

LITERATURE REVIEW INFORMATION EXTRACTION PADA TEKS BIOMEDICAL DENGAN PENDEKATAN MACHINE LEARNING DAN NAMED ENTITY RECOGNITION

Rizki Ariyani

Ilmu Komputer dan Teknologi Informasi, rizkiariyani@staff.gunadarma.ac.id, Universitas Gunadarma

ABSTRAK

Extraction of information or information extraction is a field of science for natural language processing, by converting unstructured into information in a structured form. Information that is useful and requires data that can be used as a reference in making decisions. Machine Learning Algorithms have been used to build basic classification models to identify pediatrics in biomedical texts and have been developed to provide more accurate information. Named Entity Recognition (NER) is useful to help identify and identify a word. The biomedical field has a lot of literature so that NER is highly demanded in the biomedical domain. Biomedical texts are not homogeneous. Medical notes or records written differently from scientific articles, sequence annotations, or other recent public health. Various methods in biomedicine can be recognized, the meaning of which has been developed using machine learning. This study discusses previous research on NER in the biomedical sector.

Keywords: *Biomedical, Information Extraction, Machine Learning, Named Entity Recognition.*

Abstrak

Ekstraksi informasi atau *Information Extraction* merupakan suatu bidang ilmu untuk pengolahan bahasa alami, dengan cara mengubah teks tidak terstruktur menjadi informasi dalam bentuk terstruktur. Informasi yang berguna dan bermakna membutuhkan data yang dapat dijadikan acuan dalam pengambilan keputusan. Algoritma *Machine Learning* telah digunakan untuk membangun model klasifikasi biner dasar untuk mengidentifikasi teks pediatrik dalam teks biomedical dan semakin dikembangkan agar mampu memberikan informasi yang lebih akurat. *Named Entity Recognition* (NER) berguna untuk membantu mengidentifikasi dan mendeteksi entitas dari suatu kata. Bidang biomedis memiliki banyak pustaka sehingga NER sangat dituntut dalam domain biomedical. Teks biomedical bukan termasuk homogen. Catatan atau record medis ditulis berbeda dari artikel ilmiah, anotasi urutan, atau pedoman kesehatan masyarakat lainnya. Berbagai metode dalam biomedical untuk dapat dikenali maknanya sudah banyak dikembangkan dengan menggunakan *machine learning*. Penelitian ini membahas penelitian yang sudah dilakukan sebelumnya mengenai NER dalam bidang biomedical.

Kata Kunci: *Biomedical, Information Extraction, Machine Learning, Named Entity Recognition.*

1. PENDAHULUAN

Named Entity Recognition (NER) adalah komponen dari ekstraksi informasi yang berfungsi untuk mengenali entitas nama (nama orang, lokasi, organisasi), ekspresi waktu (tanggal, waktu, durasi) dan ekspresi bilangan (uang, persen, numerik, kardinal) pada kumpulan teks [1]. *Information extraction* atau ekstraksi informasi merupakan bagian dari *Natural Language Processing* (NLP). Sistem ekstraksi informasi adalah sebuah proses penemuan informasi dari kumpulan dokumen atau teks berbahasa alami sebagai masukannya dan menghasilkan informasi yang berguna berupa informasi yang terstruktur dengan format tertentu. Ekstraksi informasi yang utuh harus melewati lima tahapan yaitu *named entity recognition*, *coreference resolution*, *template element construction*, *template relation construction* dan *scenario template production* [2].

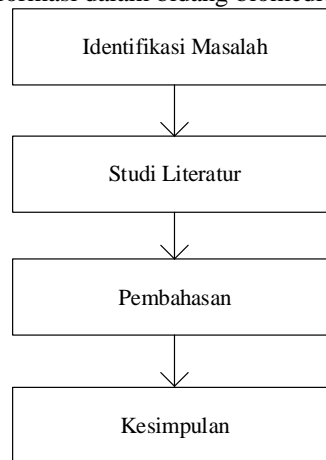
NER diimplementasikan dalam banyak bidang, antara lain dalam *machine translation*, *question-answering machine system*, *indexing* pada *information retrieval*, klasifikasi dan juga dalam *automatic summarization*. Beberapa pendekatan yang dipakai dalam NER antara lain *Rule Based*, *Machine Learning Based* yang memanfaatkan *Hidden Markov Model* [3], *Maximum Entropy* [4], *Decision Tree* [5], *Support*

Vector Machine [6], *Conditional Random Fields* [7] dan pendekatan *Hybrid* [8]. Tujuan yang diharapkan dari proses dalam NER adalah untuk melakukan ekstraksi dan klasifikasi nama ke dalam beberapa kategori dengan mengacu kepada makna yang tepat.

Metode *machine learning* dalam NER digunakan untuk melakukan klasifikasi dan menggunakan model klasifikasi statistik untuk mengenali *named-entity*. Pada metode ini, sistem mencari *patern/pola* dan hubungannya pada sebuah text untuk membuat model dengan pendekatan statistik dan algoritma *machine learning*. Sistem tersebut digunakan untuk mengidentifikasi dan mengklasifikasi kata benda ke dalam beberapa kelas seperti orang, lokasi, waktu [9]. Pendekatan *machine learning* dapat dikelompokkan ke dalam model *supervised* dan *unsupervised*. *Supervised learning* menggunakan pendekatan pembelajaran dengan menggunakan data yang sudah diberi label untuk menghasilkan *feature* dalam klasifikasi. Model ini akan menghasilkan *performance* yang bagus jika sistem di *training* dengan menggunakan data label yang berkualitas dan dalam jumlah data yang besar.

2. TINJAUAN PUSTAKA

Dalam penulisan ini diperlukan kerangka kerja dalam melaksanakan penelitian untuk melakukan *literature review* mengenai ekstraksi informasi dalam bidang biomedical seperti pada Gambar 1.



Gambar 1. Metode Penelitian

Beberapa tahun belakangan ini banyak peneliti yang membahas dan mengembangkan tentang teks biomedical. Pada penulisan ini akan dilakukan kajian ilmu dengan cara *literature review* terhadap tujuh penelitian yang telah dilakukan dengan *Named Entity Recognition* (NER) dalam teks biomedical. Berikut ini adalah acuan perbandingan yang digunakan dalam penulisan ini:

Tabel 1. Perbandingan Penelitian Sebelumnya (*Literature Review*)

No.	Penulis	Judul	Metode	Hasil
1.	Zhang, J., et al (2000) [10]	<i>Enhancing HMM-based biomedical named entity recognition by studying special phenomena</i>	<i>Hidden Markov Model (HMM)</i>	Hasil evaluasi pada GENIA V3.02 dan V1.1 menunjukkan bahwa sistem kami masing-masing mencapai 66,5 dan 62,5 <i>F-measure</i> , dan mengungguli sistem terbaik yang diterbitkan sebelumnya dengan 8,1 <i>F-measure</i> pada pengaturan eksperimental yang sama.
2.	Lee, Ki-Joong, et al (2004) [11]	<i>Biomedical named entity recognition using two-phase model based on SVMs</i>	<i>Two-phase SVM, identification phase, and classification phase</i>	NER telah menjadi salah satu tugas paling mendasar dalam akuisisi pengetahuan biomedis. Dalam penelitian ini, menyajikan dua fase SVM. Hasil percobaan pada GENIA corpus menunjukkan bahwa metode yang diusulkan efektif tidak hanya dalam mengurangi biaya komputasi

				tetapi juga dalam meningkatkan kinerja. Skor-F ($\beta = 1$) untuk identifikasi batas adalah 74.8 dan skor-F untuk klasifikasi semantic adalah 66,7.
3.	Finkel, J., et al (2005) [12]	<i>Exploring the boundaries: gene and protein identification in biomedical text</i>	<i>Maximum-entropy based</i>	Sistem ini dimasukkan dalam evaluasi komparatif BioCreative dan mencapai presisi 0,83 dan perolehan 0,84 dalam evaluasi "terbuka" dan presisi 0,78 dan perolehan 0,85 dalam evaluasi "tertutup".
4.	Goulart, R. R. V., et al (2011) [13]	<i>A systematic review of named entity recognition in biomedical texts</i>	<i>Kopora, NER, Machine Learning</i>	Hasil penelitian ini mengidentifikasi metode utama dalam NER biomedit, fitur dan metodologi untuk implementasi sistem NER. Selain dari kecenderungan yang teridentifikasi, beberapa kesenjangan terdeteksi yang mungkin merupakan peluang studi baru di area tersebut.
5.	Alshaikhdeeb, B., and Ahmad, K. (2016) [14]	<i>Biomedical Named Entity Recognition A Review</i>	<i>BNERR Features (Morphological, Lexical, Distance-based, Dictionary-based)</i>	<i>Biomedical Named Entity Recognition (BNER)</i> adalah tugas untuk mengidentifikasi contoh biomedis seperti senyawa kimia, gen, protein, virus, gangguan, DNA, dan RNA. Tantangan utama dibali BNER terletak pada metode yang akan digunakan mengekstraksi entitas tersebut. Sebagian besar metode yang digunakan untuk BNER mengandalkan <i>Supervised machine Learning (SML)</i> .
6.	Wulandari, DW., et al., (2018) [15]	<i>Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier</i>	<i>Rule Based dan Naïve Bayes Classifier (NBC)</i>	Didapatkan rata-rata precision, recall dan f-measure tertinggi pada rule based yaitu 0,85 dengan micro average. Dengan macro average recall dan f-measure tertinggi didapatkan kombinasi yaitu 0,66 dan 0,45, sedangkan precision tertinggi didapatkan <i>rule based</i> yaitu 0,39.
7.	Lee, W., et al., (2019) [16]	<i>Information Extraction from Biomedical Text Using Machine Learning</i>	<i>Machine Learning</i>	<i>Natural Language Processing</i> dan <i>Machine Learning</i> untuk mengimplementasikan sistem <i>Information Extraction</i> untuk teks biomedis menunjukkan hasil bahwa prototipe memiliki perolehan 0,93 dan presisi 0,86. Ini menunjukkan keefektifan algoritma pembelajaran mesin dalam mengklasifikasikan dan mengekstrak informasi yang relevan dari teks biomedis.

Penelitian pertama memperkenalkan peningkatan pengenalan entitas bernama biomedis berbasis *Hidden Markov Model (HMM)* dengan mempelajari berbagai fenomena khusus, seperti singkatan, entitas

bernama bertingkat, dll. Bertujuan untuk meningkatkan pengenalan entitas bernama berbasis HMM di domain biomedis. Pertama, menganalisis karakteristik entitas bernama biomedis. Kemudian, mengusulkan serangkaian fitur yang kaya, termasuk fitur pemicu ortografik, morfologis, *part-of-speech*, dan semantik. Semua fitur ini terintegrasi melalui *Hidden Markov Model* dengan pemodelan *back-off*. Selanjutnya, mengusulkan metode untuk pengenalan singkatan biomedis dan dua metode untuk pengenalan entitas berjenjang. Evaluasi pada GENIA V3.02 dan V1.1 menunjukkan bahwa sistem kami masing-masing mencapai 66,5 dan 62,5 *F-measure*, dan mengungguli sistem terbaik yang diterbitkan sebelumnya dengan 8,1 *F-measure* pada pengaturan eksperimental yang sama. Kontribusi utama dari makalah ini terletak pada kumpulan fitur yang kaya yang dirancang khusus untuk domain biomedis dan metode efektif untuk singkatan dan pengenalan entitas bernama berjenjang. Untuk pengetahuan terbaik kami, penelitian ini adalah yang pertama menangani fenomena berjenjang atau bertingkat dalam domain biomedis. Berdasarkan fitur dan metode yang kaya, sistem kami berhasil diadaptasi ke domain biomedis dan mencapai kinerja yang jauh lebih baik daripada sistem terbaik yang dipublikasikan sebelumnya. Batasan metode ini adalah bahwa beberapa konstruksi yang rumit, seperti " dan / atau, 'mungkin tidak dapat ditangani secara efektif. Dalam waktu dekat, eksplorasi lebih lanjut dapat dilakukan pada konstruksi rumit dalam dokumen biomedis. Salah satu cara yang mungkin adalah mengembangkan beberapa pola efektif untuk konstruksi konjungsi dan disjungsi. Selain itu, sumber daya dan database publik yang ada dapat diintegrasikan dalam sistem ini [10].

Penelitian kedua telah mengusulkan metode baru bernama biomedis dua fase (*two-phase biomedical*) berdasarkan SVM. Pada fase pertama, diidentifikasi batas-batas setiap entitas dengan satu klasifikasi SVM dan pasca-proses dengan pencarian kamus sederhana untuk memperbaiki kesalahan SVM. Pada fase kedua, diklasifikasikan entitas yang diidentifikasi ke dalam kelas semantiknya dengan menggunakan SVM hirarkis. Penelitian ini dengan membagi tugas menjadi dua sub-tugas, pengenalan, dan klasifikasi semantik, dapat memilih fitur yang lebih relevan untuk setiap tugas dan mengadopsi metode klasifikasi yang sesuai dengan tugas tersebut. Proses dua fase ini menghasilkan pengurangan biaya pelatihan SVM dan pengurangan masalah distribusi kelas yang tidak seimbang [11].

Fokus penelitian ketiga pada masalah khusus *Named Entity Recognition* (NER) yang membutuhkan identifikasi nama yang sesuai dengan kategori semantik dangkal. Mengharuskan peserta untuk mengidentifikasi entitas tunggal "NEWGENE" yang secara kasar sesuai dengan nama gen dan protein dalam abstrak medis. NER merupakan komponen penting untuk tugas ekstraksi informasi yang lebih kompleks seperti ekstraksi otomatis informasi interaksi protein-protein. Menyajikan sistem berdasarkan *maximum-entropy sequence* yang mencapai kinerja mutakhir dalam evaluasi komparatif BioCreative. Dalam implementasi sistem dengan *machine learning* menggunakan tagger urutan yang terlatih secara diskriminatif. Lalu menguraikan model *machine learning*, fase preprocessing, dan kemudian merinci set fitur lengkap. Kemudian memberikan detail implementasi dari prosedur pelatihan, dan terakhir menjelaskan penandaan dan fase pasca-pemrosesan yang bertujuan untuk meningkatkan deteksi batas. Penelitian ini mengevaluasi hasil BioCreative ini sebesar 83,2 dengan pencetak MUC dan mendapatkan skor 85,62. Hasil tersebut tetap jauh dibawah hasil terbaik yang dilaporkan tugas NER yaitu kisaran 87 hingga 89. Namun demikian, hasil yang lebih rendah sama-sama mencerminkan bahwa menemukan batas entitas yang benar dalam domain biomedis adalah tugas yang sangat sulit [12].

Penelitian selanjutnya mengemukakan bahwa kebanyakan studi ilmiah dalam NER Biomedical (92.1% diantara referensi yang dipelajari) sedang dikembangkan melalui karya eksperimental yang didasarkan pada teknik korpora dan pembelajaran mesin (*machine learning*). Pendekatan lain hanya mewakili 7,89% dari studi ini pada periode 2007-2009. Ini tidak berbeda dari penelitian sebelumnya yang telah disajikan [13].

Penelitian kelima ini mengemukakan telah memberikan tinjauan luas mengenai fitur-fitur BNER di mana taksonomi telah diidentifikasi berdasarkan pada representasi (yaitu nominal, numerik, dan Boolean). Selain itu, diskusi telah dilakukan untuk memberikan analisis kritis untuk setiap fitur. Fitur Boolean morfologis menunjukkan keunggulan dibandingkan dengan fitur lainnya. Membangun studi banding menggunakan fitur-fitur ini akan menjadi peluang besar untuk masa depan penelitian dalam hal mengidentifikasi kinerja untuk mengekstraksi entitas biomedis. Penelitian ini bertujuan untuk mengakomodasi studi tinjauan pada fitur yang dapat digunakan untuk BNER dimana berbagai jenis fitur akan diperiksa, termasuk fitur morfologi, fitur berbasis kamus, fitur leksikal, dan fitur berbasis jarak [14].

Dalam penelitian ini 19 dokumen latih diproses dan dianotasi manual untuk mencari Named Entity (NE) dan didapat 1135 data latih berbentuk kata. Dokumen uji ditokenisasi dan diberi POS Tag oleh tagger site terlebih dulu yang kemudian di cari *bigram* dan *trigram*. Selanjutnya proses *rule based*, jika dalam rule based tidak ditemukan solusi, maka akan masuk pada proses ekstraksi fitur dan NBC. Menggunakan 16 NE class, 18 aturan, dan 7 fitur dilakukan pengujian dengan tiga skenario yaitu pengujian *rule based*, NBC, dan

kombinasi keduanya. Didapatkan rata-rata precision, recall dan f-measure tertinggi pada rule based yaitu 0,85 dengan *micro average*. Dengan *macro average recall* dan *f-measure* tertinggi didapatkan kombinasi yaitu 0,66 dan 0,45, sedangkan precision tertinggi didapatkan rule based yaitu 0,39. Pada pendekatan *rule based* hasil kurang maksimal didapatkan karena tidak semua NE *class* memiliki aturan sehingga beberapa NE *class* tidak dikenali dan bahkan tidak didefinisikan sehingga menyebabkan NE *class* tersebut tidak memiliki anggota sama sekali. Pada metode *Naïve Bayes* hasil buruk didapatkan karena kurang dan tidak imbangnya data latih. Terdapat NE *class* yang memiliki anggota sampai ratusan namun ada NE *class* yang hanya memiliki anggota 1 saja. Beberapa kata juga masuk ke dalam NE *class* yang salah karena kata tersebut belum pernah muncul dalam data latih yang mana kebanyakan masuk ke dalam CELL COMPONENT karena NE *class* ini memiliki prior yang besar [15].

Penelitian ketujuh melakukan eksperimen pada data percobaan obat yang tidak memadai dan penggunaan obat yang tidak berlisensi dapat menyebabkan reaksi obat yang merugikan, terutama pada populasi anak. Penelitian ini bertujuan agar dokter dapat membuat keputusan yang tepat tentang keamanan dan efektivitas obat ini untuk anak-anak, ada kebutuhan untuk menganalisis label obat yang kompleks dan seringkali tidak terstruktur. Analisis eksplorasi label obat menggunakan *pipeline Natural Language Processing*. Kemudian algoritma *Machine Learning* digunakan untuk membangun model klasifikasi biner dasar untuk mengidentifikasi teks pediatrik dalam label obat tidak terstruktur. Dan serangkaian eksperimen telah dilakukan untuk mengevaluasi keakuratan model. Dalam penelitian ini telah digunakan *Natural Language Processing* dan *Machine Learning* untuk mengimplementasikan sistem *Information Extraction* untuk teks biomedis menggunakan Python. 500 label obat dari DailyMed diproses dan pengklasifikasi untuk mengidentifikasi dan mengekstrak teks dari bagian penggunaan pediatrik dari label obat dibangun. Peninjauan hasil menunjukkan bahwa prototipe memiliki perolehan 0,93 dan presisi 0,86. Ini menunjukkan keefektifan algoritma pembelajaran mesin dalam mengklasifikasikan dan mengekstrak informasi yang relevan dari teks biomedis. Hasil dari solusi ini menunjukkan bahwa ada kemungkinan penggunaan teknik *Machine Learning* untuk mengidentifikasi teks terkait pediatrik dari label obat tidak terstruktur. Karena klasifikasi yang dihasilkan dari solusi memiliki TP tinggi dan juga FP, pengklasifikasi sekunder yang dilatih hanya pada teks non-pediatrik dapat diterapkan. Ini akan membantu menyaring teks non-pediatrik dari hasil. Teks pediatrik yang diekstrak dari solusi ini dapat digunakan sebagai kumpulan data untuk mengidentifikasi apakah obat tersebut aman dan efektif pada populasi anak. Dalam solusinya, setiap kalimat dipetakan ke 1 tupel. Teknik jendela geser untuk mengelompokkan 2 atau lebih kalimat untuk membentuk tupel juga dapat digunakan. Selain itu, *Deep Learning* dengan vektor kata (*word2vec*) untuk menambah penemuan fitur otomatis untuk pemrosesan bahasa alami dapat diimplementasikan [16].

3. KESIMPULAN DAN SARAN

Dari beberapa penelitian yang telah dilakukan oleh para peneliti didapati hasil pengamatan dari berbagai metode berhasil dilakukan dalam *Information Extraction* pada Teks Biomedical dengan Pendekatan *Machine Learning* dan *Named Entity Recognition (NER)*. NER dan klasifikasi adalah tugas mengidentifikasi teks makna khusus dan mengklasifikasikan ke dalam beberapa kategori yang telah ditentukan. Beberapa kategori-kategori seperti orang, lokasi, organisasi, jumlah, dan lain sebagainya. Salah satu penelitian yang dikembangkan seperti yang dibahas pada kajian ilmu yaitu Teks Biomedical. Dalam metode NER yang menggunakan *Natural Language Processing (NLP)* juga dapat menggunakan metode *Conditional Random Field, Hidden Markov Model, Maximum Entropy, Support Vector Machine (SVM), Rule Based, Naïve Bayes Classifier (NBC), Machine Learning*, dan lain-lain. Hal ini menunjukkan bahwa studi dilakukan untuk mencari akurasi, presisi, nilai F1-Score untuk mencapai nilai optimal dalam ekstraksi informasi, sehingga dapat digunakan oleh pengguna dan penerimanya. Penelitian dimasa datang penelitian lebih mendalam tentang teks biomedical dengan menggunakan *Deep Learning* dibutuhkan agar mendapatkan hasil yang lebih sempurna dari penelitian sebelumnya.

DAFTAR PUSTAKA

- [1] Chincor, N., Brown, E., Ferro, L., and Robinson, P. "Named Entity Task Definition", *Version 1.4 The MITRE Corporation and SAIC*. 1999.
- [2] Chincor, N., Brown, E., Ferro, L., and Robinson, P. MUC-7 Information Extraction Task Definition, *The MITRE Corporation and SAIC*. 1998.
- [3] Daniel M. Bikel, Scott Miller, Richard Schwartz and Ralph Weischedel. "Nymble: a highperformance learning name-finder", *in the proceedings of the fifth conference on Applied natural language processing*, pages 194-201, San Francisco, CA, USA Morgan Kaufmann Publishers Inc. 1997.

- [4] Andrew Borthwick. "Maximum Entropy Approach to Named Entity Recognition" Ph.D. thesis, New York University. 1999.
- [5] Hideki Isozaki. "Japanese named entity recognition based on a simple rule generator and decision tree learning" in *the proceedings of the Association for Computational Linguistics*, pages 306-313. India. 2001.
- [6] Takeuchi K., and Collier N. "Use of Support Vector Machines in extended named entity recognition" in *the proceedings of the sixth Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan, China. 2002.
- [7] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" in *the proceedings of International Conference on Machine Learning*, pages 282-289, Williams College, Williamstown, MA, USA. 2001.
- [8] Kaur, D., and Gupta, V. "A Survey of Named Entity Recognition in English and othe Indian Language", *IJCSI International Journal of Computer Science Issues*, Vol. 7, Issue 6, ISSN (Online): 1694-0814. 2010.
- [9] Mansouri, L. S. Affendey, A Mamat. "Named Entity Recognition Approaches", *IJCSNS International Journal of Computer Science and Network Security*, Vol.8 No.2. 2008.
- [10] Zhang, J., Shen, D., Zhou, G., Su, J., and Tan, CL., "Enhancing HMM-based biomedical named entity recognition by studying special phenomena", *Journal of Biomedical Informatics*, Vol. 37, pages 411-422. 2004.
- [11] Lee, Ki-Joong., Hwang, Young-Sook., Kim, S., and Rim, Hae-Chang. "Biomedical named entity recognition using two-phase model based on SVMs", *Journal of Biomedical Informatics* Vol. 37, Issue 6, Dec, pp. 436-447, <https://doi.org/10.1016/j.jbi.2004.08.012>. 2004.
- [12] Finkel, J., Dingare, S., Manning, CD., Nissim, M., Alex, B., and Glover, C. "Exploring the boundaries: gene and protein identification in biomedical text", *BMC Bioinformatics*, doi:10.1186/1471-2105-6-S1-S5. 2005.
- [13] Goulart, R. R. V., Lima, V. L. S. d., and Xavier, C. "A systematic review of named entity recognition in biomedical texts", *Journal of the Brazilian Computer Society* 17 (2) pages 103-116. DOI: 10.1007/s13173-011-0031-9. 2011.
- [14] AlshaiKhdeeb, B., and Ahmad, K. "Biomedical Named Entity Recognition: A Review", *International Journal on Advanced Science Engineering and Information Technology*, Vol. 6, No. 6, pages. 889-895. ISSN. 2088-53342016. 2016.
- [15] Wulandari, DW., Adikara, PP., dan Adinugroho, S. "Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier", *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol. 2, No. 11, Hlm. 4555-4563. ISSN. 2548-9646X. 2018.
- [16] Lee, W., Khuri, N., and Wu, M. "Information Extraction from Biomedical Text Using Machine Learning", *Master's Theses and Graduate Research*, San Jose State University. 2019.