penggunaan data minning dengan algoritma C4.5 dapat digunakan pada data set UCI Machine Learning Repository, dan dapat kita lihat bahwa topik sangat berpengaruh terhadap munculnya seorang professional blogger.
Adapun saran untuk penelitian selanjutnya adalah dapat menggunakan Dataset yang berbeda yang dapat di ambil dari *UCI Machine Learning Repository*, dapat menggunakan data preprocessing seperti menambahkan fitur selection dan

menggunakan model Agortima yang berbeda dengan dataset yang sama.

DAFTAR PUSTAKA

[1] Ardiyansyah, Rahayuningsih, P. A., & Maulana, R. (2018). Analisis Perbandingan Algoritma Klasifikasi Data Mining. *JURNAL KHATULISTIWA INFORMATIKA, VOL. VI, NO. 1 JUNI 2018*, 20-28.

[2] Gharehchopogh, F. S., & Khaze, S. R. (2012). Data Mining Application for Cyber Space Users. *International ournal of Computer Applications (0975 – 888) Volume 47– No.18*, 40-46.

[3] Luik, J. E. (t.thn.). BLOGGING AS EMPOWERMENT: Self-presentation of bloggers in Surabaya, Indonesia.

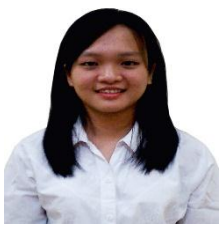
[4] Mortensen, T., & Walker, J. (2002). Blogging thoughts: personal publication as an online research tool. *Researching ICTs in Context*, 249-279.

[5] Santoso, T. B. (2014). ANALISA DAN PENERAPAN METODE C4.5 UNTUK PREDIKSI LOYALITAS PELANGGAN. *Jurnal Ilmiah Fakultas Teknik LIMIT'S Vol. 10 No.1*.

[6] Saragih, Rejoice Iboy Erwin, and Darsono Nababan. "Penerapan Algoritma Genetika Pada Pengenalan Paragraf." *Journal Information System Development (ISD)* 4.1 (2019).

BIODATA PENULIS

Robby
Mahasiswa Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Pelita Harapan Kampus Medan, saat ini aktif sebagai programmer pada Pusat Sistem Informasi UPH Medan



Lavinia
Mahasiswa Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Pelita Harapan Kampus Medan, tertarik dengan data analytic saat sedang menyelesaiakan penelitian tentang datamining.

Darsono Nababan
Dosen tetap pada Fakultas Ilmu Komputer Universitas Pelita Harapan Medan , mengajar matakuliah Business Inteligence and Data Analytic dan aktif melakukan penelitian dengan topik datamining dan Security.