Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 11, November 2018, hlm. 4555-4563 http://j-ptiik.ub.ac.id

Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan *Rule Based* dan *Naïve Bayes Classifier*

Dayinta Warih Wulandari[1], Putra Pandu Adikara[2], Sigit Adinugroho[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]dayintaw@student.ub.ac.id, [2]adikara.putra@ub.ac.id, [3]sigit.adinu@ub.ac.id

Abstrak

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 biomedis. Karena biomedis memiliki skala yang luas, penelitian hanya akan berfokus pada dokumen biologi sel. *Rule based* adalah metode yang aturan dalam sistem dibuat sendiri berdasarkan pengetahuan linguistik. *Naïve Bayes Classifier* (NBC) merupakan jenis klasifikasi statistik dengan teori utamanya adalah memprediksi probabilitas keanggotaan kelas. Penelitian ini akan menggunakan *rule based* dan NBC untuk NER dalam dokumen biologi sel. Dengan 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.

Kata kunci: *named entity recognition, NER, rule based, naïve bayes classifier, biologi sel.*

Abstract

Named Entity Recognition (NER) is useful to help identify and detect entities of a word. The biomedical field has many literature so NER is highly demanded in this domain. Since biomedical has a large scale, research will only focus on biology cell documents. This research will use rule based and Naïve Bayes Classifier for NER in biology cell documents. With 19 training documents which processed and annotated manually to search for Named Entity (NE) and obtain 1135 word training data. Test documents are denoted and tagged by tagger site then search for bigram and trigram. Furthermore, rule-based process, if in the rule based not found solution, it will continue on feature extraction process and NBC. Using 16 NE classes, 18 rules, and 7 features were tested with three scenarios: rule based testing, NBC, and a combination of both. The highest average precision, recall and f-measure with micro average on rule based is 0.85. With macro average the highest recall and f-measure obtained combination is 0.66 and 0.45, while the highest precision obtained rule based is 0.39.

Keywords: *named entity recognition, NER, rule based, naïve bayes classifier, biology cell.*

1. PENDAHULUAN

Natural Language Processing (NLP) atau pengolahan bahasa alami merupakan suatu bidang ilmu *Artificial Intelligence* (AI) yang mempelajari komunikasi antara manusia dengan komputer melalui Bahasa alami (Suciadi, 2001). *Natural Language Understanding* (NLU) dan *Natural Language Generation* adalah dua tugas utama dari Natural *Language Processing (NLP).* Selain NLG dan NLU, NLP juga bertugas untuk beberapa hal antara lain *automatic summarization* (peringkasan otomatis), *Information Extraction* (IE), *Information Retrieval* (IR), *Named Entity Recognition* (NER), dan lain sebagainya (Amarappa & Sathyanarayana, 2015).

NER adalah bagian dari proses *text mining* dan *natural language processing* sangat berguna pada proses ekstraksi informasi. Tugas utama dari NER adalah mengidentifikasi dan mengklasifikasikan nama dalam teks ke dalam kelas-kelas yang telah ditentukan (Zhang, J., et al, 2004). Pada penelitan ini, penulis ingin menerapkan teknik NER pada dokumen biologi berbahasa Indonesia.

Bidang biomedis memiliki banyak literatur sehingga NER sangat dituntut dalam domain biomedis. Teknik ini bermanfaat untuk banyak aplikasi, seperti *text mining* di domain biomedis, alat bioinformatika, pembangunan *database* biomedis, dan lain-lain (Zhang, J., et al, 2004). Ilmu biomedis mengandung berbagai macam cabang ilmu biologi, diantaranya biologi sel, biologi molekuler, biokimia dan sebagainya. Karena biomedis memiliki skala yang sangat luas, penelitian ini hanya akan berfokus pada dokumen biologi sel.

Rule based adalah sebuah metode dimana aturan yang ada di dalam sistem dibuat sendiri berdasarkan pengetahuan linguistik. Selain menggunakan kamus dan klasifikasi, NER juga bisa menggunakan aturan-aturan tertentu. Dalam penelian ini, berdasarkan dokumen latih beberapa NE memiliki pola-pola tertentu, sehingga bisa dibuat polanya yang akan menjadi aturan dalan *rule-based* NER.

Naïve Bayes Classifier merupakan jenis klasifikasi statistik. Teori utama dari *Naïve Bayes* adalah memprediksi probabilitas keanggotaan kelas (Han & Kember, 2006). Pada penelitian sebelumnya *Using Naïve Bayes Sequence Classification for Named Entity Tagging* (Shah & Tariq, 2014) menunjukkan bahwa performa *Naïve Bayes* akan semakin meningkat ketika fitur ditambah. Penelitian tersebut menghasilkan *precision* terendah 0,52 dan tertinggi 0,88, *recall* terendah 0,68 dan tertinggi 0,98, dan *F-Measure* terendah 0,61 dan tertinggi 0,87. Sedangkan pada penelitian yang dilakukan oleh Mahmudy & Widodo (2014) dengan judul Klasifikasi Artikel Berita secara Otomatis Menggunakan Metode *Naïve Bayes Classifier* yang Dimodifikasi menunjukkan bahwa metode ini dimodifikasi ataupun tidak akan menghasilkan akurasi yang semakin baik jika data latih semakin banyak.

Dari penjabaran di atas, penelitian ini akan menggunakan *Rule Based* dan *Naïve Bayes Classifier* untuk mengklasifikasikan entitas dari suatu kata dalam dokumen biologi sel dengan pertimbangan bahwa akurasi dari *Naïve Bayes Classifier* tidak buruk jika terdapat data latih dan fitur yang mendukung serta *rule based* diharapkan mampu meningkatkan performa sistem.

2. DASAR TEORI

2.1. Pemrosesan Bahasa Alami (*Natural Language Processing*)

Natural Language Processing (NLP) merupakan salah satu aplikasi *Artificial Intelligence* (AI) yang dikembangkan agar komputer mengerti dan memahami bahasa alami yang diberikan dan memberi respon hasil pengolahan sesuai yang diinginkan (Danuari, 2013). NLP mencoba membuat agar komputer mengerti bahasa manunisa dengan cara memberi pengetahuan kepada komputer tentang bahasa manusia. Sistem Pemrosesan Bahasa Alami atau *Natural Language Processing* (NLP) adalah *software* yang mengolah bahasa manusia. NLP adalah teknik komputasi untuk menganalisis dan merepresentasikan teks alami pada satu atau lebih tingkat analisis linguistik dengan tujuan mencapai pemrosesan bahasa seperti manusia untuk berbagai tugas atau aplikasi.

2.2. *Named Entity Recognition* (NER)

Named Entity adalah frasa benda (*noun phrase*) yang memiliki tipe spesifik. NER bertujuan untuk menemukan dan menentukan jenis *named entity* pada teks. NER dapat digunakan untuk mengetahui relasi antar *named entity* dan *question answering system*. Tugas utama NER adalah untuk mencari *named entiy* dan menentukan tipe *named entity*. Cara dasar untuk mengenali *named entity* adalah dengan mencari jenis dari setiap kata pada teks menggunakan kamus. Namun penggunaan kamus dalam menentukan *named entity* memiliki beberapa permasalahan, salah satunya yaitu ambiguitas.

Penelitian ini berpedoman pada penelitian sebelumnya yaitu *Enhancing HMM-based Biomedical Named Entity Eecognition by Studying Special Phenomena* (Zhang, J., et al., 2004) dalam menentukan NE *class* yang akan digunakan. Dalam penelitian tersebut terdapat 22 NE *class*, namun setelah didiskusikan lebih lanjut dengan pakar, penelitian ini hanya akan berfokus pada biologi sel sehingga NE *class* disederhanakan menjadi 16 *class*.

Tabel 1. Daftar NE *class*

Named Entity	Contoh
CELL COMPONENT	Membran, Nukleus, Ribosom
CELL TYPE	Prokariotik, eukariotik
PROTEIN	Lipoprotein, enzim
CELL LINE	HEK 293
MULTI CELL	Tumbuhan, paramecium, fungi
RNA	Rrna
MONO CELL	Bakteri, archea
VIRUS	Adenovirus
LIPID	Fosfolipid
CARBOHYDRAT	Glukosa, pentose, deoksiribosa
PEPTIDE	Polipeptida, ACV-Tripeptida
DNA	mtDNA
NUCLEOTIDE	Guanina, Nukleobasa, polinukleotida
TISSUE	Embrio, pektin
ATOM	Hidrogen, kalsium
INORGANIC	Air, O2, CO2

2.3. *Rule Based*

Rule based adalah sebuah metode dimana aturan yang ada di dalam sistem dibuat sendiri berdasarkan pengetahuan linguistik. Analisis dilakukan pada tingkatan sintaksis dan semantik secara lebih mendalam merupakan kelebihan dari metode ini (Utami & Hartati, 2007). Pengetahuan linguistik meliputi tata bahasa dan leksikon bahasa serta melibatkan algoritma yang menentukan secara rinci masing-masing operasi yang terlibat dalam analisis (Shaalan, 2010). Dalam penelitian yang dilakukan oleh Shaalan (2010) menggunakan pendekatan *rule based* perlu adanya *rule* sebanyak mungkin karena dianggap akan meningkatkan performa sistem.

Tabel 2. Daftar aturan

No	Rule	NE
1	~lipid	LIPID
2	~protein	PROTEIN
3	~RNA	RNA
4	~virus	VIRUS
5	~viridae	VIRUS
6	~DNA	DNA
7	~osa	CARBOHYDRAT
8	~rida	CARBOHYDRAT
9	~ina	NUCLEOTIDE
10	~nukleotida	NUCLEOTIDE
11	~riotik	CELL TYPE
12	~riota	CELL TYPE
13	Protein ~	PROTEIN
14	Jaringan ~	TISSUE
15	~ Virus	VIRUS
16	Virus ~	VIRUS
17	~ sel	CELL COMPONENT
18	~ lemak	LIPID

2.4. *Naïve Bayes Classifier* (NBC)

Naïve Bayes merupakan sebuah metode klasifikasi dengan menggunakan probabititas sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan (Bustami, 2013).

Persamaan dari teorema bayes dapat dilihat pada Persamaan 1:

$$P(C \mid X) = \frac{(PX \mid C) \times P(C)}{P(X)} \quad (1)$$

Keterangan:
X: data dengan *class* yang belum diketahui
C: Hipotesis data merupakan suatu *class* spesifik
P(C|X): Probabilitas hipotesis H berdasarkan kondisi X (*posterior*)
P(C): Probabilitas hipotesis H (*prior* probabilitas)
P(X|C): Probabilitas X berdasarkan kondisi pada hipotesis C (*likelihood*)
P(X): Probabilitas X

Untuk mencari Likelihood dapat dilakukan seperti pada Persamaan 2

$$P(wi \mid Cj) = \frac{count(wi, Cj)}{\sum count(w, Cj)} \quad (2)$$

Keterangan:
$P(w_i \mid C_j)$: *Likelihood*
$count(w_i, C_j)$: Jumlah banyaknya sebuah kata w_i muncul di kelas C_j
$\sum count(w, C_j)$: Jumlah seluruh kata w dalam kelas C_j

2.5. *Laplacian Smoothing*

Keterbatasan data menyebabkan munculnya nilai nol pada perhitungan probabilitas seperti *naïve bayes*. Hal ini dapat menyebabkan *naïve bayes* tidak bisa melakukan klasifikasi pada data, sehingga munculnya nilai nol harus dihindari. *Smoothing* adalah solusi untuk menghindari nilai nol pada perhitungan probabilitas (Nugraha et al, 2013).

Laplacian Smoothing adalah salah satu metode *smoothing* yang dapat diterapkan pada

algoritme *naïve bayes*. Rumus *Laplatian Smoothing* dapat dilihat pada Persamaan 3.

$$P(wi|Cj) = \frac{count(wi,Cj)+i}{(\sum_{w \in V} count(w,Cj))+|C|} \quad (3)$$

Keterangan:

$P(w_i \mid C_j)$: *Likelihood*

$count(w_i, C_j)$: Jumlah banyaknya sebuah kata w_i muncul di kelas C_j

$\sum_{w \in V} count(w, C_j)$: Jumlah seluruh kata w dalam kelas C_j

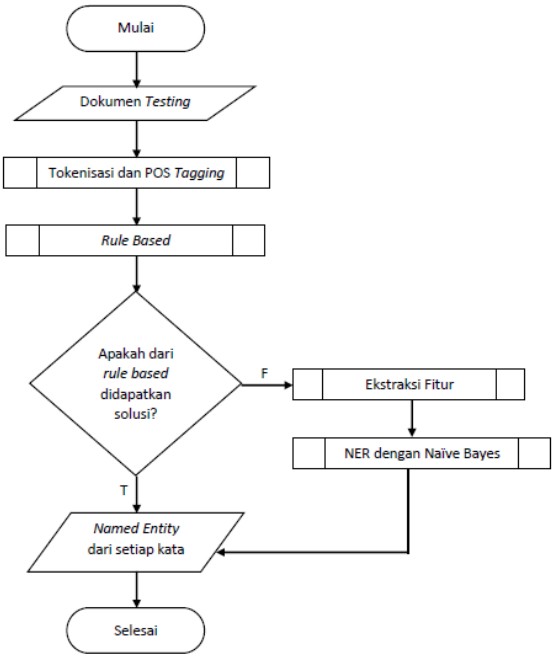
$|v|$: Banyaknya kata unik

3. PERANCANGAN DAN IMPLEMENTASI

Pendekatan *rule based* dan algoritme n*aïve bayes* akan digunakan untuk menemukan *named entity* dari setiap kata dalam dokumen biologi. Algoritme *naïve bayes* merupakan algoritme yang sudah sering digunakan dalam permasalahn teks dan terbukti memiliki akurasi yang baik asalkan memiliki data latih yang lengkap atau banyak. Dokumen latih yang digunakan mempunyai 7 *features* dengan dokumen latih sebanyak 20 dan dokumen *testing* sebanyak 1. Dengan menggunakan acuan jurnal dan pakar terdapat 16 NE *class*. Sebanyak 18 aturan telah didapatkan berdasarkan pengamatan pada data latih yang mana aturan ini akan digunakan dalam *rule based*.

Sebelum dilakukan implementasi pada sistem, tahap pertama yang dilakukan adalah melakukan tokenisasi pada dokumen latih secara manual. Selanjutnya dilakukan pos *tag* secara manual juga. Setelah ketiga tahap tersebut selesai, maka selanjutnya adalah mencari dan menentukan *named entity* dari setiap kata. Selanjutnya kata-kata tersebutlah yang dijadikan acuan dalam perhitungan dengan algoritme Naïve Bayes dengan fitur term, pos tag kata, kata sebelumnya, *suffix* dengan 3 huruf, *prefix* dengan 3 huruf, jumlah kata, dan ada tidaknya angka dalam kata tersebut. Hal yang sama dilakukan terhadap data *testing*, yaitu dilakukan tokenisasi, pos tag dengan mengambil API dari *Tag site*. Dengan demikian akan didapatkan fitur term, pos tag kata, kata sebelumnya, *suffix* dengan 3 huruf, *prefix* dengan 3 huruf, jumlah kata, dan ada tidaknya angka dalam kata tersebut. Dalam sistem, *rule based* dilakukan terlebih dahula dengan menggunakan aturan yang berpacu pada jurnal dengan penambahan oleh pakar. Jika dalam *rule based* tidak ditemukan solusi maka akan dilakukan perhitungan dengan algoritme *Naïve Bayes* untuk menentukan *named* entity dari kata tersebut. Hasil keluaran atau output akan dilakukan evaluasi.

Berikut adalah rancangan diagram alir kerja sistem yang ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alir Kerja Sistem

3.1. Tokenisasi dan POS Tag

Proses tokenisasi dan POS Tag merupakan proses yang digabung menjadi satu dengan menggunakan API dari *Tagger* Site http://bahasa.cs.ui.ac.id/postag/API/tag?.
Tagger Site tersebut sudah menyediakan tokenisasi dan POS Tag sehingga akan langsung didapatkan hasil tokenisasi dan POS Tagnya masing-masing kata. Selain menggunakan *tagger site* akan dilakukan tokenisasi lagi untuk mencari *bigram* dan *trigram*. Hal ini dilakukan dengan harapan memperbaiki tokenisasi yang dilakukan oleh *tagger site* dan untuk menyesuaikan dengan aturan pada *rule based* dimana ada kata yang berupa frasa. Sedangkan untuk POS Tag tetap mengikuti hasil dari *tagger site*.

3.2. *Rule Based*

1. Setelah mendapatkan token dan POS tag maka langkah selanjutnya adalah NER menggunakan *rule based. Rule based* akan memeriksa kecocokan tiap token dengan tiap aturan (*rule* 1 sampai 18). Jika salah 1 aturan sudah tepat/cocok maka akan langsung didapatkan hasilnya yaitu NE dari token tersebut, namun jika tidak ada aturan

yang cocok maka akan masuk ke proses ekstraksi fitur dan *Naïve Bayes Classifier*

3.3. Ekstraksi Fitur dan *Naïve Bayes Classifier* (NBC)

Tahap terakhir dalah NER dengan Naïve Bayes. Proses ini akan dilakukan jika *rule based* tidak mampu memberikan solusi. Namun, sebelum masuk pada *Naïve Bayes* terlebih dulu dilakukan ekstraksi fitur. Proses ini melakukan pemilihan fitur-fitur pada setiap kata (term, POS Tag, kata sebelum, sufiks, prefiks, jumlah kata, angka). Jika ekstraksi fitur telah selesai maka akan digunakan dalam perhitungan menggunakan *Naïve Bayes.*

4. PENGUJIAN DAN ANALISIS

Dokumen latih dan dokumen uji diambil dari Wikipedia. Dokumen latih berjumlah 19 dan dokumen uji berjumlah 1 dokumen dengan sumber dan pembahasan yang berbeda dengan harapan setiap NE mempunyai anggota sehingga dapat ditemukan *precision, recall* dan *f-measure* yang baik. Tabel 3 menunjukkan judul dokumen yang digunakan.

Tabel 3 Daftar dokumen latih dan uji

No.	Judul	Akses
1	Badan Golgi	07/02/2018
2	Flagelium	07/02/2018
3	Immortalized Cell Line	07/02/2018
4	Inti Sel	07/02/2018
5	Kloroplas	07/02/2018
6	Membran Sel	07/02/2018
7	Mitokondria	07/02/2018
8	Nukleotida	07/02/2018
9	Organel	07/02/2018
10	Retikulum Endoplasma	07/02/2018
11	Ribosom	07/02/2018
12	Silia	07/02/2018
13	Sitoplasma	07/02/2018
14	Sitoskeleton	07/02/2018
15	Vesikel	07/02/2018
16	Asam Deoksiribonukleat	07/02/2018
17	Asam Ribonukleat	07/02/2018
18	Peptida Nonribosomal	07/02/2018
19	Virus DNA	07/02/2018
20	Biologi Sel	27/02/2018
21	Transpor Pasif	27/02/2018
22	Sel (Biologi)	27/02/2018
23	Jaringan	27/02/2018
24	Virus	27/02/2018
25	Basa Nukleotida	27/02/2018

Pengujian dilakukan dengan 3 kali percobaan atau 3 skenario. Setiap percobaan dihitung *precision, recall* dan *f-measure* masing-masing kelas yang kemudian akan diambil nilai tengahnya (rata-ratanya) untuk mendapatkan kesimpulan dari hasil sistem ini. Skenario pertama adalah pengujian terhadap pendekatan *rule based*, skenario kedua dengan menggunakan *Naïve Bayes Classifier*, dan yang ketiga adalah kombinasi *rule based* dan *Naïve Bayes Classifier*. Dikarenakan adanya ketidak seimbangan data, maka pengambilan nilai rata-rata dengan *micro average* akan digunakan. *Micro average* akan menggabungkan kontribusi semua kelas untuk menghitung matriks rata-rata. Dalam klasifikasi multi kelas, *micro average* lebih baik jika ada ketidakseimbangan kelas. Nilai rata-rata dengan *macro average* juga akan dihitung sebagai pembanding.

Dokumen latih yang sudah ditokenisasi dan mendapat POS Tag dari *tagger site* kemudian dianotasi secara manual dan didapatkan beberapa token yang merupakan NE. Hal demikian juga dilakukan pada dokumen uji yang kemudian akan digunakan sebagai *ground truth.* Penelitian ini menggunakan 19 dokumen latih dan setelah diproses menjadi 1135 data latih dalam bentuk token.

4.1. Hasil Pengujian

1. *Rule Based*

Tabel 4 menunjukkan hasil *precision, recall*¸ dan *f-measure* tiap NE *class* dalam pengujian *rule based*.

Tabel 4. Hasil pengujian *rule based*

NE	*Precision*	*Recall*	*F-Measure*
CELL COMPONENT	0,5	0,1529	0,2342
CELL TYPE	0,8846	0,793103	0,8363
PROTEIN	0,1666	1	0,2857
CELL LINE	0	0	0
MULTI CELL	0	0	0
RNA	0,5	1	0,6666
MONO CELL	0	0	0
VIRUS	0,75	0,75	0,75
LIPID	0	0	0
CARBOHYDRAT	1	0,2	0,3333

PEPTIDE	0	0	0
DNA	0	0	0
NUCLEOTIDE	1	0,7777	0,875
TISSUE	0,6666	1	0,8
ATOM	0	0	0
INORGANIC	0	0	0
OTHERS	0,8694	0,97954	0,9212

2. *Naive Bayes Classifier*

Tabel 5 menunjukkan hasil *precision, recall*¸ dan *f-measure* tiap NE *class* dalam pengujian *Naive Bayes*.

Tabel 5. Hasil pengujian *Naive Bayes*

NE	*Precision*	*Recall*	*F-Measure*
CELL COMPONENT	0,5142	0,8470	0,64
CELL TYPE	0,6097	0,8620	0,7142
PROTEIN	0,0344	1	0,0666
CELL LINE	0	0	0
MULTI CELL	0,3157	0,8571	0,4615
RNA	0,1428	1	0,25
MONO CELL	0,3333	0,3333	0,3333
VIRUS	0,6666	0,5	0,5714
LIPID	0,125	1	0,2222
CARBOHYDRAT	0,1666	0,2	0,1818
PEPTIDE	0	0	0
DNA	0	0	0
NUCLEOTIDE	0,8421	0,8888	0,86486
TISSUE	0	0	0
ATOM	0,2	0,5	0,2857
INORGANIC	0	0	0
OTHERS	1	0,8209	0,9016

3. *Rule* based dan *Naive Bayes* (Kombinasi)

Tabel 6 menunjukkan hasil *precision, recall*¸ dan *f-measure* tiap NE *class* dalam pengujian *rule based* dan *Naive Bayes* (kombinasi)

Tabel 6. Hasil pengujian kombinasi

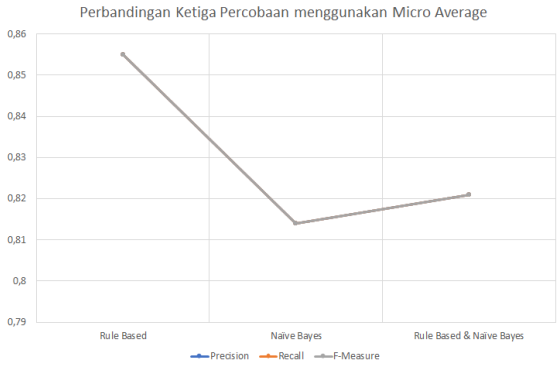
NE	*Precision*	*Recall*	*F-Measure*
CELL COMPONENT	0,4905	0,9176	0,6393
CELL TYPE	0,5853	0,8275	0,6857
PROTEIN	0,1666	1	0,2857
CELL LINE	0	0	0
MULTI CELL	0,3636	0,8571	0,5106
RNA	0,1428	1	0,25
MONO CELL	0,8181	0,75	0,7826
VIRUS	0,375	0,75	0,5
LIPID	0,1111	1	0,2
CARBOHYDRAT	0,1666	0,2	0,1818
PEPTIDE	0	0	0
DNA	0	0	0
NUCLEOTIDE	0,8571	1	0,9230
TISSUE	0,25	1	0,4
ATOM	0,1666	0,5	0,25
INORGANIC	0	0	0
OTHERS	1	0,8184	0,9001

4.2. Analisis

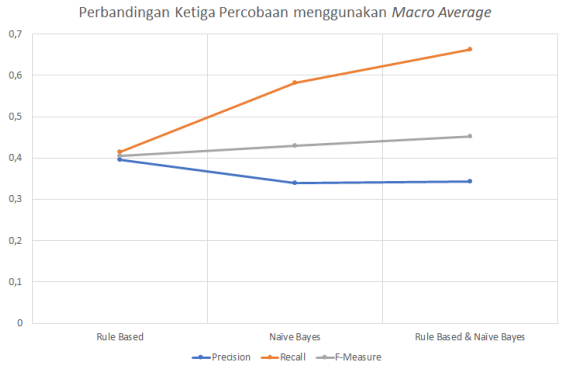
Hasil perbandingan rata-rata *precision, recall* dan *F-Measure* menggunakan *micro average* ditunjukkan pada Tabel 7. Gambar 2 dan Gambar 3 menunjukkan grafik perbedaan hasil rata-rata *precision, recall* dan *F-Measure* menggunakan *micro* dan *macro average.*

Tabel 7. Hasil ketiga percobaan

	Precision		*Recall*		*F-Measure*	
	mic	*mac*	*mic*	*mac*	*mic*	*mac*
Rule Based	0,85	0,39	0,85	0,41	0,85	0,40
Naïve Bayes	0,81	0,34	0,81	0,58	0,81	0,42
***Rule Based & Naïve Bayes* (Kombinasi)**	0,82	0,34	0,82	0,66	0,82	0,45



Gambar 2. Grafik rata-rata pecision, recall dan f-measure menggunakan micro average



Gambar 3 Grafik rata-rata *pecision, recall* dan *f-measure* menggunakan macro average

Dari hasil tersebut, dapat disimpulkan bahwa nilai dari *precision, recall* dan *f-measure* dengan menggunakan *micro average* cukup bagus untuk ketiga percobaan. Ketiga nilai mempunyai nilai yang sama pada masing-masing pengujian. Dapat dilihat dari hasil pengujian tersebut bahwa *rule based* dapat meningkatkan nilai *precision, recall* dan *f-measure* dari *Naïve Bayes* dengan kenaikan pada kombinasi yang tidak signifikan yaitu hanya sebesar 0,85% untuk *precision, recall, f-measure*. Namun hasil dari kombinasi tidak lebih baik dari pendekatan *rule based* sendiri. Hal ini disebabkan karena rendahnya nilai *presicion, recall* dan *f-measure* dari *Naïve Bayes* sendiri. Meskipun demikian, perbedaan hasil dari ketiga percobaan tidaklah besar, hasilnya sama-sama berada pada angka 0,8. Nilai *precision, recall* dan *f-measure* menggunakan *micro average* tertinggi diraih oleh *rule based* dengan nilai sama yaitu 0,855.

Nilai dari *precision, recall* dan *f-measure* dengan menggunakan *macro average* kurang bagus karena nilai dibagi sama rata dengan kelas-kelas yang tidak memiliki anggota. Nilai *precision* tertinggi didapatkan *rule based* yaitu 0,39. Nilai *recall* dan *f-measure* tertinggi didapatkan oleh kombinasi *rule based* dan *Naïve Bayes* yaitu 0,66 untuk *recall* dan 0,45 untuk *f-measure*. Dengan menggunakan *macro average*, *rule based* dapat meningkatkan performa dari *Naïve Bayes* pada kombinasi keduanya dengan kenaikan nilai *precision, recall,* dan *f-measure* secara berturut yaitu 0,88%, 14% dan 5%. Begitu juga dengan sebaliknya, nilai *recall* dan *f-measure* dari *rule based* juga meningkat sebanyak 59% untuk *recall* dan 11% untuk *f-measure*, namun *precision* kombinasi tidak lebih baik dari *rule based* sendiri.

Perbedaan hasil antara *micro* dan *macro average* ini terjadi karena ketika menggunakan *micro average* yang dihitung hanyalah nilai dari kelas yang memiliki anggota sehingga nilai *precision, recall,* dan *f-measure* akan bagus namun tidak memperhatikan persebaran data sehingga saat di *rule based* hanya akan menghitung kelas-kelas yang memiliki anggota, di *Naïve Bayes* dan kombinasi juga demikian. Ketika menggunakan *macro average* nilai *precision, recall,* dan *f-measure* buruk, namun *macro average* memperhatikan persebaran data sehingga dapat diketahui kelas-kelas yang tidak memiliki anggota di *rule based* dan *Naïve Bayes* telah memiliki anggota saat di kombinasi, hal inilah yang menyebabkan hasil *precision, recall,* dan *f-measure* menggunakan *macro average* mengalami kenaikan.

Ada beberapa faktor yang menyebabkan kurang maksimalnya hasil penelitian ini selain faktor *micro* dan *macro average.* Faktor secara keseluruhan adalah POS Tag yang menggunakan *tagger site* dengan akurasi 79% masih menunjukkan beberapa kesalahan pada hasil POS Tag sehingga menyebabkan kata yang seharusnya bukan nomina menjadi nomina dan berakibat pada kata tersebut masuk dalam NE *Class.* Selain itu juga berimbas pada pembuatan *bigram* dan *trigram* dimana hanya dilakukan dengan cara menggabungkan kata bersebelahan yang merupakan nomina.

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.

5. KESIMPULAN

Berdasarkan hasil penelitian *Named Entity Recognition* pada Dokumen Biologi menggunakan *Rule Based* dan *Naive Bayes Classifier* didapatkan kesimpulan sebagai berikut:

1. *Rule based* dan *Naïve Bayes Classifier* dapat digunakan dalam *Named Entitty*

Recognition pada dokumen Biologi berbahasa Indonesia. *Rule based* memberikan hasil berdasarkan kecocokan dengan aturan. *Naïve Bayes Classifier* memberikan hasil berdasarkan peluang kemungkinan terjadi atau dalam kasus ini peluang kemungkinan memiliki sebuah NE. *Named Entitty Recognition* dengan *Rule Based* dan *Naïve Bayes Classifier* dilakukan menggunakan 18 aturan dan 1135 data *training* berupa kata. Aturan disusun dengan acuan jurnal dan data latih. Data *training* diambil dari Wikipedia lalu dokumen tersebut kemudian ditokenisasi dan di POS Tag dengan menggunakan API Tag Site http://bahasa.cs.ui.ac.id/postag/API/tag?. Dari hasil tersebut kemudian dilakukan pencarian dan pemberian *named entity*. Selanjutnya data akan diolah sedemikian rupa sehingga dapat menjadi data *training* pada program. Untuk data *testing* juga diambil dari sumber yang sama. Cara kerja pengujiannya yaitu dengan memasukkan satu dokumen kemudian akan ditokenisasi dan di POS *Tagging* dengan menggunakan API Tag *Site* http://bahasa.cs.ui.ac.id/postag/API/tag?. Setelah itu akan dilakukan pengecekan *bigram* dan *trigram.* Setelah selesai maka akan masuk proses *rule based* dengan cara mencari kecocokan setiap token dengan semua aturan. Jika ada aturan yang cocok maka akan langsung didapatkan hasilnya yaitu NE dari token tersebut, namun jika semua aturan tidak memenuhi akan dilanjutkan pada proses *Naïve Bayes.* Langkah pertama dalam proses inti ini adalah menghitung *prior*. Selanjutnya menghitung *likelihood* dari setiap fitur yang berarti akan ada 7 fitur. Langkah selanjutnya yaitu menghitung *posterior* dengan mengalikan *prior* dan semua *likelihood*. Langkah terakhir yaitu memilih NE dengan *posterior* tertinggi.

2. Dari penelitian ini didapatkan hasil nilai rata-rata menggunakan *micro average* yaitu *precision:* 0,855, *recall*: 0,855 dan *f-measure*: 0,855 pada percobaan *rule based* saja. Pada percobaan *Naïve Bayes* didapatkan *precision:* 0,814, *recall*: 0,814 dan *f-measure*: 0,814. Pada kombinasi *rule* based dan *Naïve Bayes* didapatkan *precision:* 0,821, *recall*: 0,821 dan *f-measure*: 0,821. Sedangkan ketika menggunakan *macro average* didapatkan nilai rata-rata *precision:* 0,396, *recall*: 0,415 dan *F-Measure*: 0,405 pada percobaan *rule based* saja. Pada percobaan *Naïve Bayes* didapatkan *precision:* 0,340, *recall*: 0,581 dan *f-feasure*: 0,429. Pada kombinasi *rule* based dan *Naïve Bayes* didapatkan *precision:* 0,343, *recall*: 0,663 dan *f-measure*: 0,452. Hal ini dapat disebabkan oleh banyak faktor, yaitu tidak adanya aturan untuk beberapa NE *class* untuk *rule based* dan kurangnya data *training* mengingat *Naïve Bayes* sangat bergantung pada data *training.* Selain itu pada fitur POS Tag hanya memiliki akurasi 79% ternyata masih memiliki beberapa kesalahan POS Tag.

DAFTAR PUSTAKA

Amarappa, S. & Sathyanarayana, S.V., 2015. *Kannada Named Entity Recognition And Classification (NERC) Based On Multinomial Naïve Bayes (MNB) Classifier. International Journal on Natural Language Computing*, Vol. 4, No. 4.

Bustami, 2013. Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2.

Danuari, 2013. *Natural Language Processing* untuk *Structured Query Language* pada DBMS MySQL. Bengkalis: Politeknik - Negeri Bengkalis.

Han, J. & Kamber, M., 2006. *Data Mining Concept and Techniques 2nd Edition*. San Francisco: Morgan Kaufmann Publisher.

Mahmudy, W.F. & Widodo, A.W., 2014. Klasifikasi Artikel Berita Secara Otomatis Menggunakan Metode *Naïve Bayes Classifier* yang Dimodifikasi. TEKNO, Vol. 21 Maret 2015.

Nugraha, P.A., et al., 2013. Perbandingan Metode Probabilistik *Naïve Bayes Classifier* dan Jaringan Syaraf Tiruan *Learning Vector Quantization* dalam Kasus Klasifikasi Penyakit Kandungan. JURNAL ITSMART, Vol. 2, No. 2.

Rashel, F., et al., 2012. *Building an Indonesian Rule-Based Part-of-Speech Tagger*.

International Conference on Asian Language Processing. Kuching, Malaysia, 20-22 October 2014. Depok: Indonesia.

Shaalan, K., 2010. *Rule-based Approach in Arabic Natural Language Processing. International Journal on Information and Communication Technologies*, Vol. 3, No. 3.

Shah, M.A. & Tariq, S., 2014. *Using Naive Bayes Sequence Classification for Named Entity Tagging*. [online]. Tersedia di: < http://www.contrib.andrew.cmu.edu/~ mshah1/textfiles/nlp-report.pdf> [Diakses 12 Oktober 2017]

Suciadi, J., 2001. Studi Analisis Metode-Metode *Parsing* dan Intrepretasi Semantik pada *Natural Language Processing*. Jurnal Informatika, Vol. 2, No.1.

Utami, E. & Hartati, S., 2007. Pendekatan Metode *Rule Based* Dalam Mengalihbahasakan Teks Bahasa Inggris Ke Teks Bahasa Indonesia. Jurnal Informatika, Vol. 8, No. 1.

Zhang, J., et al., 2004. *Enchancing HMM-based biomedical named entity recognition by studying special phenomena. Journal of Biomedical Informatics* 37 (2004) 411-422.