RABIT : Jurnal Teknologi dan Sistem Informasi Univrab ISSN CETAK : 2477-2062
Volume 6 No. 2 | Juli 2021 : Hal : 130-137 ISSN ONLINE : 2502-891X

MENGIDENTIFIKASI HOAX PADA HASIL PENCARIAN BERITA ONLINE DENGAN TEKNIK WEB SCRAPING DAN ALGORITMA C4.5

[1)] Diki Arisandi, [2)] Zul indra, [3)] Kartini
[1,2*,3]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Abdurrab
[1,2*,3,4)]Jl. Riau Ujung No 73 Pekanbaru – Riau - Indonesia
E-mail : *diki@univrab.ac.id, *zul.indra@univrab.ac.id, kartini@student.univrab.ac.id*

ABSTRAK
Berita online merupakan salah satu produk jurnalistik yang melaporkan fakta atau peristiwa yang diproduksi dan didistribusikan melalui internet. Namun tidak seluruh informasi yang disebarkan melalui media online berupa fakta atau sering disebut hoax. Banyaknya terjadi pemalsuan berita tentunya sangat berdampak pada masyarakat yang mengkonsumsi berita tersebut, sehingga bisa menimbulkan kesalahan persepsi maupun tindakanyang tidak semestinya. Pada penelitian ini peneliti menggunakan teknik web scraping untuk mengekstraksi konten dari hasil pencarian pada search engine, dan dilanjutkan dengan penggunaan algoritma C4.5 untuk proses klasifikasi. Ada tiga parameter yang menjadi acuan yaitu ajakan untuk menyebarkan berita, kredibilitas sumber, dan judul yang memprovokasi. Hasil dari penelitian ini berupa pohon keputusan yang dapat mengklasifikasikan suatu konten berita tergolong hoax atau bukan. Dari hasil percobaan yang dilakukan pada penelitian ini, hasil akurasi dari klasifikasi hoax dengan teknik web scraping dan algoritma C4.5 mencapai angka keberhasilan sebesar 80%.

Kata Kunci: berita online, *hoax*, web scraping, algoritma C4.5, pohon keputusan

ABSTRACT
Online news is a journalistic product reports the facts or events that are produced and distributed via internet. However, not all of the information through online media is a real facts, also described as hoax. The large number of hoax news occurs, of course, deliver the impact on the people who look on the news, so it could cause misperceptions or inappropriate actions. We exploit a web scraping technique to extract the content from search search engines results. Furthermore, we employ the C4.5 algorithm for the classification process. There were three parameters as references: invitation to spread the news, credibility of the sources, and provoking title. The results of this work were a decision tree, that able to classify a news content as a hoax or legitimate. From the experiments which carried out, the accuracy of classification using the web scraping and C4.5 algorithm achieved 80% of success rate in determining the hoax.

Keyword: *online news, hoax, web scraping, C4.5 algorithm, decission tree.*

PENDAHULUAN

Banyaknya media berita online saat ini memberikan manfaat tersendiri bagi seluruh masyarakat di Indonesia terutama dari sisi kemudahan akses [1], selain memberikan informasi, berita online juga dapat dijadikan sebagai wadah dalam memberikan masukan, kritik maupun saran dalam pembangunan [2]. Di sisi lain perlu adanya dorongan kepada semua lapisan masyarakat agar memiliki etika bagaimana memanfaatkan media *online* [3]. Banyak sekali pengguna media online yang memanfaatkan media untuk hal-hal yang sifatnya negatif dan dapat merugikan semua pihak, baik itu pemerintah maupun masyarakat itu sendiri[4]. Saat ini di Indonesia marak terjadi peristiwa penyebaran berita palsu atau yang disebut *hoax*. Adapun dalam hal penyebarannya, berita *hoax* sangat banyak tersebar dari media berita *online* [5].

Dampak yang dihasilkan oleh berita *hoax* merupakan dampak yang tidak bisa disadari secara langsung, karena berita hoax akan langsung menyerang pemikiran pembacanya dan jika tidak berhati-hati akan mempengaruhi cara berpikir pembacanya pula [6]. Banyaknya berita *hoax* tentunya sangat berdampak pada masyarakat yang mengkonsumsi berita tersebut [7].

Hasil penelitian terdahulu menyebutkan bahwa sumber berita *online* dapat diklasifikasikan agar lebih mudah untuk diinterpretasikan [8],

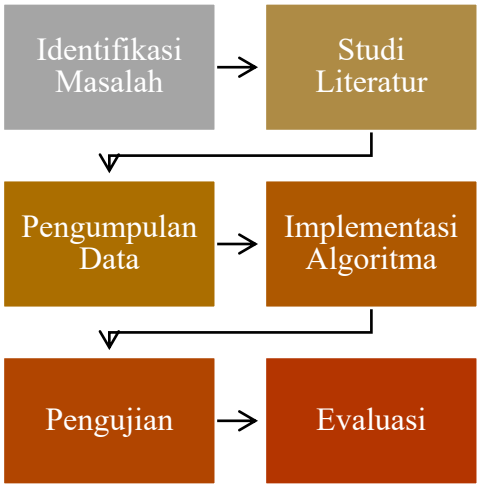
sehingga maknanya tidak bias dan dikategorikan sebagai *hoax* [9]. Namun banyaknya sumber berita *online* pada saat ini, mendukung upaya untuk dikembangkannya sebuah pengumpul data, yang kemudian hasilnya dapat dimanfaatkan untuk berbagai macam seperti *opinion mining, topic classification, text summarization* dan juga identifikasi *hoax* [10].

Pada penelitian ini, peneliti menggunakan teknik *scraping* yang dipadukan dengan Algoritma C4.5. Teknik *scraping* merupakan proses pengambilan sebuah dokumen semi-terstruktur dari internet, kemudian diekstrak untuk diambil data tertentu dari halaman tersebut agar bisa digunakan bagi kepentingan lain. Manfaatnya ialah agar informasi yang diambil atau digunakan lebih terfokus sehingga memudahkan dalam melakukan pencarian sesuatu [11]. Teknik *scrapping* dapat dilakukan melaui teknik *regular ekspression*, yang ditentukan oleh pola yang mengawali dan mengakhiri suatu konten utama pada halaman situs. Secara singkat *regular expression* menyediakan cara untuk memanipulasi dan mencocokan string sesuai dengan formula yang dibuat [12].

Algoritma C4.5 merupakan salah satu teknik yang sering digunakan untuk menghasilkan beberapa aturan-aturan dan sebuah pohon keputusan dengan tujuan untuk meningkatkan keakuratan dari prediksi yang sedang dilakukan. Penggunaan algoritma ini banyak diaplikasikan pada bidang ekonomi, kesehatan, pendidikan, dan lain sebagainya [13]. Algoritma C4.5 berbentuk pohon keputusan dimana terdapat *node internal* yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Secara umum pohon keputusan dapat memiliki akurasi yang baik, tergantung pada kesediaan kualitas data yang diolah [14].

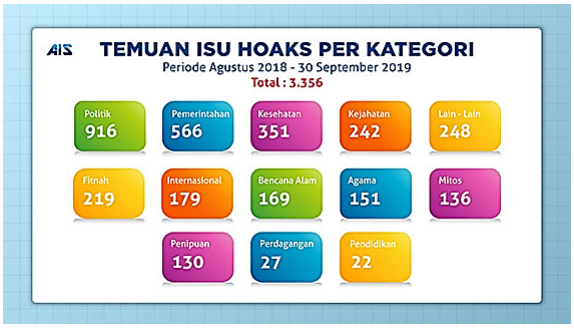
METODE

Pada penelitian ini, peneliti mengikuti tahapan kerja sebagai berikut:



Gambar 1. Tahapan Penelitian

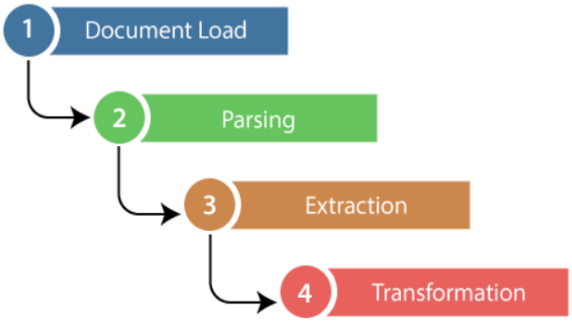
Tahapan identifikasi masalah pada penelitian ini yaitu peneliti melakukan identifikasi permasalahan yang ada, dimana Kementrian Komunikasi dan Informatika Republik Indonesia melaporkan ribuan sumber alamat penyebar hoax dan menemukan ratusan isu hoax yang berpotensi menyesatkan pembaca [5] seperti terlihat pada gambar 2 berikut:



Gambar 2. Temuan Isu Hoax

Berdasarkan identifikasi yang telah dilakukan, penulis melakukan penelusuran referensi terkait bagaimana cara mengidentifikasi *hoax* dengan teknik scraping dan beberapa algoritma klasifikasi. Teknik *scraping* dipilih karena informasi yang akan diekplorasi berjumlah sangat banyak, sehingga perlu dilakukan

ekstraksi data dan disimpan dalam format yang lebih mudah untuk diolah [15] seperti terlihat pada gambar 3.



Gambar 3. Tahapan Teknik Scraping

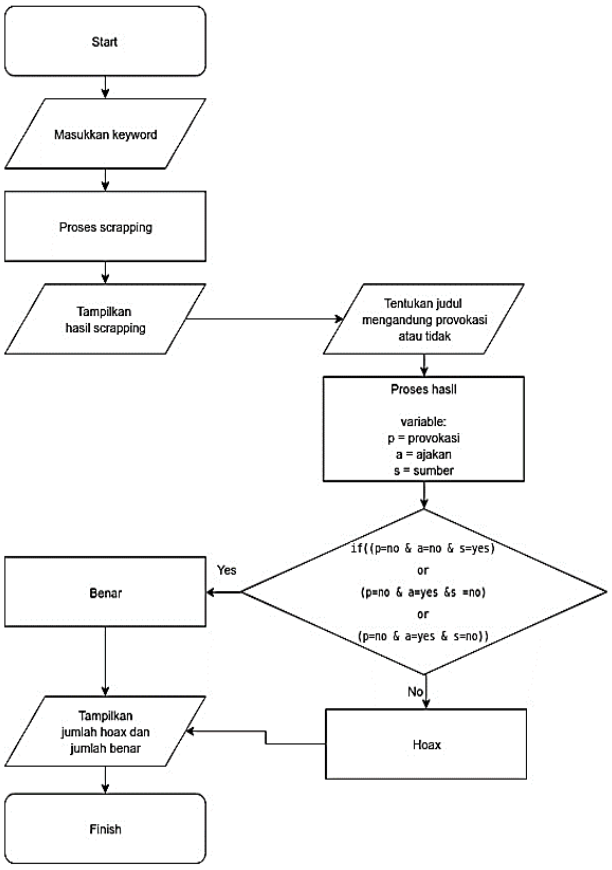
Data yang diekstrak dari teknik *scraping* berasal dari website pencarian bing sebanyak dua puluh halaman hasil pencarian yang akan berfungsi sebagai data *training*. Sedangkan untuk data *testing*, peneliti mengambil sampel informasi dari *website* https://turnbackhoax.id/. Setelah itu, proses algoritma dilanjutkan dengan menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai indeks *entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai index entropy (persamaan 1).

$$entropy(i) = -\sum_{j=1}^{m} \quad f(i,j).\log_2 f[(i.j)] \ldots . (1)$$

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada seperti pada persamaan 2, lalu diulangi langkah ini hingga semua *record* terpartisi.

$$Gain(S,A) = Entropy(S) \sum_{i=1}^{m} \quad .\frac{|Si|}{|S|} * Entropy\ Si (2)$$

Dalam proses klasifiksi *hoax*, algoritma C4.5 menjadi pilihan untuk diimplementasikan. Hal ini dikarenakan algoritma ini tidak memerlukan proses komputasi yang lama [16] dan mempunyai tingkat akurasi yang bagus dibandingkan dengan dengan algoritma sejenis [17],[18]. Atribut yang digunakan untuk penentuan *hoax* atau tidak nya suatu konten berita adalah judul provokasi (P), ajakan menyebarkan (A), dan kredibilitas sumber (S). Hasil ekstraksi data dan klasifikasi dengan algoritma C4.5 lalu akan diuji dan dievaluasi agar mendapatkan hasil kesimpulan terhadap identifikasi hoax yang dilakukan sesuai dengan *flowchart* berikut:



Gambar 4. Alur Identifikasi Hoax

HASIL

Data training diambil dari *website* https://turnbackhoax.id/ sebanyak 70 data yang berguna untuk menghitung nilai *Entropy* dan nilai *Gain* pada Algoritma C4.5. Berikut adalah sampel berita yang digunakan sebagai data *training*.

Tabel 1. Sampel Berita
https://turnbackhoax.id/

Kode Berita	Judul Berita	Url
1	Biar melek ahk yang mau bubarin fpi	https://turnbackhoax.id/2019/05/09/salah-biar-melek-ahk-yg-mau-bubarin-fpi/
2	Saksi PKS Pekanbaru Bernama Hatta Zalliyus Dirawat di Rumah Sakit Karena Keracunan Cyanida	https://turnbackhoax.id/2019/05/09/salah-saksi-pks-pekanbaru-bernama-hatta-zalliyus-dirawat-di-rumah-sakit-karena-keracunan-cyanida/
3	Ketua KPUD Bekasi Meninggal Dunia	https://turnbackhoax.id/2019/05/09/salah-ketua-kpud-bekasi-meninggal-dunia/
4	Viralkan Berulang-Ulang Karena Ini Masalah Nyawa	https://turnbackhoax.id/2019/05/09/salah-viralkan-berulang-ulang-karena-ini-masalah-nyawa/
5	Eks Kapolsek Cabut Pernyataan Soal Diarahkan Dukung Jokowi	https://turnbackhoax.id/2019/05/09/berita-eks-kapolsek-cabut-pernyataan-soal-diarahkan-dukung-jokowi/
6	Kapolri Nyatakan KPI Tidak Membahayakan Negara	https://turnbackhoax.id/2019/05/09/salah-kapolri-nyatakan-pki-tidak membahayakan-negara/
7	Buntut Panjang Penabrakan Kapal Perang RI di Selat Natuna	https://turnbackhoax.id/2019/05/08/berita-buntut-panjang-penabrakan-kapal-perang-ri-di-selat-natuna/
8	curang kok jangan diviralkan	https://turnbackhoax.id/2019/05/08/salah-curang-kok-jangan-diviralkan/
9	Klarifikasi dari Rektor UIN Suska Riau terkait Surat Pemecatan Ustaz Abdul Somad Sebagai Dosen	https://turnbackhoax.id/2019/05/08/benar-klarifikasi-dari-rektor-uin-suksa-riau-terkait-surat-pemecatan-ustaz-abdul-somad-sebagai-dosen/
10	Ulama di Banten Dibacok Simpatisan PKI	https://turnbackhoax.id/2019/05/08/salah-ulama-di-banten-dibacok-simpatisan-pki/

Berdasarkan data *training* pada tabel 1, maka dihasilkan keputusan berdasarkan atribut-atribut yang telah ditentukan seperti berikut:

Tabel 2. Hasil Keputusan

Kode Berita	A	S	P	Status
B1	No	No	Yes	Salah
B2	No	No	No	Salah
B3	No	No	Yes	Salah
B4	Yes	No	Yes	Salah
B5	No	Yes	No	Benar
B6	No	No	No	Salah
B7	No	Yes	No	Benar
B8	Yes	No	Yes	Salah
B9	No	Yes	Yes	Benar
B10	Yes	Yes	No	Salah

Selanjutnya akan diteruskan dengan menghitung jumlah *Entropy* dan *Gain*:

E(S) $=\Sigma_{1=1}^{c} - p_i \log_2 p_i$,

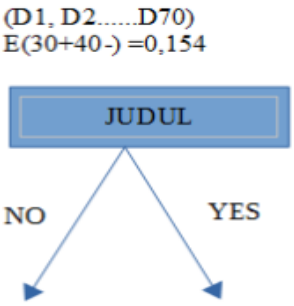
= E ([30+40-])

$$= [(-{}^{30}/_{70} \log_2 {}^{30}/_{70} +$$

$$-{}^{40}/_{70} \log_2 {}^{40}/_{70})]$$

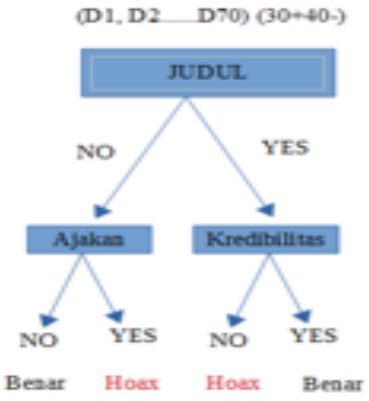
$$= 0, 154 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Berdasarkan perhitungan nilai entropy keseluruhan (3), maka didapat nilai *entropy* dari masing-masing atribut yaitu Ajakan (A) = 0,078, Kredibilitas (S)= 0,025 dan Judul (P) = 0,112. Karena P mempunyai nilai *entropy* tertinggi, maka P merupakan pohon keputusan pertama seperti terlihat pada gambar 5.



Gambar 5. Pohon Keputusan Pertama

Untuk penentuan *entropy* Judul (P) = Yes, didapat hasil 0,0032. Kemudian dilanjutkan dengan nilai *gain* dari Kredibilitas (S), didapatkan nilai S sebesar 0,082. Untuk penentuan *entropy* judul (P) = no, didapat hasil 0,049. Kemudian dilanjutkan dengan nilai *gain* dari ajakan (A), didapatkan nilai S sebesar 0,052. Berdasarkan nilai *entropy* dan *gain*, maka dihasilkan pohon keputusan akhir sebagai berikut:

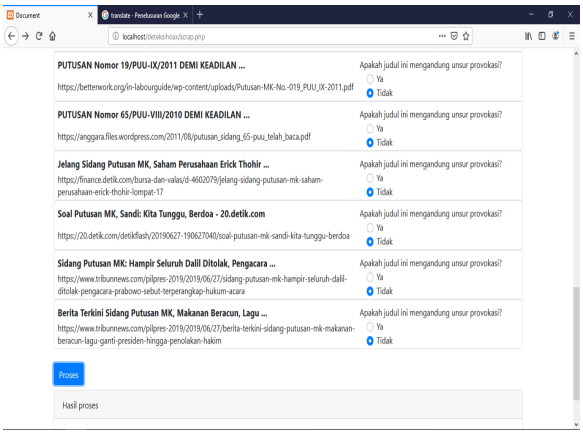


Gambar 6. Pohon Keputusan Akhir

Berdasarkan pohon keputusan akhir maka di dapatkan aturan sebagai berikut:

a. IF (judul = no AND ajakan = no) or (kredibilitas = yes) then Berita Benar
b. IF (judul = no AND ajakan = yes) or (kredibilitas = yes) then Berita Benar
c. IF (judul = no AND ajakan = yes) or (kredibilitas = no) then Berita Benar
d. IF (judul = no AND ajakan = no) or (kredibilitas = no) then Berita Hoax
e. IF (judul = yes AND ajakan = yes) or (kredibilitas = yes) then Berita Hoax
f. IF (judul = yes AND ajakan = no) or (kredibilitas = no) then Berita Hoax

Pohon keputusan dan rule yang telah disusun kemudian diimplementasikan pada sebuah sistem untuk mengetahui sejauh mana keberhasilan algoritma C4.5 dan teknik scraping dalam mengidentifikasi *hoax*. Pengujian dilakukan dengan memasukkan dua puluh kata kunci secara bergantian. Setelah proses *scraping*, dilanjutkan dengan pengisian informasi dari atribut judul dengan memilih "ya" atau "tidak".

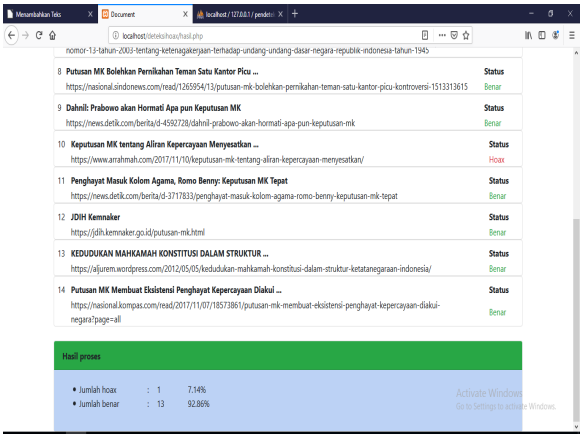


Gambar 7. Pengisian Informasi Atribut Judul

Selanjutnya, setelah melalui pemrosesan maka sistem akan memunculkan jumlah berita yang teridentifikasi sebagai *hoax* dan berita yang benar seperti pada gambar 8.

Gambar 8. Hasil Identifikasi Pada Salah Satu

Kata Kunci



Tabel 3. Komparasi Hasil Sistem dengan Fakta

No	keywords	Sistem	Fakta	Akurasi
1	Cuaca Indonesia	Benar	Benar	1
2	Kecelakaan Mobil	Benar	Benar	1
3	Putusan MK	Benar	Benar	1
4	DPR Korupsi	Benar	Benar	1
5	Sistem Zonasi Sekolah	Benar	Hoax	0
6	Jualan online	Benar	Benar	1
7	Saffron Palsu	Benar	Benar	1
8	Liga Champions	Benar	Benar	0
9	Bayar Pajak	Benar	Benar	1
10	Vpn Luar negri Gratis	Benar	Hoax	0
11	Erupsi Gunung Tangkuban Perahu	Hoax	Benar	0
12	Hilangnya Thoriq di gunung	Benar	Benar	1
13	Kecurangan Pemilu 2019	Benar	Benar	1
14	Penyerang Anggota TNI di Papua	Hoax	Hoax	1
15	Prabowo Gugat ke MK	Benar	Benar	1
16	Foto Syukuran Jokowi 2 Periode	Hoax	Hoax	1
17	Saksi PKS Keracunan Sianida	Hoax	Hoax	1
18	Mentri Kelautan Lepaskan Penyu	Benar	Benar	1
19	Kapal Tenggelam di Danau Toba	Benar	Benar	1
20	Rupiah Makin Menurun	Benar	Benar	1

Berdasarkan pengujian pada tabel 3 dengan menggunakan dua puluh kata kunci, didapatkan nilai akurasi sebesar 80%. Perhitungan akurasi berdasarkan Hasil Sistem dan Fakta, jika nilai Hasil Sistem dan Fakta itu adalah "Benar" dan "Benar" atau "Hoax" Dan "Hoax" maka nilai akurasinya adalah (1), dan jika Hasil Sistem dan Fakta bernilai "Hoax" dan "Benar" atau sebaliknya maka nilai akurasinya adalah (0). Beberapa *keywords* terdapat sedikit perbedaan hasil antara pengujian sistem dengan pengujian fakta karena pada pengujian sistem penentuan judul informasi masih dilakukan manual. Kemudian untuk nilai akurasi yang belum bisa mencapai angka persentase atas mendekati sempurna, dimungkinkan karena data training masih di perbarui secara manual.

KESIMPULAN

Sebagai sarana dalam mendapatkan informasi,

berita online menjadi salah satu pilihan yang saat ini bisa didapat dengan mudah. Banyaknya berita online juga tidak menutup kemungkinan hadirnya hoax yang dapat mempengaruhi pikiran pembacanya. Identifikasi hoax dapat dilakukan dengan bantuan algoritma klasifikasi seperti C4.5 dan teknik scraping. Atribut yang digunakan untuk mengidentifikasi hoax adalah judul provokasi, ajakan menyebarkan, dan kredibilitas sumber. Hasil pengujian menunjukkan bahwa pada penelitian ini, algoritma dan teknik yang digunakan mampu mendeteksi hoax sebesar 80% dengan skema penentuan judul secara manual.

DAFTAR PUSTAKA

[1] Z. Indra, N. Zamin, and J. Jaafar, "A clustering technique using single pass clustering algorithm for search engine," in *2014 4th World Congress on Information and Communication Technologies, WICT 2014*, 2014, pp. 182–187.

[2] R. Mustika, "Etika Berkomunikasi Di Media Online Dalam Menangkal Hoax," *Diakom J. Media dan Komun.*, vol. 1, no. 2, pp. 43–50, 2018.

[3] A. N. Desga, "Upaya Media Massa Online dalam Menghadapi Berita Hoax," *J. Kaji. Media*, vol. 2, no. 2, pp. 97–101, 2018.

[4] kominfo, "Kominfo Temukan 3.356 Hoaks, Terbanyak saat Pemilu 2019," *kominfo.go.id*, 2019. [Online]. Available: https://kominfo.go.id/content/detail/21876/kominfo-temukan-3356-hoaks-terbanyak-saat-pemilu-2019/0/berita_satker. [Accessed: 14-Apr-2021].

[5] A. Yuliani, "Ada 800.000 Situs Penyebar Hoax di Indonesia," *kominfo.go.id*, 2017. [Online]. Available: https://kominfo.go.id/content/detail/12008/ada-800000-situs-penyebar-hoax-di-indonesia/0/sorotan_media. [Accessed: 14-Apr-2021].

[6] A. Budiman, "Berita Bohong (Hoax) Di Media Sosial Dan Pembentukan Opini Publik," *Pusat Penelitian Badan Keahlian DPR RI*, vol. IX, no. 01, pp. 2009–2012, 2017.

[7] M. Iqbal, "Efektifitas Hukum dan Upaya Menangkal Hoax Sebagai Konsekuensi Negatif Perkembangan Interaksi Manusia," *Literasi Huk.*, vol. 3, no. 2, pp. 1–9, 2019.

[8] Z. Indra, J. Jaafar, N. Zamin, and Z. A. Bakar, "A language identifier for Indonesian and Malay text document," in *2015 International Symposium on Mathematical Sciences and Computing Research, iSMSC 2015 - Proceedings*, 2016, vol. 2015, pp. 127–131.

[9] J. Jaafar, Z. Indra, and N. Zamin, "A category classification algorithm for Indonesian and Malay news documents," *J. Teknol.*, vol. 78, no. 8–2, pp. 121–132, 2016.

[10] Z. Indra and L. Trisnawati, "Pengembangan Intelligent Data Collector Untuk Analisis Big Data Artikel Berita Online," *RABIT J. Teknol. dan Sist. Inf. Univrab*, vol. 3, no. 1, pp. 47–57, 2018.

[11] S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis, *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons, 2014.

[12] A. V Saurkar and S. A. Gode, "An Overview On Web Scraping Techniques And Tools," *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol. 4, no. 4, pp. 363–367, 2018.

[13] A. Cherfi, K. Nouira, and A. Ferchichi, "Very fast C4. 5 decision tree algorithm," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 119–137, 2018.

[14] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019.

[15] dev0928, "Getting started with web

scraping in Python," 2020. [Online]. Available: https://dev.to/dev0928/getting-started-with-web-scraping-in-python-1joi. [Accessed: 14-Apr-2021].

[16] C. A. Sugianto, "Analisis Komparasi Algoritma Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan," *Techno.Com*, vol. 14, no. 4, pp. 336–342, 2015.

[17] Sunaryono, "Penelitian Komparasi Algoritma Klasifikasi dalam Menentukan Website Palsu," *Teknikom*, vol. 1, no. 1, pp. 1–12, 2017.

[18] N. Frastian, S. Hendrian, and V. H. Valentino, "Komparasi Algoritma Klasifikasi Menentukan Kelulusan Mata Kuliah Pada Universitas," *Fakt. Exacta*, vol. 11, no. 1, p. 66, 2018.