Jurnal Mobile and Forensics (MF)
Vol. 2, No. 2, September 2020, pp. 71-83
P-ISSN: 2656-6257, E-ISSN: 2714-6685, DOI: https://doi.org/10.12928/mf.v2i2.2450

PELABELAN KELAS KATA PADA BAHASA JAWA DENGAN MENGGUNAKAN HIDDEN MARKOV MODEL

[1,*] Mohammad Mursyit, [2]Aji Prasetya Wibawa, [3]Ilham Ari Elbaith Zaeni, [4]Harits Ar Rosyid
[1234]Teknik Elektro, Universitas Negeri Malang
Jl. Semarang No. 5, (0341) 551312, Malang, Jawa Timur, Indonesia
e-mail: [1]mohammadmursyit@gmail.com, [2] aji.prasetya.ft@um.ac.id, [3]ilham.ari.ft@um.ac.id, [4]harits.ar.ft@um.ac.id
[*)] correspondence email

Abstrak

Part of Speech Tagging atau *POS Tagging* adalah proses memberikan label pada setiap kata dalam sebuah kalimat secara otomatis. Penelitian ini menggunakan algoritma *Hidden Markov Model* (HMM) untuk proses *POS Tagging*. Perlakuan untuk *unknown words* menggunakan *Most Probable POS-Tag*. *Dataset* yang digunakan berupa 10 cerita pendek berbahasa Jawa terdiri dari 10.180 kata yang telah diberikan *tagset* Bahasa Jawa. Pada penelitian ini proses *POS Tagging* menggunakan dua skenario. Skenario pertama yaitu menggunakan algoritma *Hidden Markov Model* (HMM) tanpa menggunakan perlakuan untuk *unknown words*. Skenario yang kedua menggunakan HMM dan *Most Probable POS-Tag* untuk *perlakuan unknown words*. Hasil menunjukan skenario pertama menghasilkan akurasi sebesar 45.5% dan skenario kedua menghasilkan akurasi sebesar 70.78%. *Most Probable POS-Tag* dapat meningkatkan akurasi pada *POS Tagging* tetapi tidak selalu menunjukan hasil yang benar dalam pemberian label. *Most Probable POS-Tag* dapat menghilangkan probabilitas bernilai Nol dari *POS Tagging Hidden Markov Model*. Hasil penelitian ini menunjukan bahwa *POS Tagging* dengan menggunakan *Hidden Markov Model* dipengaruhi oleh perlakuan terhadap *unknown words*, perbendaharaan kata dan hubungan label kata pada *dataset*.

Part of Speech Tagging or POS Tagging is the process of automatically giving labels to each word in a sentence. This study uses the Hidden Markov Model (HMM) algorithm for the POS Tagging process. Treatment for unknown words uses the Most Probable POS-Tag. The dataset used is in the form of 10 short stories in Javanese consisting of 10,180 words which have been given the Javanese tagset. In this study, the POS Tagging process uses two scenarios. The first scenario is using the Hidden Markov Model (HMM) algorithm without using treatment for unknown words. The second scenario uses HMM and Most Probable POS-Tag for treatment of unknown words. The results show that the first scenario produces an accuracy of 45.5% and the second scenario produces an accuracy of 70.78%. Most Probable POS-Tag can improve accuracy in POS Tagging but does not always produce correct labels. Most Probable POS-Tag can remove zero-value probability from POS Tagging Hidden Markov Model. The results of this study indicate that POS Tagging using the Hidden Markov Model is influenced by the treatment of unknown words, vocabulary and word label relationships in the dataset.

Kata Kunci: Bahasa Jawa, *Hidden Markov Model*, Kelas Kata, *Most Probable POS-Tag*, *POS Tagging*.

PENDAHULUAN

Bahasa sebagai identitas budaya bangsa yang digunakan untuk berkomunikasi dengan orang lain serta media dalam seni[1]. Bahasa merupakan alat komunikasi antar anggota masyarakat pemakai bahasa yang bersangkutan. Begitu pula bahasa Jawa, dalam kehidupan sehari-hari bahasa Jawa dipergunakan sebagai alat komunikasi masyarakat Jawa[2]. Bahasa Jawa adalah bahasa daerah terbesar di Indonesia[3]. Menurut catatan, jumlah penutur bahasa ± 80 juta orang, kira-kira 40% dari jumlah penduduk Indonesia. Orang jawa berasal dari Jawa Tengah dan Jawa Timur tetapi banyak orang jawa tinggal di Sumatera dan ditempat lain di Indonesia[4]. Beberapa wilayah persebaran Bahasa Jawa diluar negeri seperti Suriname, New Caledonia dan

Received July 1, 2020; Revised August 15, 2020; Accepted August 16, 2020

Pantai Barat Johor[5]. Persebaran Bahasa Jawa di luar Jawa Timur dan Jawa Tengah umumnya terjadi karena migrasi penduduk ke daerah daerah tersebut. Seiring perkembangan zaman penggunaan Bahasa Jawa mengalami penurunan

Penggunaan bahasa Jawa mengalami proses penurunan. Situasi kontak dengan bahasa Indonesia yang secara politis lebih dominan telah menyebabkan penurunan frekuensi pemakaian bahasa Jawa[6]. Selain itu juga disebabkan pergeseran penggunaan Bahasa Jawa pada masyarakat Multilingual[7]. Pada masyarakat yang Multilingual sering terjadi penggunaan variasi bahasa lain dan interferensi sehingga dapat mengiring bahasa kepada ambang kepunahan[8]. Maka dibutuhkan upaya untuk melestarikan kembali bahasa daerah.

Salah satu cara untuk melestarikan bahasa daerah terutama bahasa jawa yaitu dengan menerapkan teknologi informasi dalam pelabelan kelas kata atau (*Part of Speech Tagging*) pada bahasa Jawa. *Part of Speech Tagging* (*POS Tagging*) adalah suatu proses memberikan label kelas (anotasi) kata secara otomatis pada suatu kata dalam kalimat. Sebuah jenis kata (*Part of Speech*) dapat memberitahu tentang bagaimana kata tersebut dilafalkan. *POS Tagging* telah dikembangkan menggunakan metode statistik, aturan linguistik dan hibrida statistik dan aturan. Hasil penelitian *Part-of-Speech Tagging* pada dokumen dapat memberikan manfaat tidak hanya informasi tetapi juga masalah NLP seperti *Text Chungking, Syntactic Parsing, Semantic Role, Labeling* dan *Semantic Parsing*[9] dan digunakan sebagai dasar penelitian dalam *Natural Language Processing* lainnya, seperti *Language Generator, Information Retrieval, Text Summarization, Question and Answering*, dan *Machine Translation* [10].

Salah satu algoritma *POS Tagging* adalah *Hidden Markov Model*[11][12]. Metode ini memiliki kelebihan tingkat akurasi yang tinggi dan tidak memerlukan banyak pengetahuan tentang bahasa yang dikembangkan[13]. Berbagai penelitian sebelumnya *POS Tagging* dengan menggunakan *Hidden Markov Model* menunjukan hasil yang baik. Pada Bahasa Indonesia menunjukan hasil akurasi 98%[14], pada Bahasa Melayu 94%[15], dan pada Bahasa Bengali 95%[16]. Penggunaan POS Tagging dalam Bahasa Jawa pernah dilakukan penelitian berbasis aturan dan distribusi probabilitas *Maximum Entropy* untuk Bahasa Jawa Krama dengan menghasilkan akurasi tertinggi 97.67%[17]. Sedangkan *POS Tagging* pada Bahasa Jawa dengan menggunakan *Hidden Markov Model* belum pernah dilakukan. Hasil implementasi di masa mendatang dari rancangan algoritma ini dapat digunakan sebagai dasar penelitian dalam pemrosesan bahasa natural terutama yang berkaitan dengan Bahasa Jawa. Batasan pada penelitian ini hanya untuk menentukan label kelas kata pada setiap satu kata.

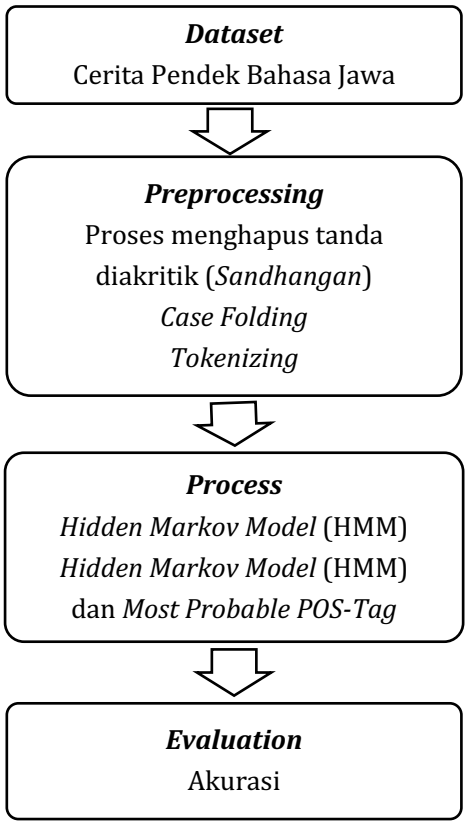
Artikel ini akan membahas tentang penerapan algoritma *POS Tagging Hidden Markov Model* untuk Bahasa Jawa. Dengan adanya penelitian ini diharapkan dapat memberikan pelabelan kelas kata yang lebih baik. Oleh karena itu artikel ini ditulis dengan sistematis untuk membahas metode penelitian, hasil pembahasan serta kesimpulan dan saran pengembangan di masa yang akan datang.

METODE PENELITIAN

Dalam penelitian *POS Tagging Hidden Markov Model* terdiri dari 4 tahapan yaitu (1) *Dataset*, (2) *Preprocessing*, (3) *Process*, (4) *Evaluation*.

Dataset berupa cerita pendek berbahasa Jawa yang dikumpulkan pada bulan Desember 2019 sampai Januari 2020 dari *website* ki-demang.com dan cerkak.com. *Website* dipilih karena tata bahasa yang digunakan telah sesuai dengan tata Bahasa Jawa. *Dataset* yang telah terkumpul dilakukan *preprocessing* untuk mengubah data yang belum terstruktur menjadi data terstruktur sesuai dengan kebutuhan pada penelitian.

Preprocessing terdiri dari tiga tahap yaitu menghapus tanda diakritik pada huruf vokal, merubah semua huruf menjadi huruf kecil (*Case Folding*), dan melakukan pemenggalan pada kalimat menjadi kata dengan simbol dan spasi sebagai pemenggalan (*Tokenizing*). Pada penelitian ini algoritma yang digunakan untuk *POS Tagging* yaitu *Hidden Markov Model* (HMM). Implementasi *Hidden Markov Model* digunakan untuk mencari probabilitas terbesar dari kemunculan label kelas kata yang dimiliki suatu kata dalam kalimat. Penentuan label kelas kata dengan HMM menghitung probabilitas dari semua kemungkinan label kelas kata dari kata yang dicari labelnya pada *dataset.* Probabilitas terbesar HMM berdasarkan banyaknya kata yang muncul pada label kelas kata yang dicari dengan total label kelas kata tersebut dan banyaknya hubungan label kelas kata yang dicari dengan label kelas kata sebelumnya. Hasil probabilitas label terbesar dari perhitungan HMM digunakan sebagai label kelas kata untuk kata yang dicari labelnya. Pada penelitian ini menggunakan dua skenario. Skenario yang pertama yaitu menggunakan algoritma *Hidden Markov Model* (HMM) tanpa menggunakan perlakukan untuk *unknown words*. Skenario yang kedua menggunakan *Hidden Markov Model* dan *Most Probable POS-Tag* untuk perlakuan *unknown words*. Setelah proses *POS Tagging* dilakukan tahap evaluasi dengan menghitung akurasi dari hasil pelabelan kelas kata (*Part of Speech Tagging*) menggunakan 10 *Cross Validation*. Alur penelitian ditunjukan pada Gambar 1.



Gambar 1. Desain Penelitian

Dataset

Dataset yang digunakan dalam penelitian ini berasal dari *website* ki-demang.com dan cerkak.com. *website* tersebut berisi karya berupa bahan/naskah, gambar atau

informasi berbahasa jawa. Dataset terdiri atas sepuluh cerita pendek berbahasa jawa dengan jumlah kata sebanyak 10.180 kata yang telah diberi label kelas kata Bahasa Jawa. Informasi terkait *dataset* yang digunakan ditunjukan pada Tabel 1.

Tabel 1. Rincian *Dataset*

No	Judul	Jumlah Kata	Jumlah Kata Unik	Jumlah Kalimat	Tag yang Digunakan
1	*Jangan Tumpang*	1130	543	61	15
2	*Dhuh Gusti*	1304	559	149	16
3	*Skripsi*	1247	610	93	15
4	*Goroh*	667	347	46	16
5	*Lemah*	740	348	53	15
6	*Guru*	1056	469	87	16
7	*Sketsa Klawu*	738	336	74	14
8	*Untu Palsu*	1245	634	78	17
9	*Kontraktor*	1021	512	101	15
10	*Pak Guru*	1032	565	81	17

Pada Tabel 1 memuat informasi rincian *dataset* yang digunakan pada penelitian ini. Baris jumlah kata unik memuat jumlah kata unik yang terdapat pada *dataset*. Kata unik adalah setiap kata yang muncul satu kali dalam *dataset* atau setiap kata yang muncul lebih dari satu kali dalam *dataset*, maka terhitung satu. Contoh kata "*adhem*" pada dokumen "Pak Guru" hanya muncul satu kali atau kata "*adoh*" muncul sebanyak 6 kali tetapi terhitung satu kali. Kelas kata adalah unsur kategorial yang merupakan tataran kedua yang tingkat keabstrakannya lebih rendah daripada fungsi [18]. Unsur kategorial yang dimaksud adalah kategori sintaksis, yakni klasifikasi satuan-satuan gramatikal berdasarkan bentuk, fungsi, serta perilakunya dalam sebuah konstruksi [19] [20]. Label kata atau tagset yang diberikan ke suatu kata dalam suatu kalimat menunjukkan kelas kata (*word class*) dari kata yang bersangkutan, dalam konteks kalimat tersebut. Kelas kata ini juga disebut sebagai *Part of Speech*. Kumpulan atau koleksi label atau *tag part of speech* atau kelas kata disebut sebagai *tagset* [11]. Pada Tabel 2 merupakan sampel kalimat *dataset* yang digunakan dalam penelitian ini. *Sampel dataset* telah diberikan label kelas kata Bahasa Jawa.

Tabel 2. *Sample Dataset*

Judul	Dataset
Jangan Tumpang	Parwati/TA kang/TP nomer/TKN loro/TW ora/TKG kalah/TKG karo/TP kangmase/TS ./TWP
Para Guru	Pak/TT Guru/TA Parjo/TA ,/PL lengkape/TKG bapak/TA doktorandus/TA Parjo/TA .TWP

Pada Tabel 3 merupakan sampel data uji yang digunakan dalam penelitian ini. Sampel data uji yang telah diberikan label kelas kata Bahasa Jawa.

Tabel 3. *Sample* Data Uji

Judul	Dataset
Guru	Jalaran/TP aku/TS iki/TS guru/TA ,/PL sanajan/TP guru/TA bantu/TKG ./TWP

Dataset pada setiap baris berisi kalimat dengan *Tagset* mengikuti setiap kata dipisahkan dengan tanda garis miring (/). *Tagset* setiap kata mengikuti makna kata dalam setiap kalimat sehingga memungkinkan satu kata memiliki *Tagset* yang berbeda. Pada Tabel 4 merupakan label kelas kata (*Tagset*) yang digunakan dalam penelitian ini. Label kelas kata Bahasa Jawa dengan menggunakan nama dari setiap istilah dalam Bahasa Jawa pada buku Paramasastra Bahasa Jawa [21].

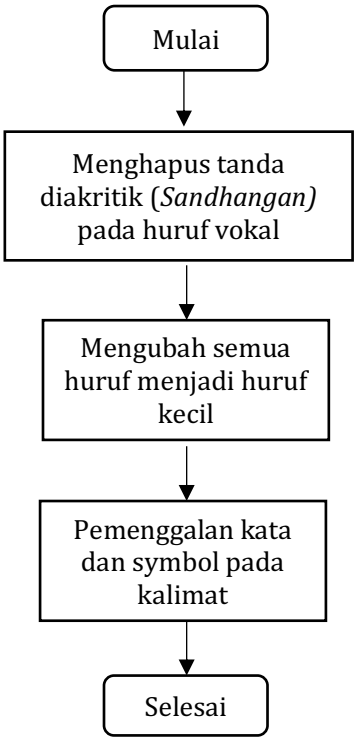
Tabel 4. *Tagset* Bahasa Jawa

No	Tag	Deskripsi	Contoh
1	TT	(Kata Sandang) (Tembung Tetenger)	Si Sang Ponang
2	TP	(Kata Hubung) (Tembung Penyambung)	Karo Banjur
3	PL	(Tanda Koma) (Pada Lingsa)	,
4	TW	(Bilangan) (Tembung Wilangan)	5 Siji-siji glintir
5	TM	(Kata Asing/Serapan) (Tembung Manca)	Dan Gara-gara efektif
6	TPP	(Kata Depan) (Tembung Pangarep)	Ing Saka
7	TKN	(Kata Sifat) (Tembung Kaanan)	Pinter Seneng Males
8	TA	(Kata benda) (Tembung Aran)	Buku Topi Meja
9	TS	(Kata Ganti) (Tembung Sesulih)	Aku Iki
10	TKG	(Kata Keterangan) (Tembung Katrangan)	Kerep Arang Wis
11	TG	(Kata Tugas) (Tembung Tugas)	Kok Mbok
12	SYM	(Simbol dan Notasi) (Simbol lan Notasi)	$ #
13	TPN	(Kata Perintah/Seru) (Tembung Penyeru)	Adhuh Lho
14	TK	(Kata Kerja) (Tembung Kriya)	Tuku Nyeterika Nggawa

15	KP	(Tanda Kutip) (Kutipan)	“ ” ‘ ’
16	KG	(Tanda Kurung) (Kurungan)	()[]
17	TWP	(Tanda di akhir Kalimat) (Tandha Wacan Pungkasan)	. ! ?
18	TWT	(Tanda di tengah Kalimat) (Tandha Wacan Tengah)	: ; - /

Preprocessing

Preprocessing merupakan proses awal yang akan mentransformasikan data masukan menjadi data dengan format yang sesuai dan siap untuk diproses [22]. *Preprocessing* digunakan untuk mengubah data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan pada penelitian. Pada penelitian ini preprocessing yang dilakukan terdiri dari tiga tahap. Tahap pertama yaitu menghapus tanda diakritik pada huruf vokal. Tanda diakritik adalah tanda baca tambahan pada huruf yang sedikit banyak mengubah nilai fonetis huruf tersebut, misalnya tanda [´] pada huruf é [23]. Tahap kedua *Case Folding* yaitu merubah semua huruf dalam kalimat menjadi huruf kecil dan melakukan *Tokenizing* yaitu melakukan pemenggalan pada kalimat menjadi kata dengan simbol dan spasi sebagai pemenggalan. Untuk lebih jelasnya tahapan *preprocessing* ditunjukkan pada Gambar 2.



Gambar 2. *Preprocessing*

Hidden Markov Model (HMM)

Hidden Markov Model (HMM) adalah sebuah model statistik dari sebuah sistem yang melakukan perhitungan probabilitas dari suatu kejadian yang tidak dapat diamati berdasarkan kejadian yang dapat diamati [24]. Perhitungan probabilitas dilakukan dengan melihat kejadian-kejadian lain yang dapat diamati secara langsung [14]. Dalam

Hidden Markov Model memiliki 2 macam bagian yaitu *observed state* dan *hidden state*. *Observed state* merupakan bagian yang dapat diamati secara langsung dan *hidden state* merupakan bagian yang tidak dapat diamati [25]. *Hidden states* mempresentasikan label kelas kata (*Tag*) dan *observation states* mempresentasikan kata (*Words*). Probabilitas transisi bergantung pada pasangan *Tag* sedangkan probabilitas emisi bergantung pada *Tag* saat ini [26]. Pada *markov* terdapat dua asumsi, asumsi pertama suatu kata bergantung pada dirinya sendiri tanpa memperhitungkan kelas kata sekitarnya. Asumsi kedua yaitu suatu kemunculan kata hanya bergantung pada kelas kata sebelumnya dalam penelitian ini menggunakan *bigram assumption* atau *first order markov chain*.

Persamaan *Hidden Markov Model* dalam pelabelan kelas kata (*Part of Speech Tagging*), seperti dalam persamaan (1)

$$Tag_n = Max\ (P(word_i \mid tag_i)\ x\ P(tag_i \mid tag_{i-1})) \quad (1)$$

Tag_n	: Kelas kata yang dicari
tag_i	: Kelas kata dari *word* i yang ada di *corpus*.
word_i	: Kata yang dicari kelas katanya
tag_(i-1)	: Kelas kata sebelum kelas kata dari *word* i yang ada di *corpus* sebanyak 1
P	: Probabilitas
P(word_i | tag_i)	: Probabilitas emisi
P(tag_i| tag_(i-1))	: Probabilitas transisi

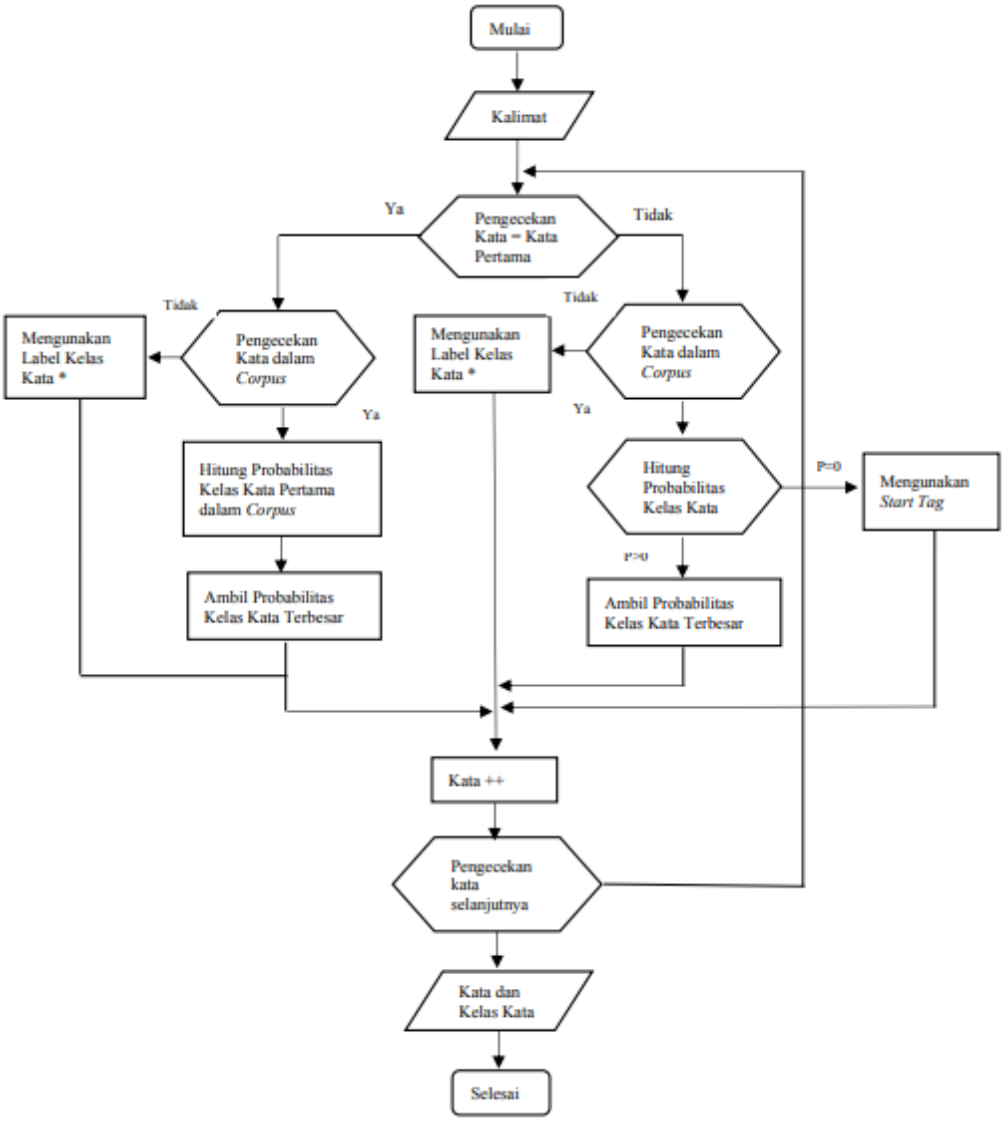
Most Probable POS-Tag

Unknown word atau kata yang tidak dikenal dalam model statistik, *Unknown word* tidak terdapat dalam *training corpus* sehingga mempengaruhi proses dan hasil *Part of Speech Tagging*. *Unknown words* akan menghasilkan probabilitas emisi bernilai nol. Terdapat berbagai cara dalam menangani *unknown words* salah satunya yaitu *Most Probable POS-Tag*. *Most probable POS-Tag* yaitu dengan melihat *training corpus* untuk mendapatkan label kelas kata yang paling banyak atau paling sering digunakan. Label kelas kata tersebut diasumsikan bahwa kata-kata yang tidak dikenal selalu memiliki label kelas kata tersebut [27].

Flowchart

Gambar 3 merupakan *Flowchart* skenario pertama yaitu *Hidden Markov Model* tanpa perlakukan untuk *unknown words* yang digunakan dalam penelitian ini. Berikut merupakan tahapan penerapan *Hidden Markov Model* secara detail:

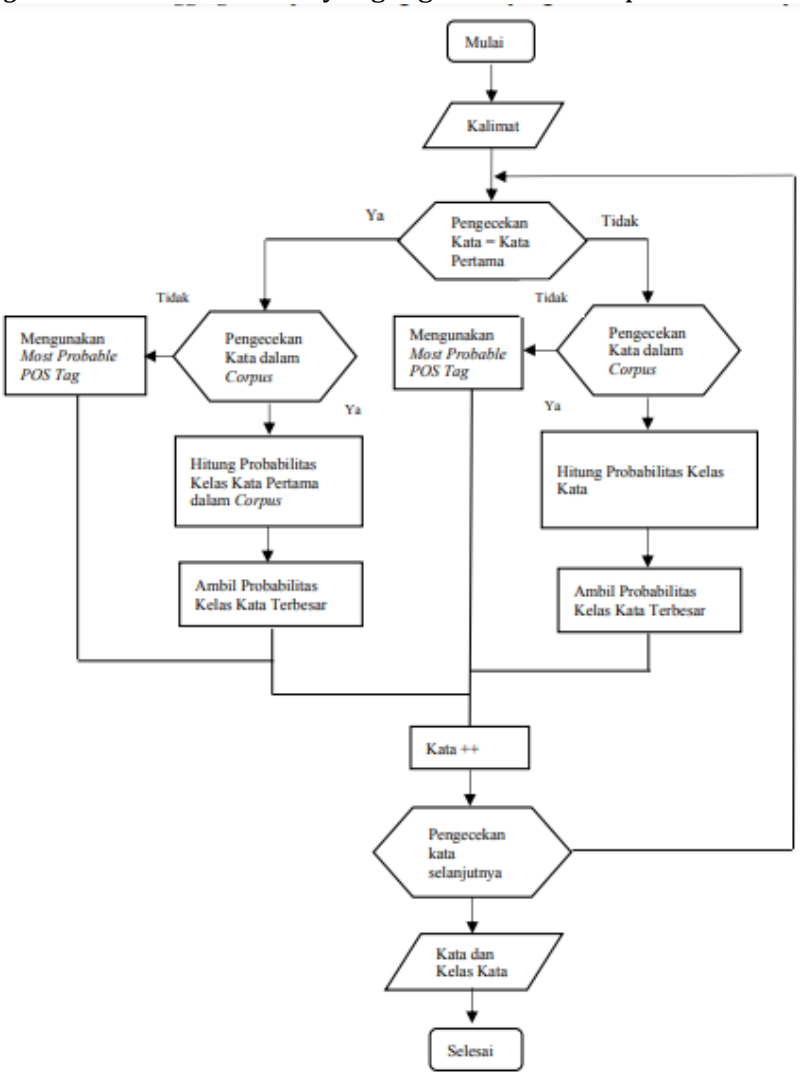
1. Proses *POS Tagging Hidden Markov Model* dimulai dengan pengecekan kata pertama pada suatu kalimat.
2. Jika kata termasuk kata pertama pada kalimat, maka selanjutnya dilakukan pengecekan setiap kata dalam *corpus*.
3. Dilakukan pengecekan kata pertama pada *corpus*. Jika kata pertama tidak terdapat dalam *corpus* diberikan label kelas kata khusus.
4. Jika kata pertama terdapat pada *corpus*, menghitung jumlah kemunculan kata pada label kelas kata yang dicari dengan total label kelas kata tersebut (Probabilitas Emisi).
5. Selanjutnya menghitung jumlah hubungan label kelas kata yang dicari dengan *start tag* (Probabilitas Transisi) untuk kata pertama.



Gambar 3. *Flowchart POS Tagging Hidden Markov Model*

6. Selanjutnya menghitung peluang terbesar dari hasil perhitungan perkalian probabilitas emisi dan transisinya untuk kata pertama.
7. Dilakukan pengecekan kata kedua pada *corpus*. Jika kata kedua tidak terdapat dalam *corpus* diberikan label kelas kata khusus.
8. Jika kata kedua terdapat pada *corpus*. Maka dilakukan perhitungan probabilitas kelas kata dari perkalian probabilitas emisi dan transisinya. Perhitungan probabilitas transisi berdasarkan hubungan label kelas kata yang dicari dengan label hasil *POS Tagging* kata pertama. Hasil yang diambil adalah probabilitas terbesar dari perhitungan probabilitas emisi dan transisi.
9. Jika kata kedua menghasilkan probabilitas bernilai nol maka akan memberikan hasil label *start tag*.
10. Kata ketiga sampai kata terakhir dalam kalimat dihitung sama seperti proses pada pencarian probabilitas kata kedua.
11. Hasil dari *POS Tagging* berupa kata dan kelas kata.

Gambar 4 adalah *Flowchart* skenario kedua *Hidden Markov Model* dan *Most Probable POS-Tag* untuk *unknown words* yang digunakan dalam penelitian ini.



Gambar 4. *Flowchart POS Tagging Hidden Markov Model* dan *Most Probable POS-Tag*

1. Proses *POS Tagging Hidden Markov Model* dan *Most Probable POS-Tag* dimulai dengan pengecekan kata pertama pada suatu kalimat.
2. Jika kata termasuk kata pertama pada kalimat, maka selanjutnya dilakukan pengecekan setiap kata dalam *corpus*.
3. Dilakukan pengecekan kata pertama pada *corpus*. Jika kata pertama tidak terdapat dalam *corpus* maka menggunakan label dari *Most Probable POS-Tag*.
4. Jika kata terdapat pada *corpus*, menghitung Probabilitas Emisi dari kata pertama.
5. Selanjutnya menghitung Probabilitas Transisi untuk kata pertama.
6. Selanjutnya menghitung peluang terbesar dari hasil perhitungan perkalian probabilitas emisi dan transisinya untuk kata pertama.
7. Dilakukan pengecekan kata kedua pada *corpus*. Jika kata kedua tidak terdapat dalam *corpus* maka menggunakan label dari *Most Probable POS-Tag.*

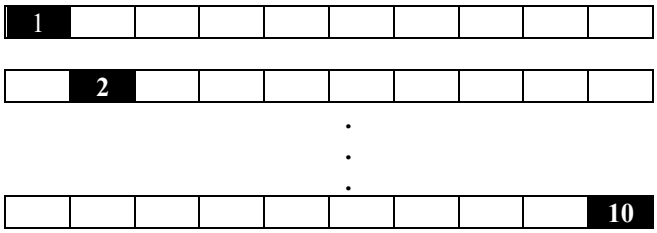
8. Jika kata kedua terdapat pada *corpus*. Maka dilakukan perhitungan probabilitas kelas kata dari perkalian probabilitas emisi dan transisinya. Perhitungan probabilitas transisi berdasarkan hubungan label kelas kata yang dicari dengan label hasil *POS Tagging* kata pertama. Hasil yang diambil adalah probabilitas terbesar dari perhitungan probabilitas emisi dan transisi.
9. Kata ketiga sampai kata terakhir dalam kalimat dihitung sama seperti proses pada pencarian probabilitas label kelas kata kedua.
10. Hasil dari *POS Tagging* berupa kata dan kelas kata

Evaluation

Pada tahap evaluasi dilakukan perhitungan akurasi untuk mengetahui kinerja dari metode *Hidden Markov Model* untuk *POS Tagging* Bahasa Jawa. Pada proses perhitungan akurasi yaitu menghitung akurasi pada setiap dokumen data uji. Perhitungan nilai akurasi pada dokumen menggunakan rumus seperti dalam persamaan (2)

$$Accuracy = (Jumlah\ Kata\ Benar\ x\ Jumlah\ Kata\ Keseluruhan)\ x\ 100\% \quad (2)$$

Hasil perhitungan akurasi dokumen digunakan untuk menghitung akurasi penelitian dengan menggunakan 10 *Fold Cross Validation*. Konsep 10 *Fold Cross Validation* ditunjukan pada Gambar 5.



Gambar 5. 10-*Fold Cross Validation*

HASIL DAN PEMBAHASAN

Dari hasil pengujian menggunakan *dataset* yang berisi cerita pendek Bahasa Jawa telah mampu melakukan proses *POS Tagging* pada kalimat bahasa jawa. Dari 10 dokumen yang diujikan menggunakan konsep 10 *cross validation*, skenario pertama yaitu *POS Tagging* dengan menggunakan *Hidden Markov Model* tanpa perlakuan untuk *unknown word* menghasilkan akurasi sebesar 45.5%. Skenario kedua yaitu *POS Tagging* HMM dengan menggunakan *Most Probable POS-Tag* menghasilkan akurasi sebesar 70.78%. Hasil pengujian terhadap *POS Tagging* dengan menggunakan *Hidden Markov Model* ditunjukkan pada Tabel 5.

Tabel 5. Hasil Pengujian

No	Uji	Akurasi	
		Hidden Markov Model	***Hidden Markov Model Most Probable POS-Tag***
1	Dokumen 1	41.58%	66.84%
2	Dokumen 2	45.11%	70.26%
3	Dokumen 3	38.47%	65.80%

4	Dokumen 4	46.93%	70.36%
5	Dokumen 5	41.77%	65.78%
6	Dokumen 6	59.81%	77.12%
7	Dokumen 7	52.55%	76.61%
8	Dokumen 8	39.82%	67.79%
9	Dokumen 9	45.33%	74.01%
10	Dokumen 10	43.59%	73.21%
Rata-rata (Mean)		***45.5%***	***70.78%***

Setelah dilakukan pengujian, hasil pelabelan kelas kata oleh algoritma *POS Tagging* dengan menggunakan *Hidden Markov Model* terdapat beberapa hal yang mempengaruhi akurasi. Kata yang tidak terdapat pada *corpus* diberikan label kelas kata *non tagset* Bahasa Jawa mempengaruhi hasil dari label untuk kata setelahnya. Probabilitas yang dihasilkan bernilai nol maka kata setelahnya akan memiliki label *start tag*. Label yang tidak memiliki transisi dengan label *start tag* juga akan memiliki probabilitas nol seperti label tanda titik dan label tanda koma. Penggunaan asumsi *Most Probable POS-Tag* tidak selalu menunjukan hasil yang benar. Penggunaan *Most Probable POS-Tag* akan memberikan label kelas kata yang sama untuk semua kata yang tidak terdapat pada *corpus*. *Most Probable POS-Tag* dapat menghilangkan probabilitas bernilai Nol dari *POS Tagging Hidden Markov Model*.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan bahwa algoritma *Hidden Markov Model* dapat digunakan untuk melakukan *POS Tagging* Bahasa Jawa dengan tingkat akurasi 45,5%. Hasil akurasi dipengaruhi *unknown words* sehingga diperlukan perlakuan untuk *unknown words* dengan menggunakan *Most Probable POS-Tag*. Hasil pengujian setelah diberikan perlakuan memperoleh akurasi sebesar 70,78%. Terjadi peningkatan akurasi tetapi penggunaan *Most Probable POS-Tag* tidak selalu menunjukan hasil label *POS Tagging* yang benar. Penggunaan *Most Probable POS-Tag* dapat menghilangkan probabilitas kelas kata yang bernilai nol.

Untuk meningkatkan hasil akurasi disarankan menambah perbendaharaan kata dan transisi antar label kelas kata. Selain itu juga untuk meminimalisir pengaruh *unknown words* pada *corpus*. Menggunakan algoritma lainnya untuk perlakuan *unknown word*. Algoritma yang bisa digunakan berbasis morfologi (*prefixes*, *suffixes*), *Laplacian Smoothing, Maximum likelihood, Overall POS distribution*, dan *Hapax Legomena*. Menggunakan atau menggabungkan algoritma lain dalam *POS Tagging*. Algoritma yang bisa digunakan *N-gram, Viterbi Algorithm, Maximum Entropy, Baum-Welch Algorithm, Neural Network, Rule Base* dan *Brill Tagger*. Kedepannya diharapkan adanya pengembangan yang dapat melakukan perbaikan kesalahan – kesalahan untuk menghasilkan pelabelan kelas kata yang baik dan optimal.

DAFTAR PUSTAKA

[1] Aji P. Wibawa, Andrew Nafalski, Neil Murray, and Wayan F. Mahmudy, "Parallel Text Processing: Alignment of Indonesian to Javanese Language," *Int. J. Inf. Control Comput. Sci.*, vol. 6.0, no. 2, Jan. 2014, doi: 10.5281/zenodo.1335958.

[2] W. E. S. Nurlina, Herawati, D. Sutono, and T. Suwondo, *Pembentukan Kata dan Pemilihan Kata dalam Bahasa Jawa*. Jakarta: Pusat Bahasa Departemen Pendidikan Nasional, 2004.
[3] G. Quinn, "Teaching Javanese respect usage to foreign learners," *Electron. J. Foreign Lang. Teach.*, vol. 8, pp. 362–370, Dec. 2011.
[4] A. K. Ogloblin, "Javanese," in *The Austronesian Languages of Asia and Madagascar*, Routledge Language Family Series, 2005, p. 590.
[5] Wedhawati, *Tata bahasa Jawa mutakhir*. Kanisius, 2006.
[6] A. Munandar, "Pemakaian Bahasa Jawa Dalam Situasi Kontak Bahasa Di Daerah Istimewa Yogyakarta," *HUMANIORA*, vol. 25, pp. 92–102, Feb. 2013.
[7] H. B. Mardikantoro, "Pergeseran Bahasa Jawa Dalam Ranah Keluarga Pada Masyarakat Multibahasa Diwilayah Kabupaten Brebes," *HUMANIORA*, vol. 19, pp. 43–51, Feb. 2007.
[8] F. H. Tondo, "Kepunahan Bahasa-Bahasa Daerah: Faktor Penyebab Dan Implikasi Etnolinguistis," *J. Masy. Dan Budaya*, vol. 11, no. 2, pp. 277–296, 2009.
[9] N. X. Bach, N. D. Linh, and T. M. Phuong, "An empirical study on POS tagging for Vietnamese social media text," *Comput. Speech Lang.*, vol. 50, pp. 1–15, Jul. 2018, doi: 10.1016/j.csl.2017.12.004.
[10] N. Sabloak, B. Agung Hardono, and D. Alamsyah, "Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi," *Unpublished*, Jul. 2016.
[11] A. Mulyanto, Y. A. Nurhuda, and N. Wiyanto, "Penyelesaian Kata Ambigu Pada Proses POS Tagging Menggunakan Algoritma Hidden Markov Model (HMM)," *Pros. Semin. Nas. Metode Kuantitatif*, vol. 0, no. 1, Nov. 2017.
[12] A. Azimizadeh, M. Arab, and S. R. Quchani, "Persian part of speech tagger based on Hidden Markov Model," 2008, pp. 121–128.
[13] L. M. S. Martínez, C. A. Cobos, and J. C. Corrales, "Memetic Algorithm Based on Global-Best Harmony Search and Hill Climbing for Part of Speech Tagging," in *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings*, A. Ghosh, R. Pal, and R. Prasath, Eds. Springer International Publishing, 2017.
[14] K. Widhiyanti and A. Harjoko, "POS Tagging Bahasa Indonesia Dengan HMM dan Rule Based," *J. Inform.*, vol. 8, no. 2, Mar. 2013, doi: 10.21460/inf.2012.82.125.
[15] H. Mohamed, N. Omar, and M. J. A. Aziz, "Statistical malay part-of-speech (POS) tagger using Hidden Markov approach," in *2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011*, 2011, pp. 231–236, doi: 10.1109/STAIR.2011.5995794.
[16] S. Dandapat, S. Sarkar, and A. Basu, "(PDF) A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali.," *Int. Conf. Comput. Intell.*, pp. 169–172.
[17] H. R. U. Pramudita Ema; Amborowati, Armadyah, "Pengaruh Part of Speech Tagging Berbasis Aturan dan Distribusi Probabilitas Maximum Entropy untuk Bahasa Jawa Krama," *J. Buana Inform.*, no. Vol 7, No 4 (2016): Jurnal Buana Informatika Volume 7 Nomor 4 Oktober 2016, 2016.
[18] J. W. M. Verhaar, *Pengantar Linguistik*. Yogyakarta: UGM Press, 1982.
[19] H. Alwi, *Tatabahasa Baku Bahasa Indonesia*. Balai Pustaka, 1993.
[20] Y. Sudaryat, "Pemarkah Pertarafan Dalam Bahasa Sunda," *Adab. J. Bhs. Dan Sastra*, vol. 12, no. 2, pp. 263–282, Dec. 2013, doi: 10.14421/ajbs.2013.12203.
[21] A. B. Setiyanto, *Parama Sastra Bahasa Jawa*. Yogyakarta: Panji Pustaka, 2007.
[22] D. Setyohadi, "Perbaikan Performansi Klasifikasi Dengan Preprocessing Iterative Partitioning Filter Algorithm," *telematika*, Apr. 2017.

[23] L. Setyowati, Bertalya, and T. W. R. Ningsih, "Aplikasi Transkripsi Fonetik Bahasa Indonesia Berdasarkan IPA (The International Phonetic Association) Untuk BIPA," *Pros. Semin. Ilm. Nas. Komput. Dan Sist. Intelijen KOMMIT 2014*, vol. 8, Oktober 2014.
[24] Jurafsky, D and Martin, J, H, *Speech and Language Processing"An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition"*. New Jersey: Prentice Hall, 2000.
[25] Y. Wibisono, "Penggunaan Hidden Markov Model untuk Kompresi Kalimat," Jan. 2008.
[26] A. Farizki Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," Jan. 2010.
[27] M. Haulrich, "Different Approaches to Unknown Words in a Hidden Markov Model Part-of Speech Tagger," *Unpublished*, May 2009.