EKSTRAKSI INFORMASI DOKUMEN KARYA ILMIAH MENGGUNAKAN MAXIMUM ENTROPY MARKOV MODEL

Dina Ilman[1], Ken Kinanti Purnamasari[2]

[1,2]Universitas Komputer Indonesia
Jl. Dipati Ukur No. 112-116, Lebakgede, Coblong, Kota Bandung, Jawa Barat 40132
E-mail: ilmandina@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRAK

Pendeteksian kategori dokumen karya ilmiah belum mampu dilakukan untuk dokumen dengan format beragam. Permasalahan tersebut dapat diatasi jika menggunakan *machine learning*. Algoritma *machine learning* yang digunakan adalah *Maximum Entropy Markov Model* (*MEMM*). Algoritma *MEMM* merupakan salah satu gabungan dari *Markov Model* dengan *Logistic Regression*. Proses pelatihan dari algoritma tersebut dokumen karya ilmiah dipelajari terlebih dahulu pola dari dataset dengan menentukan nilai *lerarning rate* dan bobot awal *theta*. Pengulangan berhenti berdasarkan nilai *cross entropy* jika sudah mendekati konvergen dan jumlah pengulangan. Pengujian berdasarkan model yang sudah dilatih dengan algoritma *viterbi* untuk mengekstraksi kategori pada dokumen karya ilmiah. Berdasarkan pengujian pada 40 dokumen karya limiah skripsi tahun 2011 sampai 2018, diperoleh rata-rata akurasi dengan pengujian token-kelas sebesar 77% dengan *learning rate* 0.001. Perolehan akurasi token-kelas disebabkan penggunaan fungsi fitur yang kurang unik untuk menentukan ciri dari setiap kategori dokumen, pelatihan bobot *theta* perlu menggunakan metode tambahan agar lebih tepat dan pengaruh *learning rate* yang menentukan seberapa banyak proses pelatihan dilakukan.

Kata kunci: Ekstraksi Informasi, *Maximume Entropy Markov Model*, *MEMM*,Dokumen Karya ilmiah, Skripsi, *Logistic Regression*

1. PENDAHULUAN

Ekstraksi informasi merupakan proses pengambilan informasi dari teks tidak terstruktur yang menghasilkan teks terstruktur dan rapi sekaligus mempermudah dalam menemukan informasi dari teks terstruktur [1]. Pengertian lain ekstraksi informasi menurut Piskorsi, ekstraksi informasi bertujuan untuk mengekstraksi sekumpulan data teks untuk mendapatkan fakta-fakta berkaitan dengan kejadian, entitas atau keterhubungan dalam bentuk informasi terstruktur sebagai masukan untuk basis data [2]. Dokumen karya ilmiah memiliki format yang beragam sehingga sulit untuk mendeteksi setiap komponen pada dokumen. Oleh karena itu, dibutuhkan metode untuk mengekstraksi informasi dari dokumen karya ilmiah sehingga aturan ekstraksi informasi sesuai dengan format dokumen karya ilmiah.

Penelitian sebelumnya menggunakan rule-based pada dokumen karya tulis ilmiah skripsi berbahasa Indonesia [3]. Pada penelitian tersebut ekstraksi informasi dilakukan dengan sistem berbasis aturan
untuk mendeteksi identitas pada dokumen skripsi, diantaranya cover, abstrak dan abstract. Berdasarkan pengujian 3 buah dokumen pada tahun 2017, akurasi yang diperoleh sebesar 100% sedangkan pengujian terhadap 50 dokumen yang beragam diperoleh rata-rata akurasi 57%. Penurunan akurasi disebabkan sistem yang tidak mampu menangani dokumen dengan format beragam. Oleh karena itu, permasalah tersebut dapat diatasi jika menggunakan machine learning.

Pada penelitian sebelumnya, penggunaan machine learning di bidang ekstraksi informasi sudah dilakukan pada makalah ilmiah [1]. Pada penelitian tersebut belum ada algoritma Maximum Entropy Markov Model (MEMM) sehingga dibutuhkan algoritma tersebut. Hasil penelitian sebelumnya oleh Susan Mengel dan Yaoquin Jing penggunaan MEMM untuk ekstraksi struktur data pada halaman web memiliki error rate yang rendah 0.14(0.08 dengan 30 halaman web) [4]. Shurthi S. melakukan studi pengenalan entitas bernama untuk bahasa malaysia menggunakan pos tag TnT dan metode MEMM dengan hasil akurasi 82.5% [5]. A. Nedjo et al. menggunakan MEMM untuk POS Tag otomatis pada bahasa Oromo dan menghasilkan akurasi 99.3% pada data latih dan 93.01% pada data uji [6].

Berdasarkan pemaparan tersebut maka dalam peneltian ini digunakan metode MEMM untuk membangun sistem ekstraksi informasi dokumen karya ilmiah berbahasa Indonesia, dengan batasan dokumen yang akan digunakan pada penelitian ini adalah dokumen karya ilmiah skripsi Program Studi Teknik Informatika Universitas Komputer Indonesia.

2. ISI PENELITIAN

Isi penelitian menjelaskan mengenai metode penelitian, ekstraksi informasi, dokumen karya

ilmiah skripsi, arsitektur sistem, tokenisasi, ekstraksi fitur, algoritma *MEMM* dan hasil pengujian

2.1. Metode Penelitian

Pada penlitian ini terdapat lima tahapan alur penelitian, diantaranya identifikasi masalah, studi literatur, pengumpulan data, pembangunan sistem ekstraksi informasi serta penarikan kesimpulan. Berkut blok diagram tahapan alur penelitian.

Gambar 1 Metode Penelitian

2.2. Ekstraksi Informasi

Teknologi ekstraksi informasi (*Information Extraction*) adalah teknologi yang berkaitan dengan cara menjadikan dokumen teks tidak terstruktur dengan domain tertentu ke dalam sebuah struktur informasi yang relevan. Dengan kata lain, tujuan utama dari IE adalah mencari informasi-informasi yang relevan dengan domain dan tidak memperdulikan informasi tidak relevan[9]. Secara garis besar, proses ekstraksi informasi terdiri dari dua tahap, yaitu mengidentifikasi data yang relevan, kemudian menyimpannya ke dalam bentuk
terstruktur untuk digunakan kemudian[10].
Algoritma mempelajari aturan-aturan ekstraksi berdasarkan dokumen teks sebagai data latih yang telah diberi anotasi mengenai entitas informasi yang akan diekstrak. Peran manusia diperlukan untuk memberikan label atau anotasi sesuai entitas yang akan diekstrak pada dokumen[10]:

1. *Preprocessing* data masukan

Data masukan berupa teks yang tidak terstruktur, *natural language text*. Informasi penting didapatkan dengan analisis linguistik, karena menghasilkan kata kunci dan fitur ciri penting untuk mengidentifikasi informasi. Analisis linguistik yang digunakan diantaranya tokenisasi, pembagian kalimat (*sentence splitting*), analisis morpologi, *parsing* dan *named entity*. Pada penelitian ini menggunakan tokenisasi dan ekstraksi fitur termasuk *named entity*.

1.1. Tokenisasi

Keadaan awal dalam bentuk karakter yang terhubung dengan tujuan untuk mengidentifikasi bagian dasar dari *natural language* seperti kata, tanda baca dan pemisah. Hasil dari token yang memiliki makna dan terhubung sebagai dasar untuk proses linguistik dan teks berikutnya.

1.2. Ekstraksi Fitur

Menurut Prihatini, ekstraksi fitur merupakan proses untuk mencari nilai - nilai fitur yang terkandung dalam dokumen [12]. Fitur dapat diartikan sebagai ciri dari setiap data yang dikenali oleh sistem sehingga menghasilkan nilai fitur. Ekstraksi fitur merupakan topik penting dalam klasifikasi, karena fitur-fitur yang baik akan sanggup meningkatkan tingkat akurasi, sementara fitur-fitur yang tidak baik cenderung memperburuk tingkat akurasi [13].

Ada 3 kelompok fitur yang akan digunakan dengan total 15 fitur, yaitu fitur lokal, fitur tata letak dan *named entity*. Fitur lokal adalah karakteristik yang terdapat dalam karakter setiap baris kalimat 7 fitur. Fitur tata letak adalah posisi suatu baris kalimat dalam bagian dokumen 3 fitur. Fitur *named enity* adalah fitur yang diekstrak dari dokumen berdasarkan aturan tertentu 3 fitur [14] sedangkan untuk fitur LOWERCASE dan EIGHTDIGIT merupakan fitur yang ditentukan oleh peneliti.

2.3. Dokumen Karya Ilmiah

Dokumen adalah surat tertulis atau tercetak yang dipakai sebagai bukti keterangan [14]. Karya ilmiah adalah karangan ilmu pengetahuan yang menyajikan fakta dan ditulis menurut metodologi penulisan yang baik dan benar [15]. Kamus Besar Bahasa Indonesia menjelaskan bahwa karya ilmiah berupa makalah berisi tulisan tentang suatu pokok yang dimaksudkan untuk dibacakan di muka umum dan sering disusun untuk diterbitkan. Pengertian karya ilmiah lainnya adalah makalah yang berisi karangan termasuk tugas peserta didik selama dalam pendidikan di sekolah [14]. Jenis karya ilmiah ada 2 yaitu hasil penelitian contoh skripsi, tesis, disertasi, buku dan makalah dan tinjauan atau usulan/gagasan sendiri contoh buku pelajaran, diktat dan modul. Skripsi termasuk hasil penelitian yang mengemukakan pendapat penulis berdasarkan pendapat orang lain. Pendapat yang diajukan harus didukung oleh data dan fakta empiris-objektif, baik berdasarkan penelitian langsung (observasi lapangan) maupun penelitian tidak langsung (studi kepustakaan) [15]. Skripsi termasuk hasil penelitian yang mengemukakan pendapat penulis berdasarkan pendapat orang lain. Pendapat yang diajukan harus didukung oleh data dan fakta empiris-objektif, baik berdasarkan penelitian langsung (observasi lapangan) maupun penelitian tidak langsung (studi kepustakaan) [15].

Pada penelitian ini dokumen karya ilmiah yang digunakan untuk data masukan adalah dokumen skripsi pada bagian halaman depan atau sampul dan abstrak dari program studi Teknik Informatika secara acak yang memiliki 17 kategori atau kelas. Pada halaman sampul terdiri dari kategori Judul Penelitian (Sampul), Jenis Penelitian, Kalimat Pengajuan, *Other*, Penulis (Sampul), NIM (Sampul), Program Studi, Fakultas, Universitas, Kota dan Tahun. Pada halaman abstrak terdiri dari Jenis Halaman, Judul Penelitian (Abstrak), *Other*, Penulis (Abstrak), NIM (Abstrak), Isi Abstrak dan Kata Kunci.

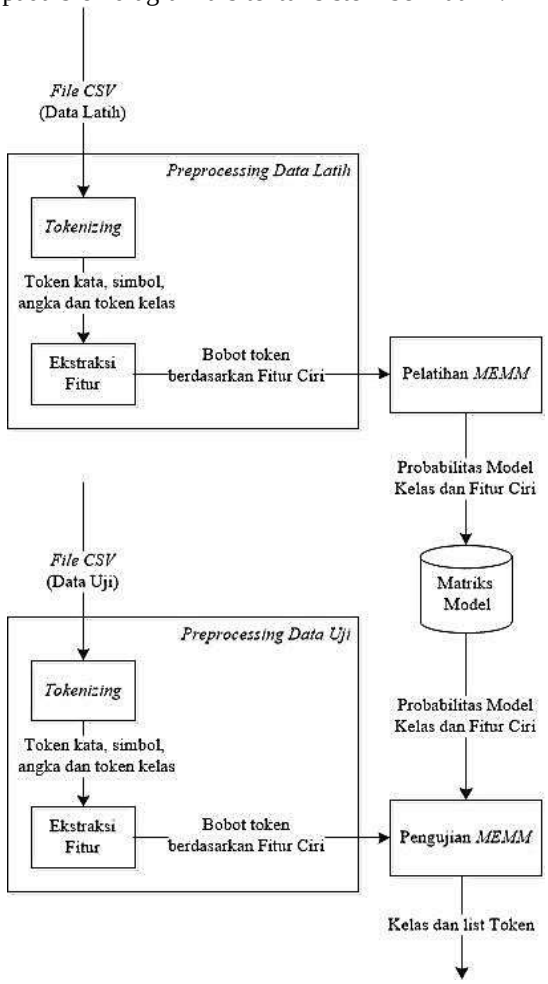
Tabel 1 Bagian-bagian Kategori dari Dokumen Karya Ilmiah

Lembar Sampul Skripsi		
No	**Kategori**	**Kelas**
1	Judul Penelitian (Sampul)	0
2	Jenis Penelitian	1
3	Kalimat Pengajuan	2
4	*Other*	15
5	Penulis (Sampul)	3
6	NIM (Sampul)	4
7	Prodi	5
8	Fakultas	6
9	Universitas	7
10	Kota	8
11	Tahun	9
Lembar Abstrak Skripsi		
No	**Kategori**	**Kelas**
12	Jenis Halaman	10
13	Judul Penelitian (Abstrak)	11
14	*Other*	15
15	Penulis (Abstrak)	12
16	NIM (Abstrak)	13
17	Isi Abstrak	14
18	Kata Kunci	16
12	Jenis Halaman	10
13	Judul Penelitian (Abstrak)	11
14	*Other*	15
15	Penulis (Abstrak)	12

Beberapa kategori pada Tabel 1 menjadi kelas dalam bentuk angka untk proses pelatihan *MEMM*. Jumlah seluruh kelas yang digunakan pada penelitian ini 17 kelas.

2.4. Arsitektur Sistem

Pembangunan sistem ekstraksi informasi terdapat beberapa proses. Proses utama yang terdapat pada sistem ekstraksi informasi menggunakan *MEMM* terdiri dari data latih dan data uji berupa *file csv*, *preprocessing* data latih dan data uji, pelatihan *MEMM* dan pengujian *MEMM*. Untuk proses lebih jelasnya, dapat dilihat pada blok diagram arsitektur sistem berikut ini.



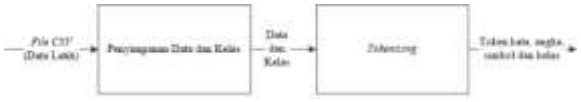
Gambar 2 Blok Diagram Arsitektur Sistem

Data latih dari hasil scan halaman sampul dan abstrak skripsi Teknik Informatika secara acak dalam bentuk *.pdf* kemudian dilakukan proses konversi kedalam format *.txt* secara manual untuk mengekstrak tulisan dengan proses OCR. Tahap *preprocessing* data latih melalui proses *tokenizing*. Tahap Ekstraksi Fitur dilakukan dengan menentukan masing-masing hasil token termasuk fitur yang mana. Terdapat 3 fitur yang dibandingkan fitur lokal, fitur tata letak dan fitur *named entity*. Tahap *Maximum Entropy Markov Model* melakukan pelatihan dengan maximum entropy kemudian pengujian untuk menentukan urutan kelas berdasarkan data dengan viterbi. Hasil pemetaan data pengujian berdasarkan model yang telah dibentuk sebelumnya.

2.5. Tokenisasi

Keadaan awal dalam bentuk karakter yang terhubung dengan tujuan untuk mengidentifikasi bagian dasar dari *natural language* seperti kata,

tanda baca dan pemisah. Hasil dari token yang memiliki makna dan terhubung sebagai dasar untuk proses linguistik dan teks berikutnya [11]. Berikut blok diagram tokenisasi.

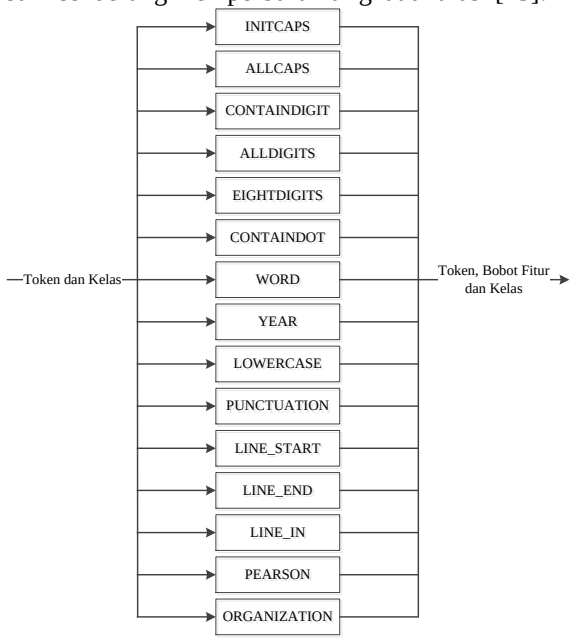


Gambar 3 Blok Diagram Tokenisasi

Proses tokenisasi pada penelitian ini tidak menghilangkan sejumlah token atau kata yang tidak penting karena menjadi ciri untuk proses ekstraksi fitur.

2.6. Ekstraksi Fitur

Menurut Prihatini, ekstraksi fitur merupakan proses untuk mencari nilai - nilai fitur yang terkandung dalam dokumen [12]. Fitur dapat diartikan sebagai ciri dari setiap data yang dikenali oleh sistem sehingga menghasilkan nilai fitur. Ekstraksi fitur merupakan topik penting dalam klasifikasi, karena fitur-fitur yang baik akan sanggup meningkatkan tingkat akurasi, sementara fitur-fitur yang tidak baik cenderung memperburuk tingkat akurasi [13].



Gambar 4 Blok Diagram Ekstraksi Fitur

Ada 3 kelompok fitur yang akan digunakan dengan total 15 fitur, yaitu fitur lokal, fitur tata letak dan *named entity*. Fitur lokal adalah karakteristik yang terdapat dalam karakter setiap baris kalimat 7 fitur. Fitur tata letak adalah posisi suatu baris kalimat dalam bagian dokumen 3 fitur. Fitur *named enity* adalah fitur yang diekstrak dari dokumen berdasarkan aturan tertentu 3 fitur [14] sedangkan untuk fitur LOWERCASE dan EIGHTDIGIT merupakan fitur yang ditentukan oleh peneliti.

Tabel 2 Ekstraksi Fitur

Nama Fitur	Deskripsi
Fitur Lokal	
INITCAPS	Dimulai dengan huruf kapital
ALLCAPS	Seluruh karakter adalah huruf kapital
CONTAINSDIGIT	Mengandung digit angka
ALLDIGITS	Seluruh karakter adalah digit angka
EIGHTDIGITS	Karakter berjumlah 8 dan digit angka
CONTAINDOT	Mengandung paling sedikit satu titik
WORD	Untuk menambah bobot perhitungan, pada kelas "KALIMAT_PENGAJUAN" termasuk fitur ini
YEAR	Untuk menandakan "TAHUN"
LOWERCASE	Seluruh karakter adalah huruf kecil
PUNCTUATION	Tanda baca
Fitur Tata Letak	
LINE_START	Berada di awal baris
LINE_END	Berada di akhir baris
LINE_IN	Berada di pertengahan baris
Fitur *Named Entity*	
PERSON	Nama orang
ORGANIZATION	Nama instansi

Setiap token yang memnuhi syarat dari salah satu ekstraksi fitur tersebut akan mendapatkan nilai bobot 1 dan jika tidak memnuhi maka nilai bobot 0.

Alat preprocessing yang tepat dalam banyak studi Natural Language Processing (NLP) sangat penting untuk dilakukan memberikan akurasi yang lebih baik. Tahap preprocessing leksikal, seperti pendeteksian kata-kata dasar (Stemming) dan deteksi jenis kata (penandaan POS) berdampak besar bagi sistem komputasi bahasa itu membutuhkan penentuan struktur kalimat. Dalam bahasa Indonesia, penelitian tentang Stemming dan Penandaan POS masih dilakukan, baik dengan menggunakan metode statistik atau aturan tertentu. Beberapa masalah yang dihadapi untuk pengolahan tersebut adalah kurangnya corpus di Indonesia dan Indonesia ketidaklengkapan aturan yang tersedia. Penelitian tentang stemming, pertama kali diterbitkan oleh Julie Beth Lovins pada tahun 1968 [21].

2.7. Maximum Entropy Markov Model

Supervised Learning disebut juga klasifikasi atau pembelajaran secara induktif dalam pembelajaran mesin (*machine learning*). Pembelajaran dilakukan dari data yang telah dikumpulkan sebelumnya dan menggambarkan keadaan sebelumnya dalam pengaplikasian di dunia nyata [15]. Menurut Zdravko, tujuan klasifikasi untuk membuat sebuah pemetaan (disebut juga model atau hipotesis) diantara sebuah set dokumen dan set label kelas. Hasil pemetaan digunakan untuk menentukan kelas dokumen baru (belum diberi label kelas awal) secara otomatis [16]. Salah satu pengklasifikasi yang akan digunakan dalam penelitian ini adalah pengklasifikasi sekuens (*Sequence Classifier*). Pengklasifikasi sekuens adalah model yang
bertugas untuk menentukan sebuah label atau kelas ke setiap unit dalam sebuah sekuens, sehingga memetakan sebuah sekuens observasi ke sekuens label [17]. Algoritma yang termasuk dalam pengklasifiksi sekuens adalah *Hidden Markov Model* (HMM) dan *Maximum Entropy Markov Model* (MEMM).

Markov Model disebut juga *Markov Chain* atau *Markov Process* merupakan bagian dari proses stokastik yang memiliki properti *Markov*, sehingga apabila diberikan data masukan keadaan saat ini, keadaan akan datang dapat diprediksi dan ia lepas dari keadaan masa lalu. Dengan kata lain, kondisi masa depan dituju dengan menggunakan probabilitas. Model bagian dari *Finite state* atau *Finite Automaton*.
Finite Automaton adalah kumpulan state yang transisi antar state-nya dilakukan berdasarkan observasi. Pada *Markov Chain*, setiap busur antar state berisi probabilitas kemungkinan jalur tersebut akan diambil. Jumlah probabilitas semua busur yang keluar dari sebuah simpul adalah satu. *Markov chain* bermanfaat
untuk menghitung probabilitas urutan kejadian yang dapat diamati. Untuk mengetahui urutan kejadian yang tersembunyi menggunakan algoritma *Maximum Entropy Markov Model*.

Maxium Entropy Markov Model atau MEMM adalah sekuens model yang diadaptasi dari *MaxEnt* (*multinomial logistic regresstion*) *classifier*. Berdasarkan *logistic regression*, MEMM adalah model sekuens yang diskriminatif sedangkan HMM (*Hidden Markov Model*) model sekuens yang generatif [17].
Pada MEMM menggunakan *Maximum Entropy* untuk proses klasifikasinya sehingga rumus pelatihan untuk menentukan kelas target terbaik $P(T|W)$ rumus (2.4) sebagai berikut [17]:

$$\hat{T} = arg\max_{T} P(T|W)$$
$$= arg\max_{T} P(t|t', w)$$
$$= \operatorname{argmax}_{T} \frac{\exp(\sum_j \theta_j f_j(w_i, t_i))}{\sum_{t' \in seluruh\ kelas\ T} \exp(\sum_j \theta_j f_j(w_i, t'))}$$

t_i adalah kelas ke-i dari data yang ditunjuk. W adalah kata dari seluruh dataset. T' adalah seluruh kelas awal yang telah ditentukan. θ adalah vektor bobot awal. $f_i(w_i, t_i)$ adalah fungsi indikator yang menghasilkan nilai 0 jika syarat tidak terpenuhi dan 1 jika syarat terpenuhi [17]. Indikator yang diambil berdasarkan kelas ke-i dan ekstraksi fitur ke-i.

Multinomial Logistic Regression atau disebut juga *Maximum Entropy Model* bagian dari pengklasifikasi exponensial atau *log-linear*, mengekstrak set kata, mengkombinasikan secara linear (mengkalikan setiap kata dengan bobot dan menjumlahkannya) kemudian menerapkan sebuah fungsi pada kombinasi tersebut [17]. Proses pelatihan dalam *logistic regression* menggunakan
fungsi objektif untuk meminimalkan nilai *error* pada data latih, yaitu *cross entropy loss function* (2.5) [17]:

$$L_{CE}(\hat{y}, y) = -\sum_{k=1}^{K} 1\{y = k\} \log p(y = k|x)$$

1{} sebagai fungsi untuk menentukan jika indeks kelas sebenernya ke k dari data latih sama dengan indeks *softmax* ke k maka nilai fungsi adalah 1 jika tidak memenuhi syarat maka nilai fungsi 0. Nilai *cross entropy* dihitung rata-rata seluruh datanya untuk mendapatkan nilai *error* secara keseluruhan. Untuk mengoptimalkan dan memperbarui nilai bobot θ berdasarkan hasil *cross entropy* jika belum mendekati 0 maka dihitung nilai *gradient descent*. Berikut proses perhitungan *gradient descent* pada rumus (2.6) [18]:

$$\frac{\partial L_{CE}}{\partial W_k} = (1\{y = k\} - p(y = k|x))X_k$$

$$\theta' = \theta - \eta \frac{\partial L_{CE}}{\partial W_k}$$

Sama dengan *cross entropy* untuk fungsi 1{}, X_k berdasarkan matriks fungsi fitur untuk setiap k kelas, η *learning rate* nilai bobot jika terlalu tinggi nilainya maka proses pelatihan akan bobot θ akan terlalu besar dan melewat nilai *cross entropy*. Jika terlalu kecil nilainya maka membuat proes pelatihan menjadi sangat lambat dan perubahan bobot yang sangat kecil. θ' perbaruan bobot [18]. Proses untuk menentukan urutan kelas berdasarkan data observasi disebut *decoding* dan sebagai metode pengujian. Pada penelitian ini menggunakan algoritma Viterbi untuk *decoder*-nya. Berikut ini rumus (2.7) untuk proses pengujian Viterbi [18]:

$$arg\max_{T} P(T|W) = \prod_{i=1} \mathrm{P}(t_i|w_1 \dots w_n, t_i\,; \theta)$$

1. Inisalisasi:

$$\pi(*,*\,,0) = 1$$

2. Rekursi:

Untuk setiap kelas $j\epsilon\{1 \dots k\}$, untuk setiap token $t\ \epsilon\ [1 \dots n]$ dan untuk kelas sebelumnya $i\ \epsilon\ [1 \dots k]$

$$v_t(j) = \max_{i=1\dots k}(v_{t-1}(i) * \mathrm{P}(\mathrm{t}_j|w_1 \dots w_t, t_i\,; \theta))$$

3. Terminasi:

Nilai tertinggi : $\mathrm{P}^* = \mathrm{argmax}_{i\ \epsilon 1\dots k}\ v_t(i)$

2.8. Hasil Pengujian

Analisis rencana pengujian dilakukan dengan melalui dua tahapan, diantaranya tahapan token-kelas
Akan dihitung tingkat kebenaran dari klasifikasi yang telah dilakukan oleh sistem, sehingga menghasilkan nilai akurasi dan error.

2.8.1. Pengujian Token-Kelas

Setiap token kata, angka, dan simbol yang telah memiliki kelas divalidasi kebenarannya, lalu dihitung tingkat kebenaran dan kesalahannya, sehingga menghasilkan nilai akurasi dan error. Berikut hasil klasifikasi pengujian token-kelas pada tabel 3.

Tabel 3 Hasil Pengujian Token-Kelas

No	Nama Dokumen	Akurasi	Error
1	1.csv	77%	23%
2	2.csv	13%	88%
3	3.csv	0%	100%
4	4.csv	0%	100%
5	5.csv	13%	88%
6	6.csv	6%	94%
7	7.csv	19%	81%
8	8.csv	0%	100%
9	9.csv	19%	81%
10	10.csv	6%	94%
11	11.csv	6%	94%
12	12.csv	0%	100%
13	13.csv	0%	100%
14	14.csv	6%	94%
15	15.csv	6%	94%
16	16.csv	13%	88%
17	17.csv	0%	100%
18	18.csv	6%	94%
19	19.csv	6%	94%
20	20.csv	13%	88%
21	21.csv	0%	100%
22	22.csv	0%	100%
23	23.csv	0%	100%
24	24.csv	0%	100%
25	25.csv	6%	94%
26	26.csv	0%	100%
27	27.csv	0%	100%
28	28.csv	6%	94%
29	29.csv	13%	88%
30	30.csv	13%	88%
31	31.csv	6%	94%
32	32.csv	6%	94%
33	33.csv	6%	94%
34	34.csv	6%	94%
35	35.csv	6%	94%
36	36.csv	0%	100%
37	37.csv	13%	88%
38	38.csv	13%	88%
39	39.csv	13%	88%
40	40.csv	13%	88%
Rata-Rata		6%	94%

2.9. Analisis Hasil Pengujian

Nilai tinggi rendahnya akurasi dan error yang diperoleh memiliki beberapa penyebab. Berikut analisis yang dilakukan terhadap hasil pengujian akurasi dan error dengan konsep token-kelas. Analisis hasil pengujian dengan konsep token-kelas Karena data masukan yang digunakan oleh sistem adalah file teks, maka terdapat keterbatasan pada fitur yang digunakan, kenaikan ataupun penurunan nilai akurasi dan error disebabkan oleh kualitas *file PDF* yang dihasilkan pada proses *scanning* sebelumnya, nilai akurasi yang rendah diperoleh karena pada dokumen *testing* terdapat beberapa simbol tidak beraturan yang dihasilkan oleh proses konversi sebelumnya, berdasarkan

pengamatan yang dilakukan, dampak yang mempengaruhi pada penurunan nilai akurasi selalu disebabkan oleh fitur *ORGANIZATION* yang dikhususkan untuk memberikan pembobotan pada kategori Program Studi, Fakultas, dan Universitas. Karena, pada algoritma *MEMM* memiliki jaringan kompetitif, maka jika Program Studi diasumsikan menang dengan pembobotan yang sama antara Fakultas dan Universitas, Fakultas dan Universitas akan terklasifikasikan pada kelas yang sama seperti Program Studi, Minimnya penggunaan ekstraksi fitur yang dapat mengenali beberapa token secara spesifik terhadap banyaknya jumlah kelas, yaitu 17 kelas, sehingga untuk memberikan bobot pada setiap token tidak signifikan.

3. PENUTUP

Berdasarkan pengujian fungsionalitas sistem dan pengukuran akurasi yang telah dilakukan, sistem ekstraksi informasi menggunakan algoritma Maximum Entropy Markov Model (MEMM) telah berhasil dibangun dengan perolehan akurasi token-kelas sebesar 77%. Berikut penyebab terhadap akurasi yang diperoleh. Perolehan akurasi token-kelas disebabkan oleh penggunaan fitur pembobotan yang tidak signifikan, sehingga algoritma MEMM tidak dapat melakukan klasifikasi secara benar

Pada penelitian ini masih memiliki beberapa kekurangan, sehingga nilai akurasi pada konsep token-kelas maupun kelas-token tidak dapat diperoleh dengan maksimal. Dengan demikian, beberapa saran akan dipaparkan untuk pengembangan lebih lanjut mengenai sistem ekstraksi informasi menggunakan machine learning, diantaranya menggunakan data masukan dengan format .docx atau .html agar dapat menggunakan beberapa fitur tambahan seperti mendeteksi bold, italic, underline, dan beberapa font style lainnya. Diperlukan suatu fitur tambahan yang signifikan untuk pembobotan kategori Fakultas, Program Studi, dan Universitas. Hal tersebut dilakukan agar ketika proses testing dilakukan, algoritma MEMM dapat menentukan kelas yang tepat untuk Fakultas, Program Studi, atau Universitas. Perlu dibuktikan algoritma *MEMM* dengan perbandingan parameter dalam melakukan ekstraksi informasi pada dokumen karya tulis ilmiah. Dibutuhkan pengecekan dan pengkoreksian kesalahan ejaan atau typo pada hasil konversi *file PDF*.

DAFTAR PUSTAKA

[1] A. I. Riaddy, S. M. Yulianti Sibaroni and S. M. Annisa Aditsania, "Ekstraksi Informasi pada Makalah Ilmiah dengan Pendekatan Supervised Learning," *e-Proceeding of Engineering*, vol. 3, no. 1, pp. 1184-1190, 2016.

[2] T. Poibeau, H. Saggion, J. Piskorski and R. Yangarber, Multi-Source, Multilingual Information Extraction and Summarization, Heidelberg: Springer, 2013.

[3] D. Mustaqwa and N. Indriani, "Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Ruled Based," 2017.

[4] S. Mengel and Y. Jing, "Extracting Structured Data from Web Pages with Maximum Entropy Segmental Markov Model," *WISE*, no. LNCS 5802, pp. 219-226, 2009.

[5] S. S, Jiljo and P. P. V, "A Study On Named Entity Recognition For Malayalam Language Using TnT Tagger & Maximum Entropy Markov Model," *International Journal of Applied Engineering Research ISSN 0973-4562*, vol. 11, no. 8, pp. 5425-5429, 2016.

[6] A. T. Nedjo, D. Huang and X. Liu, "Automatic Part-of-speech Tagging for Oromo Language Using Maximum Entropy Markov Model (MEMM)," *Journal of Information & Computational Science* , vol. 11, no. 10, pp. 3319-3334, 2014.

[7] P. D. Sugiyono, Metode Penelitian Kombinasi (Mixed Methods), Bandung: Alfabeta, 2014.

[8] I. Sommerville, Software Engineering, 9th ed., Pearson, 2011.

[9] Tellez-Valero, Alberto, M. Montes-y-Gomez and L. V. Pineda, "A Machine Learning Approach to Information Extraction," *International Conference on Intelligent Text Processing and Computational Linguistics CICLing*, pp. 539-547, 2005.

[10] C. Siefkes, "An Overview and Classification of Adaptive Approaches to Information Extraction," *Lecture Notes in Computer Science 3730, Journal Semantics IV Springer-Verlag Berlin Heidelberg,* pp. 510-521, 2005.

[11] C. Siefkes, "An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models," 16 Februari 2007. [Online]. Available: http://www.siefkes.net/talks/disputation-ie.pdf. [Accessed April 2018].

[12] P. M. Prihatini, "Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia," *Jurnal Matrix,* vol. 6, no. 3, pp. 174-178, 2016.

[13] M. Indrawijaya, Liliana and R. Adipranata, "Aplikasi Ekstraksi Fitur Citra Hufur Jawa Berdasarkan Morfologinya," *Jurnal Infra,* vol. 3, no. 1, pp. 260-266, 2015.

[14] F. Peng and A. McCallum, "Accurate Information Extraction from Research Papers Using Conditional Random Fields," *Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL),* 2004.

[15] E. Setiawan, "Kamus Besar Bahasa Indonesia," Kemdikbud Pusat Bahasa, 2018. [Online]. Available: https://kbbi.web.id. [Accessed 05 2018].

[16] P. D. E. Z. Arifin, Dasar-Dasar Penulisan Karya Ilmiah, Jakarta: Grasindo.

[17] B. Liu, Web Data Mining, University of Illinois, Chicago: Springer, 2011.

[18] Z. Markov and D. T. Larose, Data Mining The Web : Uncovering Patterns In Web Content, Structure, and Usage, New Jersey: John Wiley & Sons, Inc, 2007.

[19] D. Jurafsky and J. H. Martin, "Speech and Language Processing 3rd ed. draft," 28 August 2017. [Online]. Available: http://web.stanford.edu/~jurafsky/slp3/. [Accessed April 2018].

[20] D. L. Olson and D. Delen, Advanced Data Mining Techniques, Berlin: Springer, 2008.

[21] K. K. Purnamasari and I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," *IOP Conference Series: Materials Science and Engineering,* vol. 407, no. 012151, pp. 1-4, 2018.

[22] S. Iqbal, S. A. Qureshi, T. H. Rizvi, G. Abbas and M. M. Gulzar, "Concept Building Through Block Diagram," *Journal of The Institution of Electrical and Electronics Engineers Pakistan,* Vols. 66-67, pp. 30-34, 2010.

[23] M. Fowler, UML Distilled Edisi 3, Yogyakarta: ANDI, 2005.

[24] A. Kadir, Dasar Pemrograman Java 2, Yogyakarta: Andi, 2004.