Jupiter: Publikasi Ilmu Keteknikan Industri, Teknik Elektro dan Informatika
Vol.1, No.6 November 2023
e-ISSN: 3031-349X; p-ISSN: 3031-500X, Hal 24-35
DOI : https://doi.org/ 10.61132/jupiter.v1i6.169

Implementasi Ekstraksi Informasi Dari Abstrak Jurnal Sains Maritim Menggunakan Metode Ruled Based

Iwan Mahendro
Universitas Maritim AMNI Semarang
email : imahendro@gmail.com
Iman Mujiarto
Universitas Maritim AMNI Semarang

Abstract*. Universities certainly have a lot of important documents stored. Documents stored include scientific works in the form of journals. The journal that is kept has an identity so that if needed it will make it easier for people to find it. A library is a place to store scientific work, both in the form of journals and for storing other documents. The problem currently occurring is that there are so many documents or journals stored that it takes a long time for people who want to find information related to journals. The aim of this research is to provide a solution to make it easier for people who want to find information about many journals. The method proposed in this research is to use the information extraction method. Information extraction is the search for structured information such as entities and attributes. The stages carried out start from collecting the dataset which is then carried out preprocessing where this stage takes the form of converting documents with PDF extensions to documents with HTML extension. Apart from that, at this stage data cleaning is also carried out, meaning that at this stage it will be possible to create new paragraphs. Therefore, the new paragraph needs to be removed. Then the next stage is rule-based information extraction based on keywords and rules. The results of this research are that none of the 50 journal abstract documents failed so that the accuracy obtained was 100%. So the information extraction in this research can be used to search for information from journal abstracts.*
Keywords*: extraction, information, abstract, method, ruled based*

Abstraks. Perguruan tinggi tentunya mempunyai banyak sekali dokumen-dokumen penting yang disimpan. Dokumen yang disimpan antara lain karya ilmiah berupa jurnal. Jurnal yang disimpan mempunyai identitas sehingga bila diperlukan akan memudahkan orang untuk menemukannya. Perpustakaan merupakan tempat menyimpan karya ilmiah baik berupa jurnal maupun untuk menyimpan dokumen lainnya. Permasalahan yang terjadi saat ini adalah banyaknya dokumen atau jurnal yang disimpan sehingga masyarakat yang ingin mencari informasi terkait jurnal membutuhkan waktu yang lama. Tujuan dari penelitian ini adalah memberikan solusi untuk memudahkan masyarakat yang ingin mencari informasi tentang jurnal yang banyak. Metode yang diusulkan dalam penelitian ini adalah dengan menggunakan metode ekstraksi informasi. Ekstraksi informasi adalah pencarian informasi terstruktur seperti entitas dan atribut. Tahapan yang dilakukan dimulai dari pengumpulan dataset yang kemudian dilakukan preprocessing dimana tahap ini berupa konversi dari dokumen berekstensi pdf ke dokumen berekstensi html. Selain itu, pada tahap ini juga dilakukan pembersihan data, artinya pada tahap ini akan dimungkinkan adanya paragraf baru. Oleh karena itu, paragraf baru perlu dihilangkan. Kemudian tahap selanjutnya adalah ekstraksi informasi berbasis aturan berdasarkan kata kunci dan aturan. Hasil dari penelitian ini adalah dari 50 dokumen abstrak jurnal tidak ada satupun yang gagal sehingga keakuratan yang diperoleh adalah 100%. Sehingga ekstraksi informasi pada penelitian ini dapat digunakan untuk mencari informasi dari abstrak jurnal.
Kata kunci : ekstraksi**,** informasi, abstrak, metode, ruled based

Received Oktober 30, 2023; Revised November 15, 2023; Accepted November 30, 2023
** Iwan Mahendro, imahendro@gmail.com*

I. Pendahuluan

Saat ini banyak sekali dokumen yang beredar di masyarakat, dan dokumen yang beredar ini biasanya mempunyai format pdf maupun doc. Dengan semakin banyaknya dokumen yang sudah tersimpan, maka memerlukan waktu yang banyak pula untuk melakukan pencarian dokumen yang dikehendaki. Teknik pencarian informasi merupakan salah satu cara untuk membaca dokumen yang sangat banyak.

Suatu perguruan tinggi pastinya mempunyai perpustakaan tempat untuk menyimpan berbagai dokumen. Dokumen yang tersimpan harus diberikan identitas atau suatu informasi agar memudahkan seseorang untuk mencari dokumen yang diinginkannya. Cara memberikan identitas pada suatu dokumen yaitu seorang petugas perpustakaan atau pustakawan mengisi data yang diperlukan ke dalam suatu sistem. Cara ini tentunya akan merepotkan pustakawan apabila data yang dimasukkan sangat banyak. Selain merepotkan pustakawan, cara ini juga akan memungkinkan pustakawan melakukan salah input data ke dalam sistem dikarenakan kurang teliti ataupun karena kelelahan. Sekarang ini masalah yang terjadi karena salah input ataupun faktor kelelahan pustakawan dapat diatasi dengan cara pengisian data atau identitas dokumen secara otomatis. Salah satu cara mengisi identitas secara otomatis yaitu ekstraksi informasi. Ekstraksi informasi dapat diartikan sebagai suatu cara pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas dan atribut yang menggambarkan antar entitas dari sumber yang tidak terstruktur.[1].

Ekstraksi informasi merupakan bagian dari ilmu bahasa alami atau biasa dikenal dengan Natural Languange Processing. Pengekstraksian informasi dapat dilakukan dengan beberapa metode, salah satunya adalah dengan metode mengekstraksi adalah *rule based system*. System berbasis aturan (rule based system) adalah suatu program komputer yang memproses informasi dengan sekumpulan aturan yang terdapat di dalam basis pengetahuan untuk menghasilkan infomasi yang baru. Metode berbasis aturan dapat digunakan untuk dokumen yang terstuktur. Salah satu cara untuk mendapatkan informasi dari sebuah dokumen yaitu dengan adanya dokumen dari jurnal. Jurnal merupakan sebuah penulisan karya ilmiah yang sudah disusun secara sistematis berdasarkan aturan metode penelitian secara ilmiah. Penulisan jurnal merupakan suatu karya ilmiah dosen ataupun seseorang yang sudah melakukan penelitian sebelumnya. Struktur dari jurnal adalah judul jurnal, penulis, instansi, email, abstrak, dan kata kunci dari abstrak.

Penelitian ini menggunakan ekstraksi informasi dengan tujuan untuk mencari sebuah informasi dari sebuah abstrak. Informasi yang dicari disebut sebagai informasi target, informasi ini diambil dengan menggunakan sebuah teknik yang dapat untuk mendapatkan sebuah informasi dari teks yang tidak terstruktur. Untuk mendapatkan informasi yang spesifik maka akan digunakan machine learning dengan metode ekstraksi informasi berbasis klasifikasi. Hasil dari ekstraksi adalah kata, yang nantinya akan diklasifikasikan dalam kelas. Kelas ini akan mewakili setiap kategori dari informasi target. Kategori memerlukan karakteristik untuk mendapatkan kata. Karakteristik dalam penelitian ini adalah fitur dari setiap kata. Fitur dari kata antara lain jenis kata, kontekstual, morfologi, part of speech, dan name entity. Jika fitur sudah didapatkan maka diperlukan sebuah model yang akan diaplikasikan pada algoritma. Model ini digunakan untuk mengenali dari karakteristik, selanjutnya algoritma dapat memperhitungkan kemungkinan sebuah kata dengan karakteristik masuk dalam kategori target.

II. Teori

a. Ekstraksi informasi

Ekstraksi informasi bertujuan untuk mengidentifikasi bagian yang relevan [2]. Ekstraksi informasi adalah pencarian untuk informasi terstruktur seperti entitas, dan atribut. Ekstraksi informasi bertindak sebagai sistem yang akan mengenali data tidak terstruktur yang dimilikinya informasi yang belum dikategorikan dan tidak memiliki arti khusus. Tidak terstruktur mengandung arti bahwa informasi yang terkandung di dalamnya tidak bisa langsung diterjemahkan oleh komputer sehingga membutuhkan suatu sistem yang disebut dengan ekstraksi informasi. Ekstraksi informasi merupakan suatu pemrosesan bahasa alami yang berhubungan dengan temuan informasi seperti catatan database [3].

b. Metode Ruled based

Metode ruled based merupakan salah satu metode penerjemah bahasa alami. Ruled berarti aturan, dalam kehidupan keseharian kita banyak terdapat aturan. Tujuan dari aturan adalah untuk membatasi hal – hal yang dapat dilakukan oleh manusia. Di dalam penelitian, banyak yang sudah menggunakan ruled based untuk penelitian, misalnya penelitian berjudul Ekstraksi Nama Lokasi Dari Tweets Informasi Lalu Lintas [4]. Penelitian ini membahas tentang mengidentifikasi nama lokasi dari informasi lalu lintas melalui tweet.

c. Ekstraksi fitur

Ekstraksi fitur merupakan suatu proses pengambilan ciri objek yang menggambarkan karakteristik dari objek. Metode untuk ekstraksi fitur yaitu Principal Component Analysis (PCA) dan histogram. Ekstraksi fitur teks mengekstraksi informasi teks merupakan dasar dari pemrosesan teks dalam jumlah yang besar. Ada beberapa pendekatan untuk mengekstrak fitur yaitu pendekatan association rule mining, unsupervised pattern mining, mutual reinforcement approach, opinion lexicon, dan pattern knowledge [5].

III. Penelitian Terdahulu

Penelitian tentang ekstraksi informasi sebelumnya sudah pernah dilakukan yaitu oleh Khoirir Rosikin, Setio Basuki, dan Yusif Azhar [6]. Penelitian mereka tentang mengektrak informasi dari tweet berbahasa Indonesia berbasis klasifikasi. Mereka menggunakan 100 data uji dengan 7 fitur. Hasil dari pengujian yaitu nilai akurasi sebesar 74,07%.

Penelitian yang lain juga sudah pernah dilakukan oleh Agnieszka Konys di tahun 2018 [7]. Penelitian yang dilakukan yaitu ekstraksi informasi berbasis Ontologi (OBIE). Hasil dari penelitiannya yaitu Ontologi terbukti bisa menjadi suatu alat yang efisien untuk mengumpulkan data dan memberikan pengetahuan spesifikasi konseptual yang eksplisit.

Pada tahun 2019, juga sudah dilakukan penelitian tentang ekstraksi informasi yang dilakukan oleh Xia Xie et al [8]. Mereka melakukan ekstraksi infomasi dari konten web di China. Mereka juga menentukan beberapa metrik, struktur data, dan mengusulkan beberapa algoritma dalam data mining. Hasil dari penelitiannya adalah pendekatan dengan data mining dapat mengekstrak atribut secara akurat.

Pada tahun 2020 Donghon Ji et al melakukan penelitian tentang ekstraksi informasi. Tujuan dari penelitian mereka adalah untuk mengekstrak informasi bukti rekaman dokumen dari pengadilan [9]. Hasil dari percobaan yang telah mereka lakukan yaitu pada dataset menunjukkan keefektifan model yang diusulkan mendapatkan hasil 72,36%.

IV. Metode

A. Dataset

Dataset dalam penelitian ini berasal dari jurnal Sains dan Teknologi Maritim. Abstrak dalam jurnal ini terdapat 150 sampai 250 kata atau bahkan lebih. Dataset ini berjumlah 50 abstraksi.

B. Metode

Analisis kata kunci dan aturan

Penelitian ini menggunakan data jurnal yang berformat pdf. Untuk mengekstrak informasi akan difokuskan pada pencarian kata kunci dan aturan. Data jurnal yang akan dianalisis adalah abstrak Bahasa Indonesia dan abstrak Bahasa Inggris.

Gambar 1. Data Abstrak Jurnal

Hasil dari kata kunci dan aturan ekstraksi dapat dilihat pada tabel berikut ini.

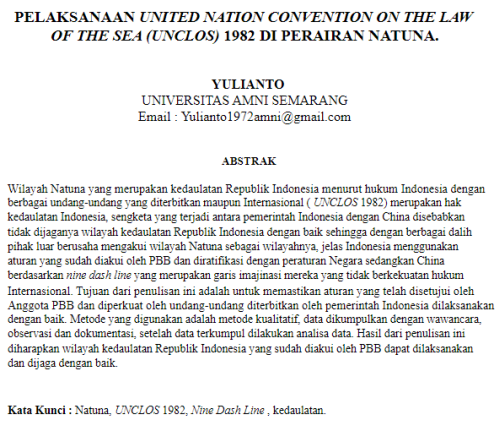
Tabel 1. Hasil Kata Kunci Dan Ekstraksi

Identifikasi	Kata kunci	Ket.
Judul jurnal	-	Mengambil judul jurnal

Nama penulis	oleh	Mencari kata oleh
Instansi	Perguruan tinggi	Mencari kata perguruan tinggi
Isi abstrak	-	Setelah menemukan kata kunci
Kata kunci	Kata kunci, keyword	Mencari kata kunci

Preprocessing

Hasil dari analisis data yang sudah dilakukan maka tahap selanjutnya adalah preprocessing yaitu melakukan konversi pdf ke html, pembersihan data abstrak jurnal. Konversi pdf ke html merupakan proses untuk mengubah data yang mempunyai format pdf menjadi format html. Hasil dari konversi pdf ke html dapat dilihat pada gambar berikut ini.



PELAKSANAAN *UNITED NATION CONVENTION ON THE LAW OF THE SEA (UNCLOS)* 1982 DI PERAIRAN NATUNA.

YULIANTO
UNIVERSITAS AMNI SEMARANG
Email : Yulianto1972amni@gmail.com

ABSTRAK

Wilayah Natuna yang merupakan kedaulatan Republik Indonesia menurut hukum Indonesia dengan berbagai undang-undang yang diterbitkan maupun Internasional (*UNCLOS* 1982) merupakan hak kedaulatan Indonesia, sengketa yang terjadi antara pemerintah Indonesia dengan China disebabkan tidak dijaganya wilayah kedaulatan Republik Indonesia dengan baik sehingga dengan berbagai dalih pihak luar berusaha mengakui wilayah Natuna sebagai wilayahnya, jelas Indonesia menggunakan aturan yang sudah diakui oleh PBB dan diratifikasi dengan peraturan Negara sedangkan China berdasarkan *nine dash line* yang merupakan garis imajinasi mereka yang tidak berkekuatan hukum Internasional. Tujuan dari penulisan ini adalah untuk memastikan aturan yang telah disetujui oleh Anggota PBB dan diperkuat oleh undang-undang diterbitkan oleh pemerintah Indonesia dilaksanakan dengan baik. Metode yang digunakan adalah metode kualitatif, data dikumpulkan dengan wawancara, observasi dan dokumentasi, setelah data terkumpul dilakukan analisa data. Hasil dari penulisan ini diharapkan wilayah kedaulatan Republik Indonesia yang sudah diakui oleh PBB dapat dilaksanakan dan dijaga dengan baik.

Kata Kunci : Natuna, *UNCLOS* 1982, *Nine Dash Line* , kedaulatan.

Gambar 2. Data Abstrak berupa pdf

```
PELAKSANAAN <span style="font-style:italic"> UNITED NATION CONVENTION ON THE LAW
</span></span>
</div>
OF THE SEA (UNCLOS)<span style="font-style:normal"> 1982 DI PERAIRAN NATUNA.</span></span>
</div>
YULIANTO</span>
</div>
UNIVERSITAS AMNI SEMARANG</span>
</div>
Email : Yulianto1972amni@gmail.com</span>
</div>
ABSTRAK</span>
</div>
Wilayah Natuna yang merupakan kedaulatan Republik Indonesia menurut hukum Indonesia dengan
</span>
berbagai undang-undang yang diterbitkan maupun Internasional (<span style="font-style:italic">
UNCLOS </span> 1982) merupakan hak </span>
</div>
kedaulatan Indonesia, sengketa yang terjadi antara pemerintah Indonesia dengan China disebabkan
</span>
</div>
<div class="pos" id="_157:418" style="top:418;left:157">
tidak dijaganya wilayah kedaulatan Republik Indonesia dengan baik sehingga dengan berbagai dalih
</span>
</div>
pihak luar berusaha mengakui wilayah Natuna sebagai wilayahnya, jelas Indonesia menggunakan
</span>
```

Gambar 3. Hasil Konversi Data Abstrak

Pembersihan data abstrak

Hasil dari konversi masih belum sesuai, ada paragrap baru muncul walaupun kata - kata masih dalam kalimat yang sama. Beberapa paragraph baru muncul, oleh karena itu perlu diadakan pembersihan data abstrak. Tahapan untuk melakukan pembersihan data abstrak yaitu : Jika ada 3 line kosong maka akan dijadikan 1 line kosong saja dan jika line tidak kosong maka line baru tersebut akan digabungkan dengan line sebelumya. Hasil dari pembersihan data abstrak terlihat seperti pada gambar berikut ini.

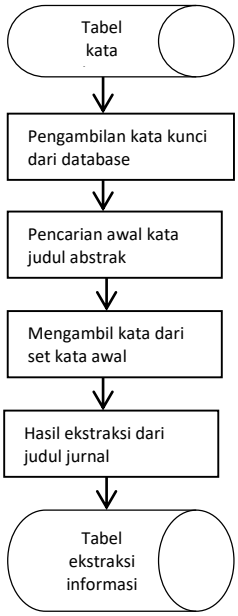
```
PELAKSANAAN UNITED NATION CONVENTION ON THE LAW </span></span>
</div>
OF THE SEA (UNCLOS)<span style="font-style:normal"> 1982 DI PERAIRAN NATUNA.</span></span>
</div>
YULIANTO</span>
</div>
UNIVERSITAS AMNI SEMARANG</span>
</div>
Email : Yulianto1972amni@gmail.com</span>
</div>
ABSTRAK</span>
</div>
Wilayah Natuna yang merupakan kedaulatan Republik Indonesia menurut hukum Indonesia dengan
berbagai undang-undang yang diterbitkan maupun Internasional merupakan hak
kedaulatan Indonesia, sengketa yang terjadi antara pemerintah Indonesia dengan China disebabkan
tidak dijaganya wilayah kedaulatan Republik Indonesia dengan baik sehingga dengan berbagai dalih
pihak luar berusaha mengakui wilayah Natuna sebagai wilayahnya, jelas Indonesia menggunakan
aturan yang sudah diakui oleh PBB dan diratifikasi dengan peraturan Negara sedangkan China
berdasarkan <span style="font-style:italic"> nine dash line</span> yang merupakan garis imajinasi
mereka yang tidak berkekuatan hukum Internasional. Tujuan dari penulisan ini adalah untuk
memastikan aturan yang telah disetujui oleh Anggota PBB dan diperkuat oleh undang-undang
diterbitkan oleh pemerintah Indonesia dilaksanakan
dengan baik. Metode yang digunakan adalah metode kualitatif, data dikumpulkan dengan wawancara,
```

Gambar 4. Hasil dari Pembersihan Data Abstrak

Ekstraksi Informasi

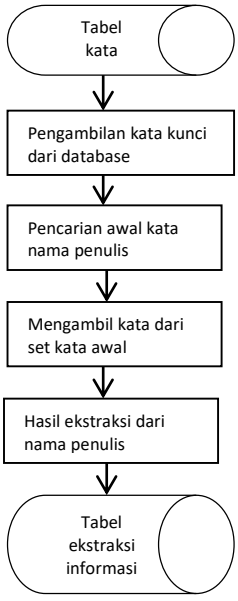
Pada penelitian ini, ekstraksi informasi akan dilakukan dengan rule based berdasarkan kata kunci dan aturan yang sudah dianalisis sebelumnya.

a. Ekstraksi judul jurnal



Gambar 5. Ekstraksi Judul Jurnal

b. Ekstraksi nama penulis



Gambar 6. Ekstraksi Nama Penulis

c. Ekstraksi instansi

Gambar 7. Ekstraksi Instansi

d. Ekstraksi isi abstrak

Gambar 8. Ekstraksi Isi Abstrak

e. Ekstraksi kata kunci

Gambar 9. Ekstraksi Kata Kunci

V. Hasil

Hasil dari penelitian ini didapatkan dari pengujian ekstraksi. Pengujian ekstraksi adalah tahap dengan tujuan untuk mengetahui performa dari metode seleksi fitur. Hasil dari ekstraksi informasi data abstrak jurnal dapat dilihat pada gambar berikut ini.

Tabel 2. Hasil Pengujian Sistem Data Jurnal Abstrak

No.	Pengujian data abstrak jurnal	Perguruan tinggi	Keterangan	Kesimpulan
1	Evyana Diah Kusumawati	Politeknik Bumi Akpelni Semarang	Ekstraksi sesuai	Ekstraksi berhasil
2	Subiyanto	Universitas Maritim Amni Semarang	Ekstraksi sesuai	Ekstraksi berhasil
3	Supriyanto	Universitas Maritim Amni Semarang	Ekstraksi sesuai	Ekstraksi berhasil
4	Yulianto	Universitas Maritim Amni Semarang	Ekstraksi sesuai	Ekstraksi berhasil
5	Kuncowati	Universitas Hang Tuah	Ekstraksi sesuai	Ekstraksi berhasil

Pada Tabel 2 terlihat bahwa fitur nama penulis berhasil untuk dilakukan ekstraksi informasi begitupun juga untuk fitur perguruan tinggi, berhasil untuk dilakukan ekstraksi informasi.

VI. Kesimpulan

Kesimpulan dari penelitian yang sudah dilakukan adalah dari 50 dokumen jurnal abstrak tidak ada yang gagal sehingga akurasi yang didapat yaitu 100%. Jadi ekstraksi informasi pada penelitian ini dapat digunakan untuk mencari sebuah informasi dari sebuah abstrak jurnal.

Daftar Pustaka

C. C. Aggarwal dan C. Zhai, Ed., *Mining Text Data*. Boston, MA: Springer US, 2012.

S. Sarawagi, *Information extraction*. Boston: Now, 2007.

J. Piskorski dan R. Yangarber, "Information Extraction: Past, Present and Future," dalam *Multi-source, Multilingual Information Extraction and Summarization*, T. Poibeau, H. Saggion, J. Piskorski, dan R. Yangarber, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, hlm. 23–49.

L. I. Tan, W. S. Phang, K. O. Chin, dan A. Patricia, "Rule-Based Sentiment Analysis for Financial News," dalam *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Kowloon, Okt 2015, hlm. 1601–1606, doi: 10.1109/SMC.2015.283.

S. S. Htay dan K. T. Lynn, "Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews," *Sci. World J.*, vol. 2013, hlm. 1–5, 2013, doi: 10.1155/2013/394758.

K. Rosikin, S. Basuki, dan Y. Azhar, "Ekstraksi Informasi Kesehatan Masyarakat Dari Tweet Berbahasa Indonesia Berbasis Klasifikasi Dengan Algoritma Naive Bayes," *J. Repos.*, vol. 2, no. 2, hlm. 193, Feb 2020, doi: 10.22219/repositor.v2i2.237.

A. Konys, "Towards Knowledge Handling in Ontology-Based Information Extraction Systems," *Procedia Comput. Sci.*, vol. 126, hlm. 2208–2218, 2018, doi: 10.1016/j. procs. 2018.07.228.

X. Xie, Y. Fu, H. Jin, Y. Zhao, dan W. Cao, "A novel text mining approach for scholar information extraction from web content in Chinese," *Future Gener. Comput. Syst.*, vol. 111, hlm. 859–872, Okt 2020, doi: 10.1016/j.future.2019.08.033.

D. Ji, P. Tao, H. Fei, dan Y. Ren, "An end-to-end joint model for evidence information extraction from court record document," *Inf. Process. Manag.*, vol. 57, no. 6, hlm. 102305, Nov 2020, doi: 10.1016/j.ipm.2020.102305.