EKSTRAKSI INFORMASI PADA KARYA TULIS ILMIAH DENGAN GENERALIZED HIDDEN MARKOV MODEL

Bhakti Junrio[1], Ken Kinanti Purnamasari[2]

[1,2] Universitas Komputer Indonesia
Jl. Dipati Ukur No.122-116, Lebakgede, Coblong, Kota Bandung, Jawa Barat 20132
E-mail : bhaktijunrio15@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRAK

Ekstraksi informasi dapat diterapkan pada berbagai macam kasus, salah satunya untuk mendeteksi setiap komponen yang ada pada dokumen karya tulis ilmiah. Dokumen karya tulis ilmiah biasanya memiliki format yang cukup beragam. Keberagaman itu menyebabkan sulit untuk mendeteksi setiap komponen-komponen yang ada di dalamnya. Pada kasus dokumen karya tulis ilmiah yang berkaiatan dengan ekstraksi informasi telah banyak dilakukan penelitian sebelumnya, salah satunya pada dokumen karya tulis ilmiah skripsi berbahasa Indonesia. Pada penelitian tersebut dilakukan pengujian terhadap 50 dokumen yang dibuat dari tahun 2013 dan juga 2017, diperoleh rata-rata akurasi sebesar 57%. Penelitian ini dilakukan menggunakan algoritma *generalized hidden markov model (GHMM)*. Penelitian sebelumnya yang telah dilakukan menggunakan algoritma *GHMM* pada kasus ekstraksi informasi web dan pengenalan DNA, *GHMM* memiliki akurasi yang unggul dari algoritma lainnya. Penerapan *GHMM* pada ekstraksi informasi dokumen karya tulis ilmiah berdasarkan dari pengujian yang telah dilakukan terhadap 40 data dokumen karya tulis ilmiah skripsi, didapat nilai rata-rata akurasi sebesar 64%. Perolehan akurasi token-kelas disebabkan adanya kesalahan ejaan atau typo dari konversi dokumen dan juga penggunaan fitur pembobotan.

Kata kunci : Ekstraksi Informasi, *Generalized Hidden Markov Model, GHMM*, Dokumen karya tulis ilmiah, Skripsi.

1. PENDAHULUAN

Ekstraksi informasi adalah suatu bidang ilmu untuk pengolahan data yang bertujuan untuk mengubah suatu data menjadi informasi yang terstruktur [1]. Tujuan lain dari ekstraksi informasi adalah untuk mendapatkan fakta-fakta dari suatu kejadian, entitas dan keterhubungan dari kumpulan data [2] data yang diolah dapat berupa teks atau juga beragam data lainnya. Pendekatan yang dilakukan pada kasus ekstraksi informasi salah satunya menggunakan *machine learning*, karena dalam proses pengembangannya tidak memerlukan banyak waktu namun diperlukannya model dan juga data latih yang besar.

Pada penelitian ini, ekstraksi informasi dilakukan pada dokumen karya tulis ilmiah skripsi, dimana terdapat beberapa dokumen yang diambil yaitu dokumen cover dan abstrak. Penelitian sebelumnya yang telah dilakukan dengan kasus ekstraksi informasi dokumen karya tulis ilmiah skripsi berbahasa Indonesia. Pada penelitian tersebut dilakukan dengan mengambil cover dan abstrak dari dokumen karya tulis ilmiah skripsi berbahasa indonesia dan didapatkan nilai akurasi dari pengujian 3 dokumen yaitu 100%. Sedangkan untuk pengujian terhadap 50 dokumen didapatkan akurasi rata-rata sebesar 57% [3]. Penelitian yang sama tetapi dengan menggunakan algoritma yang berbeda yaitu *LVQ*, didapatkan akurasi rata-rata sebesar 39% yang dilakukan terhadap 40 buah data dokumen [4].

Penelitian dengan kasus ekstraksi informasi pada karya tulis ilmiah skripsi berbahasa Indonesia menggunakan algoritma *GHMM* belum pernah diterapkan sebelumnya. Pada jurnal berbahasa inggris algoritma *GHMM* pernah diterapkan pada kasus selain ekstraksi informasi dan juga pada kasus ekstraksi informasi. Pada kasus ekstraksi informasi yaitu pada web informasi ekstraksi dengan pengujian sebuah halaman website didapatkan akurasi sebesar 85.5% [5]. Sedangkan pada kasus pengenalan gen manusia yang ada pada DNA didapatkan akurasi sebesar 85% [6].

Berdasarkan penelitian diatas diketahui bahwa algoritma *GHMM* dapat diterapkan pada berbagai macam kasus. Belum adanya penggunaan algoritma *GHMM* untuk kasus ekstraksi informasi pada dokumen karya tulis ilmiah skripsi berbahasa Indonesia, menjadi pertimbangan dimungkinkan nya untuk dilakukan pengujian. Sehingga penelitian ini bertujuan untuk membangun suatu sistem ekstraksi informasi pada dokumen karya tulis ilmiah skripsi berbahasa Indonesia dan juga untuk mengetahui tingkat akurasi yang dihasilkan dari kasus ekstraksi informasi dokumen karya tulis ilmiah skripsi berbahasa Indonesia menggunakan algoritma *GHMM*. Batasan dokumen yang digunakan pada penelitian ini adalah dokumen karya tulis ilmiah

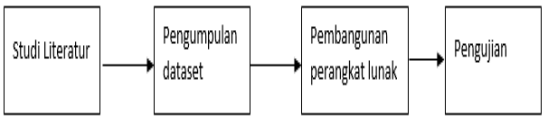
skripsi pada Universitas Komputer Indonesia (UNIKOM).

2. ISI PENELITIAN

Pada bagian dari isi penelitian ini menjelaskan beberapa hal yang meliputi metode penelitian, dokumen karya tulis ilmiah skripsi, arsitektur sistem, tokenisasi, segmentasi data, ekstraksi fitur, algoritma *GHMM*, akurasi dan error, dan juga hasil pengujian.

2.1 Metode Penelitian

Tahapan metode penelitian yang digunakan dalam penelitian ini adalah metode kuantitatif. Adapun Tahapan-tahapan yang dilakukan disini yaitu mulai dari studi literatur, pengumpulan dataset, pembangunan perangkat lunak dan juga pengujian. Blok diagram metode penelitian tersaji pada gambar berikut.



Gambar 1. Blok Diagram Metode Penelitian

2.2 Dokumen Karya Tulis Ilmiah

Dokumen adalah tulisan atau laporan tertulis yang memaparkan hasil penelitian atau pengkajian suatu masalah yang sifatnya tertulis ataupun tercetak dan dapat dipakai sebagai bukti karya tulis ilmiah [7]. karya ilmiah mesti memenuhi kaidah dan etika keilmuan yang dikukuhkan dan di taati oleh masyarakat. Data, simpulan, dan informasi lain yang terkandung dalam karya tulis ilmiah tersebut dijadikan acuan (referensi) bagi ilmuan lain dalam melaksanakan penelitian atau pengkajian selanjutnya.

Skripsi adalah karya tulis ilmiah yang mengemukakan pendapat penulis berdasarkan pendapat orang lain yang didukung oleh data dan fakta untuk melengkapi syarat guna memperoleh gelar sarjana dari suatu perguruan tinggi [8].

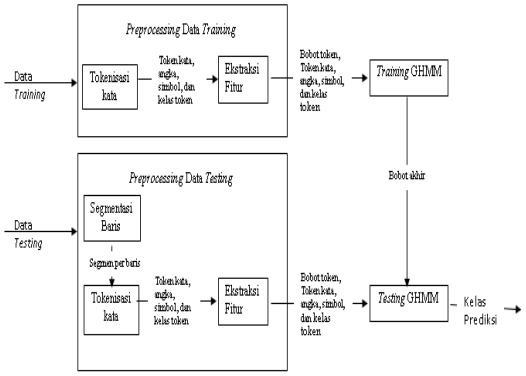
Penelitian yang akan dibuat ini menggunakan dokumen karya tulis ilmiah skripsi pada lembar cover dan abstrak sebagai data yang bertujuan untuk mengambil informasi yang ada didalamnya, karena didalam dokumen karya tulis ilmiah terdapat berbagai kategori yang bermacam-macam. Pada lembar cover memiliki kategori judul penelitian, jenis penelitian, kalimat pengajuan, nama, nim, program studi, fakultas, universitas, dan juga tahun. Sedangkan pada abtrak terdapat judul penelitian(abstrak), oleh, nama, nim, isi abstrak, dan juga kata kunci.

2.3 Natural Language Processing (NLP)

Natural language processing atau yang biasa dikenal NLP banyak sekali tools nya. NLP pada banyak kasus sangat penting untuk mendapatkan akurasi yang baik. Seperti pada pendeteksian kata-kata dasar stemming dan deteksi jenis kata (Penandaan PO). Pada banyak kasus penggunaan NLP banyak di lakukan salah satunya penelitian tentang stemming dan penandaan POS. dalam penenganannya NLP terkadang mengalami masalah salah satunya pada kasus di atas adalah kurangnya corpus di Indonesia serta tidak lengkapnya aturan yang tersedia penelitian tentang stemming, pertama kali diterbitkan oleh Julia Beth Lovins pada tahun 1968 [11].

2.4 Arsitektur sistem

Sistem ekstraksi informasi menggunakan *GHMM (Generalized Hidden Markov Model*) memiliki beberapa proses mulai dari input data masukan traning dan testing, preprocessing data training dan testing yang mana didalamnya terdapat proses segmentasi baris, tokenisasi kata, ekstraksi fitur. Hasil dari preprocessing data training akan digunakan pada testing pada *GHMM*. Adapun gambaran dari sistem ekstraksi informasi menggunakan GHMM adalah sebagai berikut.



Gambar 1. Arsitektur Sistem

Gambar 2 adalah gambaran sistem pada penelitian ini. Tahapan Preprocessing data training yang diperlukan adalah data yang berupa file bertipe .CSV. File .CSV berisi kumpulan data yang telah diberi label dari setiap kategori yang ada. Tokenisasi digunakan untuk memisahkan setiap kata yang ada pada kalimat menjadi token-token yang dapat diberikan bobot melalui proses ekstraksi fitur.

Tahapan Preprocessing data testing, yang diperlukan adalah data yang berupa file bertipe .TXT. File .TXT akan diolah menjadi tahapan segmentasi baris, tokenisasi, dan ekstraksi fitur. Pada segmentasi baris setiap kata akan dibagi sesuai baris-baris yang dibentuk dengan array. Tokenisasi membagi atau memisahkan setiap kata dari hasil segmentasi data. Ekstraksi fitur merupakan pemberian fitur pada setiap token kata.

2.5 Tokenisasi

Tokenisasi adalah suatu pemotongan atau pemisahan string dari setiap kalimat yang akan dipecah menjadi kata-kata tunggal yang berdiri sendiri. Pemisahan kalimat menjadi kata per kata dengan cara memanfaatkan spasi. Berikut adalah contoh dari tokenisasi.

Tabel 1. Tokenisasi

Teks masukan	kata
SISTEM PEMETAAN GARDU LISTRIK DI PLN UPJ CILEUNGSI BERBASIS DEKSTOP	SISTEM
	PEMETAAN
	GARDU
	LISTRIK
	DI
	PLN
	UPJ
	CILEUNGSI
	BERBASIS
	DEKSTOP

2.6 Segmentasi Data

Segmentasi data bertujuan untuk menghapus baris dan kolom sehingga menghasilkan data segmen yang telah menjadi perbaris. Segmentasi data dilakukan pada tahapan preprocessing data testing yang masukannya berupa file teks berupa txt. Berikut blok diagram dari segmentasi data.

Tabel 2. Segmentasi Data

Kalimat	Hasil Segmentasi
SISTEM PEMETAAN GARDU LISTRIK DI PLN UPJ CILEUNGSI BERBASIS DEKSTOP SKRIPSI	SISTEM PEMETAAN GARDU LISTRIK DI PLN UPJ CILEUNGSI BERBASIS DEKSTOP SKRIPSI

2.7 Ekstraksi Fitur

Adalah suatu pemberian bobot nilai pada fitur yang terkandung pada setiap token kata dari hasil tokenisasi.

Ekstraksi fitur pada penelitian ini adalah untuk memberikan ciri khusus pada suatu token kata sehingga nantinya dapat dikenali oleh sistem dan akan diberikan nilai bobot pada kata tersebut. Pemilihan fitur yang tepat dapat meningkatkan akurasi dalam pelabelan.

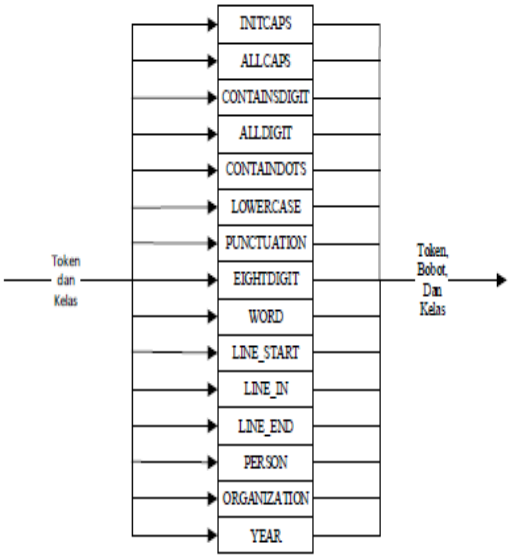
Ekstraksi fitur yang digunakan pada penelitian ini sebanyak 15 buah fitur, merujuk pada penelitian ekstraksi informasi dokumen karya tulis ilmiah menggunakan LVQ oleh Firdamdam [4]. Adapun penjelasan ekstraksi fitur sebagai berikut.

.

Tabel 3. Ekstraksi Fitur

No	Nama Fitur	Keterangan
1	INITCAPS	Mengenali setiap token yang hurufnya diawali dengan kapital.
2	ALLCAPS	Mengenali setiap token yang semua hurufnya kapital.
3	CONTAINSDIGIT	Mengenali setiap token yang mengandung digit.
4	ALLDIGIT	Mengenali setiap token yang semuanya digit.
5	CONTAINSDOTS	Mengenali setiap token yang mengandung titik
6	LOWERCASE	Mengenali setiap token yang semuanya huruf kecil.
7	PUNCTUATION	Mengenali setiap token yang mengandung tanda tertentu seperti titik, koma, titik dua, titik koma, tanda kurung, dan tanda seru.
8	EIGHTDIGIT	Fitur tambahan pada penelitian ini, fitur ini dikhususkan untuk mengenali token yang memiliki digit dengan panjang 8 digit
9	WORD	Fitur tambahan pada penelitian ini, fitur ini dikhususkan untuk memberikan bobot pada token untuk kela JENIS_PENELITIAN ,KALIMAT_PENGAJUAN.
10	LINE_START	Mengenali posisi token pada indeks array awal.
11	LINE_IN	Mengenali posisi token pada indeks array tengah.
12	LINE_END	Mengenali posisi token pada indeks array akhir.
13	PERSON	Mengenali token nama seseorang
14	ORGANIZATION	Mengenali token sebuah organisasi
15	YEAR	Mengenali ciri token tahun.

Setiap fitur yang ada akan diberi nilai bobot 1 atau 0, jika token kata tersebut termasuk salah satu fitur diatas akan diberi bobot 1, jika tidak akan diberi bobot 0. Berikut adalah gambaran ekstraksi fitur data training dan ekstraksi fitur data testing.

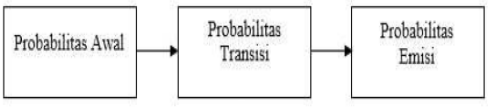


Gambar 3. Blok Digaram Ekstraksi Fitur

2.8 Algoritma GHMM

Generalized HMM termasuk pengembangan dari *HMM (hidden markov model).* Generalized hidden markov model adalah sebuah metode yang berbasi probabilitas interval yang mana probabilitas interval tersebut digunakan untuk menangkap ketidakpastian dari perhitungan yang dihasilkan oleh *HMM* biasa. Hasil yang diberikan oleh *GHMM* pada beberapa kasus yang kompleks dapat meningkatkan tingkat akurasi yang dihasilkan karena pada *GHMM* dilakukan penggabungan kesalahan yang ada selama pengumpulan data untuk tahap pembelajarannya dan pengenalan fitur.

Tahapan training pada *GHMM* mebutuhkan data masukan kelas dan fitur dan token kata. Awal permulaan dari training yaitu mencari parameter dari *GHMM*, terdapat 3 perameter utama yang harus diketahui nilainya terlebih dahulu yaitu yang termasuk kedalam parameter dari **λ = (A, B, π).** Secara umum langkah-langkah pada prosesnya digambarkan seperti Gambar 4.



Gambar 4. Skema Probabilitas GHMM

2.8.1 Training GHMM

1. Probabilitas Awal

Probabilitas awal adalah suatu peluang yang terdapat pada suatu keadaan tertentu sebagai awal dari suatu keadaan. Dimana π_i merupakan hasil nilai dari probabilitas q_i yaitu kondisi tersembunyi pada kata ke i, dan S_i jumlah keseluruhan kelas yang ada sampai kelas ke i. berikut persamaan pada 2.1 [9].

$$\pi_i = P\,(q_i = S_i\ ) \qquad (2.1)$$

Dimana :

π_i = Probabilitas awal dari kata di awal kelas ke –i

i = Kata di awal kelas ke-i

P = probabilitas

q_i = Kondisi tersembunyi pada kata di awal kelas ke-i

S_i = Jumlah keseluruhan kelas sampai ke-i

2. Probabilitas Transisi

Adalah Probabilitas dari peluang suatu keadaan bila berpindah dari kelas pertama ke kelas berikutnya. q_j adalah kondisi kelas selanjutnya yang tersembunyi pada saat q_i kondisi kelas saat ini. Probabilitas Transisi dapat dilambangkan dengan A={aij}. berikut persamaan pada 2.2 [9].

$$a_{ij} = P\,(q_j \text{ Pada } t+1 \mid q_i \text{ Pada } t) \qquad (2.2)$$

Dimana :

a_{ij} = probabilitas transisi dari i ke j

i = adalah kelas saat ini

j = adalah keadaan selanjutnya

t = merupakan state yang akan di lalui

P = adalah probabilitas

q_j = kondisi tersembunyi pada keadaan selanjutnya

q_i = kondisi tersembunyi pada keadaan saat ini

t+1 = adalah urutan indeks ke-t+1

3. Probabilitas Emisi

Adalah Probabilitas dari suatu peluang distribusi dari keadaan observasi yang mana menghasilkan suatu keadaan baru. Emisi didapatkan dari jumlah kemunculan fitur yang ada pada kelas dan dibagi dengan jumlah keseluruhan dari token kata yang mengandung fitur yang dicari. Misalkan pada emisi akan dilihat kondisi dari tiap token kata yang ada dari kelas ke 1 sampai 16, apakah termasuk kedalam allcaps jika termasuk maka diberi nilai bobot 1, lakukan pengecekan sampai kelas awal telah semua diproses dan dijumlahkan, lalu dibagi dengan jumlah token yang mengandung allcaps pada semua kelas.
Di dalam GHMM probabilitas emisi dilambangkan dengan B= (b_(I) (O_(t))). Berikut persamaan 2.3 [9].

$$b_j\,(O_t) = P(\,O \text{ pada } t \mid q_j \text{ pada } t) \qquad (2.3)$$

Dimana :

$b_j\,(O_t)$ = Probabilitas emisi ke –j pada keadaan observersi Ke-t

P = adalah probabilitas

O = adalah kondisi observasi

t = adalah indeks kata ke -t

q_j = adalah fiturnya

4. Observasi Sequence, Upper dan Lower

Pada Observasi Sequence akan dihitung nilai keseluruhan dari hasil probabilitas awal, transisi dan juga emisi. λ merupakan penampung dari nilai yang telah dicari sebelumnya yaitu pada π, B, A pada persamaan 2.1, 2.2 dan 2.3. sedangkan O adalah observasi sequence peluang muncuknya barisan observasi untuk π, B, A tersebut, dapat dilihat pada persamaan 2.4 [9]. Observasi upper merupakan batas atas dari elemen yang terkandung dari hasil pengamatan λ yang merupakan nilai dari probabilitas awal, transisi dan juga emisi, sedangkan observasi lower merupakan batas bawah dari elemen yang terkandung dari hasil pengamatan λ, dapat dilihat pada persamaan 2.4, 2.5, 2.6 [9].

Observation Sequence $P(O \mid \lambda)$ (2.4)

Observation Sequence upper $(\bar{O} \mid \lambda)$ (2.5)

Observation Sequence lower $(\underline{O} \mid \lambda)$ (2.6)

Dimana :
O = Observasi
λ = nilai dari probabilitas awal, transisi dan emisi
$\bar{O}$ = adalah observasi upper
$\underline{O}$ = adalah observasi upper

5. Generalisasi Probabilitas Awal

Generalisasi probabilitas awal disini untuk mencari nilai batas bawah dan batas atas dari elemen yang terkandung pada hasil yang didapatkan dari perhitungan probabilitas awal dari persamaan 2.1 [9]. Adapun persamaan untuk lower dan upper pada 2.7 dan 2.8 [9] :

$$\overline{\pi_i} = \overline{P}\,(\,q_1 = S_i\,)\;(\,1 \le i\;\; N \le) \quad (2.7)$$

$$\underline{\pi_i} = \underline{P}\,(\,q_1 = S_i\,)\;(\,1 \le i\;\; N \le) \quad (2.8)$$

Dimana :
$\overline{\pi_i}$ = probabilitas awal upper
$\underline{\pi_i}$ = probabilitas awal lower
P = probabilitas
q_i = Kondisi tersembunyi pada kata di awal kelas ke-i
S_i = jumlah keseluruhan kelas sampai ke-i

6. Generalisasi Probabilitas Transisi

Generalisasi probabilitas transisi disini untuk mencari nilai batas bawah dan batas atas dari elemen yang terkandung pada hasil yang didapatkan dari perhitungan probabilitas transisi dari persamaan 2.2 [9]. Adapun persamaan untuk lower dan upper pada 2.9 dan 2.10 [9] :

$$A = a_{ij=}\,p\,(q_{t+1} = S_j \mid q_t = S_i) \quad (1 \le i,j \le N) \quad (2.9)$$
$$a_{ij} \ge 0, \sum_{j=1} a_{ij} = 1\;\; (1 \le i,j \le N)$$

$$\overline{A} = \overline{a_{ij=}}\,p\,(q_{t+1} = S_j \mid q_t = S_i) \quad (1 \le i,j \le N) \quad (2.10)$$
$$a_{ij} \ge 0, \sum_{j=1} a_{ij} = 1\;\; (1 \le i,j \le N)$$

Dimana :
$\overline{a_{ij}}$ = nilai generalisasi upper dari probabilitas transisi
$\underline{a_{ij=}}$ = nilai generalisasi lower dari probabilitas transisi
q_t = waktu hidden state +1
S_j = adalah kelas selanjutnya
S_i = adalah kelasnya

7. Generalisasi Probabilitas Emisi

Generalisai probabilitas emisi untuk mencari nilai batas bawah dan batas atas dari elemen yang terkandung pada hasil didapatkan dari perhitungan probabilitas emisi dari persamaan 2.3 [9]. Adapun persamaan untuk lower dan upper pada 2.11 dan 2.12 [9] :

$$\underline{B} = \underline{Bj}\,(k) = p\,(o_t = v_k \mid q_t = S_j) \quad (2.11)$$

$$\overline{B} = \overline{Bj}(k) = p\,(o_t = v_k \mid q_t = S_j) \quad (2.12)$$

Dimana :
$\underline{Bj}$ (K) = probabilitas lower emisi ke –j pada keadaan observersi Ke-t
$\overline{Bj}(k)$ = probabilitas upper emisi ke –j pada keadaan observersi Ke-t
P = adalah probabilitas upper
0 = adah kondisi observasi
V = adalah fitur pada kelasnya
Vk = adalah fitur selanjutnya

2.8.2 Testing GHMM

Algoritma Viterbi umumnya digunakan untuk menyelesaikan masalah yang berkaitan dengan pengkodean, dapat juga untuj menyelesaikan bidang yang lain. Algoritma Viterbi pada GHMM berguna untuk menemukan nilai barisan hidden state dari suatu barisan observasi yang nilainya paling maksimal atau optimal. Pada Viterbi terdapat beberapa tahapan, berikut adalah tahapan pada viterbi.

1. Inisialisasi

Tahapan pertama yaitu inisialisasi. Pada tahap ini akan dicari nilai awal yang dihasilkan dari Viterbi. Persamaan yaitu pada 2.13 dan 2.14 [9][10]. Inisialisasi lower dirumuskan dengan :

$$\delta_1^l(i) = \pi_i b_i(\underline{o_1}) \ (1 \leq \text{i} \leq \text{N}) \ (2.13)$$

Inisialisasi upper dirumuskan dengan :

$$\delta_1^u(i) = \pi_i b_i(\overline{o_1}) \ (1 \leq \text{i} \leq \text{N}) \ (2.14)$$

$$\psi_1(i) = 0 \ (1 \leq \text{i} \leq \text{N})$$

Dimana :

$\delta_1^l(i)$ = adalah kondisi inisialisai awal lower
$\delta_1^u(i)$ = adalah kondisi inisialisasi awal upper
π_i = adalah probabilitas awal
b_i = adalah probabilitas emisi
$\underline{O}$ = observasi lower
$\overline{O}$ = observasi upper

2. Induksi

Pada tahapan induksi akan dihitung hasil yang didapat dari inisialisasi, perhitungan dilakukan sebanyak kelas yang ada. Berikut adalah rumus mencari induksi. Berikut persamaanya 2.15 dan 2.16 [9] :

Induksi lower dirumuskan dengan :

$$\delta_1^1(j) = \max_{1 \leq i \leq N} \{\delta_{t-1}^1(i) a_{ij}\} \, b_j\left(\underline{o_t}\right)$$
$$(2 \leq \text{t} \leq \text{T}, \ 1 \leq \text{j} \leq \text{N}) \qquad (2.15)$$

Induksi upper dirumuskan dengan :

$$\delta_1^u(j) = \max_{1 \leq i \leq N} \{\delta_{t-1}^u(i) a_{ij}\} \, b_j(\overline{o_t})$$
$$(2 \leq \text{t} \leq \text{T}, \ 1 \leq \text{j} \leq \text{N}) \qquad (2.16)$$

$$\psi_1(j) = argmax \left\{ \begin{matrix} \min_{1 \leq i \leq N} (pro \ \delta_{t-1}^1(i) a_{ij}), \\ \min_{1 \leq i \leq N} (pro \ \delta_{t-1}^u(i) a_{ij} \end{matrix} \right\}$$

Dimana :

$\delta_1^l(j)$ = inisialisasi setiap hasil induksi lower
$\delta_1^u(i)$ = inisialisasi setiap hasil induksi upper
a_{ij} = adalah probabilitas transisi
b_i = adalah probabilitas emisi
$pro \ \delta_{t-1}^1$ = Penginisialisasian hasil dari induksi dikurangi 1
$\psi_1(j)$ = Menyimpan hasil induksi

3. Terminasi

Pada tahapan terminasi akan dilakukan untuk mendapatkan hasil atau nilai yang terbaik. Berikut adalah rumus mencari terminasi. Berikut persamaan pada 2.17 [9].

$$P^* = \max_{1 \leq i \leq N} \{\delta_T^l(i), \delta_T^u(i)\}, \qquad (2.17)$$

$$q_T^* = argmax \left\{ \min_{1 \leq i \leq N} (pro\delta_T^l(i)), \min_{1 \leq i \leq N} (pro\delta_T^u(i)) \right\}$$
,

Dimana :

P^* = Probabilitas dari nilai yang terbaik
$\max_{1 \leq i \leq N}$ = hasil nilai maksimal dari rekursi
q_T^* = Kondisi yang tersembunyi
$argmax$ = Isi yang dihasilkan dari nilaimaksimal

4. Backtracking

Pada tahapan backtracking, akan di cek kembali urutan yang terbaik yang dihasilkan dari terminasi. Berikut adalah perhitungan untuk Backtracking pada persamaan 2.18 [9].

$$Q_t^* = \psi_{t+1}(Q_{t+1}^*), t = T-1, T-2, \ldots, 1 \qquad (2.18)$$

Dimana :

Q_t^* = Kondisi Tersembunyi
$\psi_{t+1}(Q_{t+1}^*)$ = Matriks Penyimpanan state tersembunyi
q_T^* = Kondisi yang tersembunyi
T = Jumlah T dikurangi 1

2.9 Nilai Akurasi

Nilai akurasi diperoleh dengan mengukur nilai kebenaran token yang diklarifikasikan, dibagi dengan jumlah token, dikali 100%. Berikut rumus perhitungan Nilai akurasi.

$$Akurasi = \frac{Keseluruhan \ data \ benar}{Keseluruhan \ testing \ data} \ x \ 100\%$$

2.10 Nilai Error

Nilai error merupakan nilai yang diperoleh ketika beberapa kelas tidak terklarifikasi dengan benar. Berikut rumus perhitungan error.

$$Error = \frac{Keseluruhan \ data \ salah}{Keseluruhan \ testing \ data} \ x \ 100\%$$

2.11 Hasil Pengujian

Hasil pengujian dengan data sebanyak 40 yang dilakukan dengan konsep token kelas adalah sebagai berikut. Pengujian nilai akurasi dilihat pada tabel 4.

Tabel 4. Hasil Akurasi

No	Nama Dokumen	Akurasi
1	1.txt	65%
2	2.txt	77%
3	3.txt	43%
4	4.txt	72%
5	5.txt	58%
6	6.txt	66%
7	7.txt	66%
8	8.txt	73%
9	9.txt	72%
10	10.txt	53%
11	11.txt	72%
12	12.txt	72%
13	13.txt	65%
14	14.txt	52%
15	15.txt	72%
16	16.txt	80%
17	17.txt	52%
18	18.txt	55%
19	19.txt	63%
20	20.txt	71%
21	21.txt	53%
22	22.txt	56%
23	23.txt	74%
24	24.txt	60%
25	25.txt	69%
26	26.txt	65%
27	27.txt	61%
28	28.txt	66%
29	29.txt	55%
30	30.txt	74%
31	31.txt	57%
32	32.txt	54%
33	33.txt	73%
34	34.txt	60%
35	35.txt	64%
36	36.txt	68%
37	37.txt	69%
38	38.txt	57%
39	39.txt	69%
40	40.txt	70%
	Rata-rata	**64%**

2.12 Analisis Pengujian

Pengujian nilai akurasi dan error untuk memvalidasi antara token-kelas pada data pembanding dan data prediksi. Validasi dilakukan antara data_testing.csv dari hasil plotting diluar sistem sebagai data asli dan data hasil pengujian kelas-token sebagai data prediksinya. 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. Analisi hasil pengujian dengan konsep token-kelas :

1. Karena data masukan yang digunakan oleh sistem berupa file teks, maka terdapat keterbatasan pada fitur yang digunakan.
2. Nilai akurasi yang rendah, diperoleh karena pada dokumen testing terdapat beberapa simbol yang tidak beraturan yang dihasilkam oleh proses konservasi sebelumnya.
3. Berdasarkan pengamatan yang dilakukan, dampak yang mempengaruhi pada penurunan nilai akurasi selalu disebabkan oleh fitur ORGANIZATION yang dikhususkan untuk meberikan pembobotan pada kategori Program Studi, Fakultas, dan Universitas.

3. PENUTUP

Berdasarkan pengujian fungsionalitas sistem dan pengukuran akurasi yang telah dilakukan dengan menggunakan 40 data testing, system ekstraksi informasi menggunakan algoritma Generalized Hidden Markov Model (GHMM) telah berhasil dibagun dengan didapatkannya akurasi dari token-

kelas dengan akurasi benar sebesar 64% dan error sebesar 36%.

Pada penelitian ini masih memiliki beberapa kekurangan, sehingga nilai akurasi pada konsep token-kelas tidak dapat diperoleh dengan maksimal. Dengan demikian, beberapa saran akan dipaparkan untuk pengembangan kedepannya mengenai sistem informasi menggunakan machine learning, diantaranya :

1. Menggunakan data masukan dengan format .docx atau .html agar dapat menggunakan beberapa fitur tambahan seperti mendeteksi bold, italic, underline, dan beberapa font style lainnya.
2. Perlu dibuktikannya algoritma GHMM dengan perbandingan parameter dalam melakukan ekstraksi informasi pada dokumen karya tulis ilmiah.
3. Dibutuhkan pengecekan dan pengkoreksian untuk kesalahan ejaan atau typo pada hasil konversi PDF.

DAFTAR PUSTAKA

[1] D. Jurafsky dan J. H. Martin, "*Chapter* 21 - *Information Extraction*," dalam *Speech and Language Processing*, 2017, pp. 1 - 31.

[2] E. Susanti and K. Mustofa, *Ektraski Informasi Halaman Web Menggunakan Pendekatan bootstraping pada Ontology-Based Informasi Extraction*, vol. 9, pp. 111-120, 2015.

[3] D. Mustaqwa, "Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Rule Based," Skripsi, Universitas Komputer Indonesia, Bandung, Jawa Barat, 2018..

[4] F. Sasmita, "Ekstraksi Informasi Dokumen Karya Tulis Ilmiah Menggunakan Algoritma Learning Vector Quantiation," 2018.

[5] M. E. B. Prasetyo, "Teori Dasar Hidden Markov Model," 2011.

[6] K. David, H. David and E. F. H, "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," 1996

[7] S. Lestanti dan A. D. Susana, Sistem Pengarsipan Dokumen Guru Dan Pegawai Menggunakan Metode Mixture Modelling Berbasis WEB, vol. X, 2016

[8] A. Wahyuni, Aplikasi Kriptografi untuk Pengembangan E-Dokumen Dengan Metode Hybrid : Biometrik Tandatangan dan DSA (Digital Signature Algorithm), pp. 1-123,2011

[9] X. Y. Feng, H. M. You and W. yan, "A Generalized Hidden Markov Model and Its Applications In Recognition of Cutting States," 2016.

[10] K. David, H. David and E. F. H, "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," 1996.

[11] K. K. Purnamasari and I. S. Suwadi, "Rule-Based Part Of Speech Tagger For Indonesia Language," IOP Confrence Series : Materials Science and Engineering, vol.407, no. 012151, pp. 1-4, 2018.