Jurnal Politeknik Caltex Riau
Terbit Online pada laman https://jurnal.pcr.ac.id/index.php/jkt/
| e- ISSN : 2460-5255 (Online) | p- ISSN : 2443-4159 (Print) |

Klasifikasi Topik Skripsi Berdasarkan Makna dengan Pendekatan Semantik Web

Aditya Pradana[1] dan Randy Ridwansyah[2]

[1]Universitas Padjadjaran, Program Studi Teknik Informatika, email: aditya.pradana@unpad.ac.id
[2] Universitas Padjadjaran, Program Studi Sastra Inggris, email: randy.ridwansyah@unpad.ac.id

[1] Abstrak

Semantik web merupakan teknologi yang mampu memahami konteks seperti halnya manusia memahami konteks tersebut. Teknologi semantik dapat dimanfaatkan dalam berbagai bidang, salah satunya adalah klasifikasi topik skripsi mahasiswa. Banyak skripsi membahas topik serupa, tapi dengan peristilahan berbeda. Hal tersebut seringkali menimbulkan kesulitan dalam menentukan apakah suatu topik sudah terlalu sering dibahas. Oleh sebab itu, penelitian ini bertujuan untuk mengklasifikasikan dan memetakan bidang minat skripsi mahasiswa Program Studi S1 Teknik Informatika (TI) Unpad. Langkah-langkah yang dilakukan dalam penelitian ini adalah mengumpulkan data, menentukan entitas, implementasi ontologi, dan klasifikasi data menjadi beberapa kelas berdasarkan bidang minat di Program Studi TI Unpad (Sistem Informasi dan Multimedia, Kecerdasan Buatan dan Robotika, Jaringan Komputer, dan Metode Numerik). Langkah akhirnya adalah proses *training* dan *testing* data menggunakan algoritma k-*nearest neighbour* (KNN). Klasifikasi aktual data menunjukkan bahwa mayoritas skripsi di Prodi TI Unpad adalah tentang Sistem Informasi dan Multimedia, dengan presentase 50.23%. Sedangkan, klasifikasi prediktif data dengan algoritma KNN menunjukkan persentase 47.88%. Hasil yang berbeda tersebut diperoleh karena *Confusion Matrix* menunjukkan nilai sebagai berikut: AUC (0.711), CA (0.545), F1(0.578), *Precision* (0.669), *Recall* (0.545).

Kata kunci: semantik web, *k-nearest neighbour, confusion matrix*, skripsi

[2] Abstract

The Semantic web is a technology that can understand contexts as humans do. Many fields can take advantage of the technology in solving the problems it faces, classifying student thesis topics. Many theses discuss similar topics but with different terminology. This often creates difficulties in determining whether a topic has been discussed too often. Therefore, this study aims to classify and map students' undergraduate thesis areas in Informatics Engineering (TI) Unpad. The steps taken in this study were collecting data, determining entities, implementing ontology, and classifying data into several classes based on areas of interest in the Unpad IT Study Program (Information and Multimedia Systems, Artificial Intelligence and Robotics, Computer Networks, and Numerical Methods). The final steps were training and testing data using the k-nearest neighbor (KNN) algorithm. The actual classification of the data shows that the majority of theses in the IT Study Program Unpad are about Information Systems and Multimedia, with a percentage of 50.23%. Meanwhile, the predictive classification by using the KNN algorithm shows a

Dokumen diterima pada 29 Maret, 2021
Dipublikasikan pada 31 Mei, 2021

percentage of 47.88%. This difference resulted from the Confusion Matrix showed the following values: AUC (0.711), CA (0.545), F1 (0.578), Precision (0.669), Recall (0.545).

Keywords: *web semantic, k-nearest neighbour, confusion matrix, theses*

1. Pendahuluan

Skripsi merupakan salah satu syarat wajib bagi mahasiswa dalam menempuh gelar sarjana. Banyak skripsi dengan topik kajian serupa baik dalam hal teori dan algoritma yang digunakan. Bahkan beberapa data yang digunakan sama dan hasilnya menjadi mirip karena tujuan penelitiannya yang sama. Hal tersebut menjadi perhatian dalam menentukan topik yang akan diajukan oleh mahasiswa. Semakin banyak jumlah lulusan, maka data topik skripsi pun menjadi semakin banyak, yang akhirnya, jika tidak dibuat sistem yang baik, akan sulit diarsipkan. Banyak skripsi yang mengangkat topik yang serupa, tapi menggunakan peristilahan yang berbeda, sehingga akan sulit ketika menentukan apakah topik tersebut sudah pernah dibahas atau belum. Kesulitan lain adalah ketika diperlukan referensi dari penelitian yang termasuk dalam topik yang sama, namun kata kunci yang dituliskan berbeda, maka akan memerlukan waktu tambahan untuk mencarinya. Selain itu, algoritma dan bahasa pemrograman yang sering digunakan oleh mahasiswa dalam penyusunan skripsi pun terkadang belum terdata dengan baik. Padahal, hal tersebut dapat dijadikan sebagai masukan dalam penyusunan kurikulum dan pemetaan peminatan. Beberapa permasalahan tersebut dapat diatasi dengan penggunaan teknologi semantik.

Semantik web merupakan teknologi yang mampu memahami konteks seperti halnya manusia memahami konteks tersebut. Teknologi awal dari website berupa interaksi yang satu arah dan statis. Pengguna website hanya mendapatkan informasi saja, tanpa bisa berinteraksi dengan web tersebut. Teknologi tersebut berkembang sehingga pengguna dapat berinteraksi dengan lebih baik. Teknologi semantik bisa lebih mengerti apa yang diinginkan oleh manusia.

Penelitian mengenai semantik dalam bidang pendidikan sudah banyak yang dilakukan. Diantaranya penelitian mengenai penggunaan teknologi semantik web dalam konteks pendidikan formal yang dilakukan oleh Jensen [1]. Penelitian tersebut mengidentifikasi beberapa tema dalam dari teknologi semantik web untuk digunakan pada pendidikan formal. Tema yang diidentifikasi oleh tinjauan tersebut adalah Ontologi Semantic Web; Distribusi yang efisien, aksesibilitas, pengambilan, penggunaan kembali dan kombinasi sumber daya pendidikan; Link Data; Semantic Web untuk meningkatkan lingkungan belajar virtual dan personalisasi lingkungan belajar; Objek pembelajaran Semantic Web; Evaluasi, umpan balik dan penilaian; Layanan Web Semantik; serta alat pedagogis untuk guru dan siswa.

Penelitian selanjutnya adalah mengenai implementasi kerangka kerja ekosistem *e-learning* cerdas menggunakan ontology dan SWRL yang dilakukan oleh Ouf [2]. Penelitian ini menggabungkan konsep personalisasi ke dalam ekosistem *e-learning*. Ontologi semantik web berdasarkan personalisasi pada lingkungan pembelajaran memainkan peran penting dalam membangun ekosistem *e-learning* yang cerdas. Sehingga tercipta empat ontology yang terpisah untuk paket pembelajaran lengkap berdasarkan personalisasi yang terdiri dari model pembelajar dan semua komponen proses pembelajaran (objek pembelajaran, kegiatan pembelajaran, dan metode pengajaran).

Semantik web juga dapat dimanfaatkan dalam klasifikasi metadata sebagai pencegahan pembajakan informasi. Purnamasari dkk [3] melakukan penelitian menggunakan teknologi semantik web untuk menghindari terjadinya pembajakan perangkat lunak yang masih banyak terjadi. Semantik web digunakan untuk memeriksa kesamaan informasi yang ada dalam berbagai perangkat lunak berlisensi. Metode yang dilakukan adalah dengan cara mengambil informasi dari perangkat lunak sumber dan dicocokkan dengan perangkat lunak hasil pembajakan melalui

metadata-nya. Diperlukan tangkat keamanan tertentu dalam metadata agar *raw data* dari *resource* tersebut tidak mudah diubah.

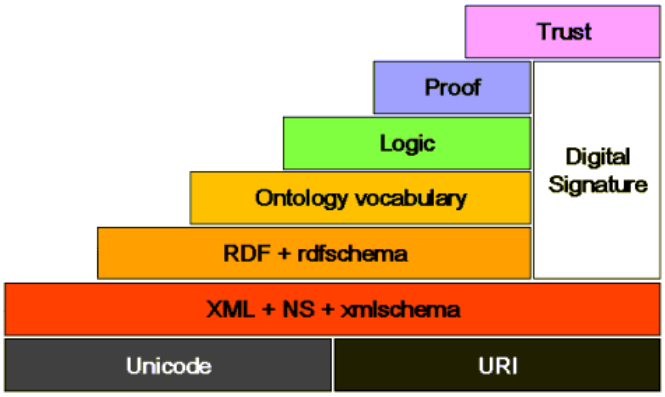
Bahkan penelitian mengenai semantik web juga dilakukan di bidang kesehatan. Penggunaan teknologi semantik web dapat digunakan untuk membantu penemuan obat baru. Penelitian yang dilakukan oleh Kanza dan Frey [4] menggunakan teknologi semantik untuk merepresentasikan data secara format, terstruktur, dapat dioperasikan, dapat dibandingkan, dan untuk menemukan hubungan yang belum ditemukan antar data obat (misalnya mengidentifikasi target obat baru atau senyawa yang relevan atau hubungan antara obat dan penyakit tertentu).

Berdasarkan pemaparan di atas, penelitian ini bertujuan untuk mengklasifikasikan dan memetakan bidang minat skripsi mahasiswa Program Studi Teknik Informatika Universitas Padjadjaran menggunakan pendekatan semantik web. Diawali dari pengambilan data, *preprocessing* data, melakukan klasifikasi berdasarkan bidang minat, kemudian melakukan *training* (melatih data) dan *testing* (menguji data) menggunakan algoritma *k-nearst neighbour*.

2. Tinjauan Pustaka

2.1 Semantik Web

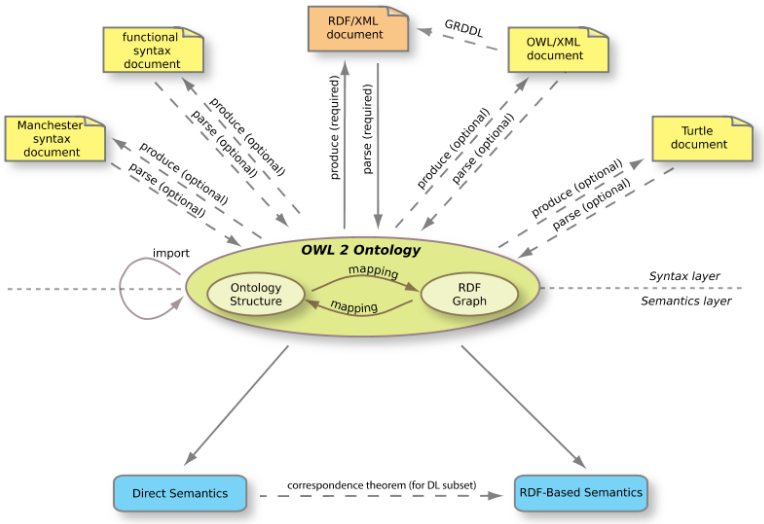
Definisi mengenai semantik web dapat ditemukan dalam artikel *Scientific American* pada bulan Mei 2001 yang berjudul "*The Semantic Web*" oleh Berners-Lee et al. [5] yang mengatakan "Web Semantik adalah perpanjangan dari web saat ini di mana informasi diberikan makna yang terdefinisi dengan baik, memungkinkan komputer dan orang bekerja sama dengan lebih baik. ". Terdapat beberapa layer dalam membangun sebuah semantik web, seperti yang ditunjukkan pada Gambar 1 berikut.



Gambar 1. Semantik web tower

2.2 Ontology Web Language

Ontologi merupakan kosakata formal yang disepakati dari istilah yang menjelaskan hubungannya dengan istilah lain, yang mencakup *domain* tertentu dan digunakan secara bersama oleh pengguna. *Ontology Web Language* versi 2 merupakan perkembangan dari OWL versi 1 yang dikembangkan oleh W3C *Web Ontology Working Group* pada tahun 2004. OWL dirancang dengan tujuan untuk membuat konten web dapat diakses oleh mesin [7].



Gambar 2. Struktur OWL 2

Gambar 2 memberikan gambaran umum dari OWL 2, di mana terdapat blok bangunan utama dan antar blok saling berhubungan satu sama lain. Elips di tengah mewakili gagasan abstrak ontologi, yang dapat dianggap sebagai struktur abstrak atau sebagai grafik RDF. Di bagian atas terdapat berbagai sintaks konkret yang dapat digunakan untuk bertukar ontologi. Di bagian bawah adalah dua spesifikasi semantik yang mendefinisikan arti ontologi OWL 2. Sebagian besar pengguna OWL 2 hanya membutuhkan satu sintaks dan satu semantik. Sehingga, diagram tersebut dapat jauh lebih sederhana, dengan hanya satu sintaks di atas, satu semantik di bawah, dan jarang perlu melihat apa yang ada di dalam elips di tengah.

Terdapat "ketidakpastian" dalam menentukan "rule" dalam menghubungkan antar entitas. Ketidakpastian tersebut mencakup berbagai aspek pengetahuan yang tidak sempurna, tidak lengkap, tidak jelas, dll. Beberapa pendekatan untuk menalar ketidakpastian tersebut diantaranya adalah teori probabilitas, logika fuzzy, k-*Nearest Neighbour*, dan banyak metode lainnya.

2.3 K-Nearest Neighbour (KNN)

Algoritma KNN merupakan algoritma yang digunakan untuk melakukan klasifikasi terhadap sebuah objek berdasarkan jumlah k buah data latih yang jaraknya paling dekat dengan objek tersebut. Jarak dekat/jauh dari data latih dengan objek yang akan diklasifikasikan dihitung menggunakan metode *cosine similarity* [10]. *Cosine Similarity* merupakan salah satu metode yang dapat digunakan untuk melihat sejauh mana kemiripan isi antar dokumen. Rumus yang digunakan umtuk menghitung jarak menggunakan *cosine similarity* pada algoritma KNN adalah:

$$\cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}} \quad \text{.....(1)}$$

Dimana:
$\cos(\theta_{QD})$ = kemiripan Q terhadap dokumen D
Q = *Training data*
D = *Testing data*
N = Banyaknya data

Untuk mendapatkan *value* fitur *training data* dan *testing data*, metode n-gram dapat digunakan. Model probabilistik n-gram, merupakan model yang digunakan untuk memprediksi kata berikutnya yang mungkin dari kata N-1 sebelumnya. Model statistika dari urutan kata ini

seringkali disebut juga sebagai model bahasa (*language models* / LMs). Model estimasi seperti n-gram memberikan probabilitas kemungkinan pada kata berikutnya yang mungkin dapat digunakan untuk melakukan kemungkinan penggabungan pada keseluruhan kalimat.

2.4 Confusion Matrix

Untuk menghitung tingkat akurasi dengan kasus kelas lebih dari dua (*multiple classifier*) dapat menggunakan *confusion matrix*. *Confusion matrix* memvisualisasikan keakuratan pengklasifikasian dengan membandingkan kelas aktual dengan kelas yang diprediksi [11].

Tabel 1. *Confusion Matrix* Perbandingan *Actual* dan *Predicted*

		Predicted	
		FALSE	TRUE
Actual	FALSE	True Negatif (TN)	False Positive (FP)
	TRUE	False Negatif (FN)	True Positive (TP)

Perhitungan akurasi dengan menggunakan *confusion matrix* dapat dilihat pada persamaan berikut

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(2)}$$

$$Precission = \frac{TP}{TP+FP} \text{.....(3)}$$

$$Recall = \frac{TP}{TP+FN} \text{.....(4)}$$

$$F1 = 2 * \frac{Recall*Precission}{Recall+Precisson} \text{.....(5)}$$

Dimana:
TP = Jika diprediksi positif, hasilnya adalah benar
TN = Jika diprediksi negatif, hasilnya adalah benar
FP = Jika diprediksi positif, hasilnya adalah salah
FN = Jika diprediksi negatif, hasilnya adalah salah

3. Metode Penelitian

Dalam mengerjakan penelitian ini, penulis melakukan beberapa langkah, yaitu pengumpulan data, menentukan entitas, implementasi ontologi, kemudian data yang didapat dikategorikan menjadi beberapa kelas sesuai dengan bidang minat yang ada pada Program Studi Teknik Informatika Unpad (Sistem Informasi dan Multimedia, Kecerdasan Buatan dan Robotika, Jaringan Komputer, dan Metode Numerik), lalu dilakukan proses *training* dan pengujian menggunakan algoritma k-*nearest neighbour*.

3.1 Pengumpulan Data

Pengumpulan data dilakukan melalui website https://informatika.unpad.ac.id/new/lulusan. Data terdiri dari nama mahasiswa, tanggal sidang, NPM, judul skripsi, pembimbing, dan penguji sidang. Data abstrak skripsi diperoleh dari perpustakaan Universitas Padjadjaran serta dari arsip skripsi mahasiswa Program Studi Teknik Informatika Unpad.

3.2 Membuat Entitas

Entitas dibuat berdasarkan data yang diperoleh. Entitas dapat berupa apapun, baik itu nama, tempat, waktu, dan lain-lain. Data yang diperoleh tersebut kemudian dikelompokkan menjadi beberapa entitas, yaitu nama dosen, kode dosen, nama mahasiswa, npm mahasiswa, judul skripsi, tanggal skripsi, topik skripsi, serta beberapa kata kunci mengenai topik skripsi tersebut. Entitas penting yang digunakan untuk menentukan klasifikasi pada penelitian ini difokuskan pada kata kunci dan kata-kata yang terkandung pada abstrak.

3.3 Implementasi Ontologi

Salah satu metode ontologi yang sering digunakan adalah *Ontology Development* 101. Konsep dari metode tersebut adalah membangun ontologi dari awal (*from stratch*) [12]. Pada prakteknya, dilakukan proses iterasi secara berkelanjutan untuk memperbaiki hasil ontologi. Proses yang dilakukan diantaranya adalah *determine scope, consider reuse, enumerate terms, define classes, define properties, define constraints,* dan *create instance.*

Tahap selanjutnya adalah pembelajaran ontologi menggunakan *ontology learning from text*, dimana prosesnya akan mengidentifikasi istilah, konsep, hubungan, serta informasi secara tekstual dan menggunakannya dalam membangun ontologi. Dalam proses otomatisasi, dapat menggunakan *Natural Language Processing* (Pemrosesan Bahasa Natural).

Metode KNN digunakan untuk melakukan klasifikasi yang diawali dengan membuat data *training* kemudian dimasukkan data *testing* untuk menguji metode yang digunakan. Hasilnya dapat dilihat menggunakan tabel *confusion matrix.*

4. Hasil dan Pembahasan

Untuk menentukan ontologi, data diambil dari judul dan abstrak skripsi, kemudian dilakukan proses *preprocessing*, di antaranya adalah mengubah penulisan semua kata menjadi huruf kecil (lowercase), menghilangkan tanda baca seperti koma, titik, serta titik koma menggunakan *regular expression*, serta menggunakan *stopword* untuk menghilangkan kata sambung dalam bahasa Indonesia.

Langkah selanjutnya adalah menentukan pemilihan data untuk digunakan dalam proses *data training*. Model n-gram merupakan model yang paling penting dalam pemrosesan suara atau bahasa, baik untuk memperkirakan probabilitas kata berikutnya maupun keseluruhan *sequence* (urutan kata). Untuk mendapatkan kombinasi kata dalam pengolahan dokumen tersebut, digunakan n-gram dengan nilai 3 (*value*=3). Hasil pengolahan data tersebut kemudian diberikan *value* yang dibagi menjadi 4 fitur, yaitu Sistem Informasi dan Multimedia, Kecerdasan Buatan (AI) dan Robotika, Jaringan Komputer, serta Metode Numerik. Hasil *preprocessing* dan n-gram dapat dilihat pada Gambar 4.

Langkah selanjutnya adalah menguji data baru dengan menggunakan hasil dari proses *data training* sebelumnya. *Testing data* terdiri dari fitur-fitur yang berisi kata yang terkandung dalam hasil pengolahan data sebelumnya.

Gambar 4. Hasil Preprocessing dan N-gram Abstrak

Dengan menggunakan data aktual, diperoleh bahwa banyak mahasiswa yang mengangkat topik mengenai Sistem Informasi dan Multimedia dalam skripsinya, dengan presentase 51,17%. Hasil lengkapnya disajikan pada Tabel 2 berikut.

Tabel 2. Presentase Data Aktual Topik Skripsi Mahasiswa

Topik	Jumlah	Presentase
Sistem Informasi dan Multimedia	107	0.502347
Kecerdasan Buatan & Robotika	75	0.352112
Jaringan Komputer	19	0.089201
Metode Numerik	12	0,056338

Setelah data aktual diperoleh, kemudian data yang sama diproses menggunakan algoritma k-*Nearest Neighbour* dengan nilai k=3. Hasil perbandingan data aktual dan data prediksi tersebut dapat dilihat pada Tabel 3 berikut.

Tabel 3. Perbandingan Data Aktual dan Data Prediksi Topik Skripsi Mahasiswa

		predicted				
		Jaringan Komputer	AI dan Robotika	Metode Numerik	Sistem Informasi dan Multimedia	∑
actual	Jaringan Komputer	4	8	1	6	**19**
	AI dan Robotika	1	54	0	20	**75**
	Metode Numerik	1	4	4	3	**12**
	Sistem Informasi dan Multimedia	5	29	0	73	**107**
	∑	**11**	**95**	**5**	**102**	**213**

Hasil dari *Confusion Matrix* menunjukkan nilai sebagai berikut: AUC (0.711), CA (0.545), F1(0.578), *Precision* (0.669), *Recall* (0.545).

5. Kesimpulan

Dari hasil penelitian yang dipaparkan di atas, ditemukan bahwa teknologi semantik web dapat membantu proses klasifikasi topik skripsi mahasiswa. Proses klasifikasi dilakukan secara aktual dan prediktif berdasarkan kata kunci dan makna yang terkandung di dalamnya. Hasil klasifikasi secara aktual menunjukkan bahwa skripsi mahasiswa TI Unpad didominasi oleh topik mengenai Sistem Informasi dan Multimedia, dengan presentase 50.23%. Sedangkan, klasifikasi prediktif dengan menggunakan algoritma KNN menghasilkan persentase 47.88%. Hasil yang berbeda tersebut diperoleh karena *Confusion Matrix* menunjukkan nilai sebagai berikut: AUC (0.711), CA (0.545), F1(0.578), *Precision* (0.669), *Recall* (0.545).

Temuan dari penelitian ini ke depannya dapat digunakan untuk membantu Program Studi TI Unpad dalam penyusunan dan evalusi kurikulum, juga pemetaan minat penelitian mahasiswa dan dosen. Pada pengembangan selanjutnya, data tersebut juga dapat digunakan untuk menyeleksi beberapa topik yang mungkin sudah banyak dibahas, agar mempunyai lebih banyak variasi dalam pengembangan dan eksplorasi mengenai teknologi terbaru.

Daftar Pustaka

[1] J. Jensen, "A systematic literature review of the use of Semantic Web technologies in formal education," *Br. J. Educ. Technol.*, vol. 50, no. 2, pp. 505–517, 2019, doi: https://doi.org/10.1111/bjet.12570.

[2] S. Ouf, M. A. Ellatif, S. E. Salama, and Y. K. Helmy, "A proposed paradigm for smart learning environment based on semantic web," *Comput. Hum. Behav.*, vol. 72, p. 79, 2017.

[3] D. Purnamasari, I. W. S. Wicaksana, A. R. Wijaya, H. Riesvicky, and W. Pratama, "PEMANFAATAN WEB SEMANTIK DALAM KLASIFIKASI METADATA SEBAGAI PENCEGAHAN PEMBAJAKAN INFORMASI," in *Prosiding Konferensi Nasional Sistem Informasi*, 2015, pp. 1126–1131.

[4] S. Kanza and J. G. Frey, "A new wave of innovation in Semantic web tools for drug discovery," *Expert Opin. Drug Discov.*, vol. 14, no. 5, pp. 433–444, 2019, doi: 10.1080/17460441.2019.1586880.

[5] W3C, "The Semantic Web Made Easy." https://www.w3.org/RDF/Metalog/docs/sw-easy.

[6] W3C, "Resource Description Framework." https://www.w3.org/RDF/.

[7] W3C, "OWL 2 Web Ontology Language Document Overview (Second Edition)." https://www.w3.org/TR/owl2-overview/.

[8] S. University, "Protege." https://protege.stanford.edu/.

[9] A. L. Asysyafii and S. Saidah, "Ontologi Web Obat Esensial Nasional Menggunakan Protégé 5.0," *J. Ilm. Teknol. dan …*, vol. 21, no. 3, pp. 196–205, 2017, [Online]. Available: http://ejournal.gunadarma.ac.id/index.php/tekno/article/view/1598.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. Waltham: Elsevier, 2012.

[11] S. Visa, B. Ramsay, A. Ralescu, and E. van der Knaap, "Confusion Matrix-based Feature Selection," in *Proceedings of the Twenty-second Midwest Artificial Intelligence and Cognitive Science Conference*, 2011, vol. 710, no. January, p. 8.

[12] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology." https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.