Identifikasi Berita HOAX Berbasis Web Menggunakan Algoritma C4.5

Edward

[1)] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta 11440 Indonesia
edwardharuntx@gmail.com

ABSTRACT

Now we can get news from anywhere and anytime, for example, such as news sites or social media that are commonly used, there are many social media options such as Facebook, Twitter, WhatsApp, and many other. But the information that is disseminated does not really contain some news that does not match the truth or is often called HOAX. HOAX is information that is untrue or real, but it is made to appear as if it were. The author's goal in creating HOAX classification system news is to help potential users to be able to identify a news item whether the news is HOAX news or not. This system is designed based on a website using the Python programming language. The C4.5 algorithm will be used to classify or determine whether a news is HOAX news or not. The results of the C4.5 method have an excellent accuracy value with an accuracy value of 99.6%

Key words

C4.5, HOAX, Social Media, Python

1. Pendahuluan

Saat ini penyebaran HOAX di media sosial terus meningkat. Sampai saat ini HOAX terus berkembang dan sedang berada di tingkat yang mengkhawatirkan. Situasi ini yang dapat menyebabkan kecemasan dan kepanikan publik. Walaupun terkadang HOAX tidak bersifat mengancam, tetapi HOAX dapat membuat persepsi baru yang dapat mempengaruhi kondisi sosial dan politik suatu negara.

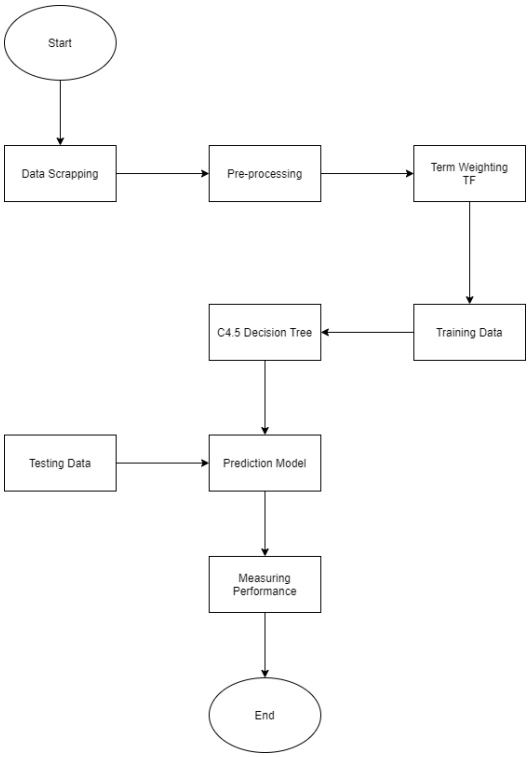
Berita HOAX mempunyai beberapa ciri khas, seperti berasal dari situs yang tidak dapat dipercaya, menggunakan kata atau kalimat provokatif, berita tidak diketahui atau memiliki sumber, dan lain-lain.

Dalam penelitian ini, penulis mengidentifikasi berita hoax dengan menggunakan metode klasifikasi *Decision Tree* C4.5 karena metode ini sangat kuat dan terkenal klasifikasi dan metode prediksi [1]. *Decision Tree* C4.5 merupakan perpanjangan dari ID3 dengan beberapa peningkatan seperti mungkin menggunakan data berkelanjutan, nilai yang hilang, dan pemangkasan. C4.5 juga dapat menggunakan atribut dengan bobot yang berbeda. Selain itu, penelitian ini juga melakukan skenario pengujian menggunakan Term Frequency untuk mencari fitur representasi nilai / bobot judul dan konten dokumen dalam dataset. Tes dilakukan pada Berita yang menggunakan bahasa Indonesia sebanyak 1.525 berita yang diambil dari proses *scrapping*. Tujuan dari ini penelitian adalah untuk menerapkan metode yang dijelaskan di atas dan mencari tahu kinerja sistem dalam mengidentifikasi berita hoax menggunakan perhitungan akurasi dengan mengacu pada tabel *Confusion Matrix*. Dari sistem yang dibangun, penulis juga ingin mencari tahu fitur apa saja yang berpengaruh mengidentifikasi berita HOAX, fitur apa yang paling sering digunakan dalam berita HOAX.

2. Metode Penelitian

2.1 Sistem *Flowchart*

Pada Gambar 1, penulis merepresentasikan *flowchart* bagaimana seluruh sistem bekerja dalam penelitian ini. Sistem dimulai dari *Data Scrapping*, lalu data tidak terstruktur akan diubah menjadi lebih terstruktur melalui *pre-processing*. Kemudian setiap kata dalam dataset akan dibobot menggunakan pembobotan TF untuk mengekstrak fitur. Sistem kemudian membagi data menjadi dua bagian, yaitu pelatihan data untuk *training* dan data *testing* untuk menguji pembelajaran hasil dari sistem dengan menggunakan empat data split yang berbeda rasio (65:35, 70:30, 75:25, 80:20). Penggunaan berbeda rasio bertujuan untuk mengetahui rata-rata akurasi kinerja sistem. Kinerja sistem diukur menggunakan akurasi *Confusion Matrix*.



Gambar 1. Flowchart

2.2 Data Scrapping

Dataset yang digunakan dalam penelitian ini diambil dari proses scrapping di detik.com dan turnbackhoax.id. Data yang diambil adalah sebanyak 1.525 data berita yang dibagi menjadi 880 data berita non-hoax dan 645 data berita hoax.

2.3 Pre-Processing

Dalam mengolah data dibutuhkan beberapa proses sehingga data mentah yang sudah didapat sebelumnya bisa diproses dan dikelola dengan mudah. Pre-processing adalah proses pengubahan bentuk data teks tidak terstruktur ke dalam bentuk terstruktur sesuai dengan kebutuhan [2]. Dalam sistem yang dibangun, ada empat tahapan preprocessing, pertama *case folding* untuk menghilangkan karakter selain a-z, lalu *tokenization* pemotongan teks berdasarkan tiap kata dengan menggunakan penanda spasi, kemudian *stopword removal* untuk membuang kata-kata yang dianggap tidak penting, dan yang terakhir adalah *stemming* untuk mengubah setiap kata dengan akar kata. Untuk melihat contoh proses *pre-processing* bisa dilihat pada tabel 1.

Tabel 1. Contoh *Pre-processing*

Process	Teks
Teks Awal	Banser buat Natal di Cilacap dan Kebumen lebih aman.
Case Folding	Banser buat natal di cilacap dan kebumen lebih aman,
Normalization	banser:1, buat:1, natal:1, di:1, cilacap:1, dan:1, kebumen:1, lebih:1, aman:1
Stopword Removal	banser:1, natal:1, cilacap:1, kebumen:1, aman:1
Stemming	banser:1, natal:1, cilacap:1, kebumen:1, aman:1

2.4 *Term Frequency* (TF)

Term Frequency digunakan untuk menunjukkan seberapa sering ekspresi sebuah (*istilah*, kata) muncul dalam sebuah dokumen. *Term Frequency* menunjukkan pentingnya *istilah* tertentu dalam keseluruhan dokumen. Hasil dari *Term Frequency* akan digunakan untuk penghitungan kepadatan *keywords*[3]. rumus dari *Term Frequency:*

$$\text{tf(t,d)} = \frac{}{\sum_{x \epsilon d} fr(x,t) \sum_{x \epsilon d} fr(x,t)} \quad (1)$$

Dimana tf(td) adalah banyak istilah dalam sebuah dokumen.

2.5 *C4.5*

Algoritma C4.5 merupakan salah satu algoritma data mining yang termasuk dalam kelompok klasifikasi. Algoritma C4.5 adalah algoritma yang digunakan untuk memproduksi sebuah *decision tree.* Algoritma C4.5 merupakan algoritma yang sangat populer yang digunakan oleh banyak peneliti di dunia, hal ini dijelaskan oleh Xindong Wu dan Vipin Kumar dalam bukunya yang berjudul The Top Ten Algorithms in Data Mining. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang diciptakan oleh Han Jiawei. Algoritma C4.5 ini merupakan ekspansi dari pendahulunya yaitu kalkulasi ID3[4]. Algoritma ini meningkatkan algoritma ID3 dengan mengatur properti yang berkelanjutan dan berlainan, missing value dan pemotongan tree setelah konstruksi. *Decision tree* yang dibuat dengan C4.5 dapat digunakan untuk mengelompokkan dan seringkali cenderung mengarah ke *statistical classifier*.

Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan decision tree yang paling kecil ukurannya, atau atribut yang dapat memisahkan objek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling *purest* (paling bersih). Ukuran *purity* dinyatakan dengan tingkat impurity, dan untuk menghitungnya, dapat dilakukan dengan menggunakan konsep Entropy, Entropy menyatakan impurity suatu kumpulan objek.
Formula mencari *entropy* sebagai berikut:

$$Entropy\ (S) = -\sum_{i=1}^{m} p_{i\ log_n\ p_i}$$
$$Entropy\ (S) = -\sum_{i=1}^{m} p_{i\ log_n\ p_i} \quad (2)$$

Dimana *S* adalah himpunan (dataset) kasus, *m* adalah banyaknya partisi S, dan *Pi* adalah probabilitas yang di dapat dari Sum(Ya) dibagi Total Kasus

Information gain adalah kriteria yang paling populer untuk pemilihan atribut. Algoritma C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Hanya saja dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan *Gain Ratio* dengan rumus:

$$gain\ ratio\ (a) = \frac{gain\ (a)}{split\ (a)}$$
$$gain\ ratio\ (a) = \frac{gain\ (a)}{split\ (a)} \quad (3)$$

Dimana *a* adalah atribut, gain(a) adalah information gain pada atribut a, dan Split(a) adalah split information pada atribut a

Atribut dengan nilai *Gain Ratio* tertinggi dipilih sebagai atribut test untuk simpul. Dengan *gain* adalah *information gain*. Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan apa yang disebut sebagai *split information*. *SplitInfo* menyatakan entropy atau informasi potensial dengan rumus:

$$SplitInfo\ (S, A) = -\sum_{i-1}^{n} \frac{S_i}{S} \log_2 \frac{Si}{s}$$
$$SplitInfo\ (S, A) = -\sum_{i-1}^{n} \frac{S_i}{S} \log_2 \frac{Si}{s} \quad (4)$$

Dimana S adalah ruang (data) sample yang digunakan untuk training, A adalah atribut, dan Si adalah jumlah sample untuk atribut i.

Keterangan Xi menyatakan sub himpunan ke-i pada sampel X. Sedangkan untuk mencari nilai Gain digunakan rumus :

$$Gain\ (S, A) = Entropy\ (S) - Entropy\ (S, A)$$
$$Gain\ (A) = Entropy\ (S) - \sum_{i=1}^{k} \frac{|Si|}{|S|}\ x\ Entropy(Si) \quad (5)$$

Dimana S adalah ruang (data) sampel yang digunakan untuk training, A adalah atribut, |Si| adalah jumlah sampel untuk nilai V, |S| adalah jumlah seluruh sample data, dan Entropy(Si) adalah entropy untuk sampel-sampel yang memiliki nilai i.

2.6 Measuring Performance

Measuring Performance adalah langkah untuk menganalisa dan mengevaluasi kinerja sistem yang

dibangun. Performa diukur dengan nilai akurasi. Confusion Matrix adalah metode yang digunakan untuk menghitung akurasi dalam konsep Data Mining. Confusion Matrix ditunjukkan pada Tabel 2.

Tabel 2. Confusion Matrix

	Predicted Class Yes	Predicted Class No
Actual Class Yes	True Positive (TP)	False Negative (FN)
Actual Class No	False Positive (FP)	True Negative (TN)

True Positive (TP) adalah saat aktual dan kelas prediksi keduanya adalah tipuan. False Positive (FP) adalah saat kelas sebenarnya bukan hoax tapi kelas yang diprediksi adalah hoax. Benar Negatif (TN) adalah saat kelas aktual dan prediksi keduanya bukan hoax. False Negative (FN) adalah saat aktual class is hoax tapi class yang diprediksi bukan hoax. Dari tabel matriks kebingungan, nilai akurasi dapat dihitung. Akurasi adalah tingkat kedekatan antara nilai prediksi dan nilai sebenarnya. Akurasi adalah digunakan untuk mengevaluasi jumlah kelas prediksi itu sesuai dengan kelas sebenarnya [15]. Berikut ini adalah rumus akurasi:

$$\text{Accuracy} = (TP+TN) \,/\, (TP+FP+TN+FN) \quad (6)$$

Dimana TP adalah True Positive, TN adalah True Negative, FP adalah False Positive, dan FN adalah False Negative.

3. Hasil Percobaan

Bagian ini menjelaskan hasil pengujian sistem yang dibangun. Pengujian dalam penelitian ini bertujuan untuk mengetahui kinerja sistem dalam mengklasifikasikan berita hoax.

Tabel 3 Hasil Perhitungan 1.525 Data Berita

Training-Testing	Akurasi
65%-35%	99.62%
70%-30%	99.52%
75%-25%	99.53%
80%-20%	99.58%
Rata-rata	99.56%

Bisa dilihat pada Tabel 3 bahwa nilai akurasi klasifikasi berita hoax menggunakan metode C4.5 memiliki nilai akurasi yang sangat tinggi sebesar 99.56%.

4. Kesimpulan

Kesimpulan yang diperoleh dalam keseluruhan proses yang dimulai dari perancangan, pembuatan, dan pengujian web "Identifikasi Berita HOAX Berbasis *Web* Menggunakan Algoritma C4.5" adalah:

1. Algoritma C4.5 dapat menghasilkan nilai akurasi yang sangat baik dengan tingkat akurasi pada data pengujian untuk data uji sebanyak 100 sebesar 99.7%, dan data uji sebanyak 1.525 sebesar 99.56%
2. Terdapat tiga fitur dari berita yang sangat mempengaruhi hasil klasifikasi yaitu url mengandung "URL dalam konten", "URL Berita", dan tanggal tepi pada berita.
3. Hasil klasifikasi berita masih belum konsisten, karena hasil klasifikasi masih bisa berubah berdasarkan ciri-ciri bentuk berita.

REFERENSI

[1] Han, Jiawei, Micheline Kamber, and Jian Pei, Data Mining Concepts and Techniques Third Edition. (Waltham:Elsevier Inc, 2012), h. 331.

[2] Siregar, Z.U., Siregar, R.R., Arianto, R., 2019. Jurnal Kilat. Klasifikasi Sentiment Analysis pada Komentar Peserta Diklat menggunakan Metode K-Nearest Neighbor, 8 (1), h. 81-92.

[3] Rasywir, E., & Purwarianti, A., 2016. Eksperimen pada sistem klasifikasi berita hoax berbahasa Indonesia berbasis pembelajaran mesin. *Jurnal Cybermatika*, h. *3*.

[4] Han, Jiawei, Micheline Kamber, and Jian Pei, Data Mining Concepts and Techniques Third Edition. (Waltham:Elsevier Inc, 2012), h. 331-332.

Edward, memperoleh gelar S. Kom dari Universitas Tarumanagara, Indonesia tahun 2021.