Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072449
Vol. 7, No. 6, Desember 2020, hlm. 1121-1128 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

POS TAGGING BAHASA MADURA DENGAN MENGGUNAKAN ALGORITMA BRILL TAGGER

Nindian Puspa Dewi*[1], Ubaidi[2]

[1,2]Informatika, Universitas Madura
Email: [1]nindianpd@unira.ac.id, [2]ubed@unira.ac.id
*Penulis Korespondensi

(Naskah masuk: 09 September 2019, diterima untuk diterbitkan: 25 November 2020)

Abstrak

Bahasa Madura adalah bahasa daerah yang selain digunakan di Pulau Madura juga digunakan di daerah lainnya seperti di kota Jember, Pasuruan, dan Probolinggo. Sebagai bahasa daerah, Bahasa Madura mulai banyak ditinggalkan khususnya di kalangan anak muda. Beberapa penyebabnya adalah adanya rasa gengsi dan tingkat kesulitan untuk mempelajari Bahasa Madura yang memiliki ragam dialek dan tingkat bahasa. Berkurangnya penggunaan Bahasa Madura dapat mengakibatkan punahnya Bahasa Madura sebagai salah satu bahasa daerah yang ada di Indonesia. Oleh karena itu, perlu adanya usaha untuk mempertahankan dan memelihara Bahasa Madura. Salah satunya adalah dengan melakukan penelitian tentang Bahasa Madura dalam bidang *Natural Language Processing* sehingga kedepannya pembelajaran tentang Bahasa Madura dapat dilakukan melalui media digital. *Part Of Speech* (POS) *Tagging* adalah dasar penelitian *text processing*, sehingga perlu untuk dibuat aplikasi POS Tagging Bahasa Madura untuk digunakan pada penelitian *Natural Languange Processing* lainnya. Dalam penelitian ini, POS Tagging dibuat dengan menggunakan Algoritma Brill Tagger dengan menggunakan *corpus* yang berisi 10.535 kata Bahasa Madura. POS Tagging dengan Brill Tagger dapat memberikan kelas kata yang sesuai pada kata dengan menggunakan aturan leksikal dan kontekstual. Brill Tagger merupakan algoritma dengan tingkat akurasi yang paling baik saat diterapkan dalam Bahasa Inggris, Bahasa Indonesia dan beberapa bahasa lainnya. Dari serangkaian percobaan dengan beberapa perubahan nilai *threshold* tanpa memperhatikan OOV *(Out Of Vocabulary),* menunjukkan rata-rata akurasi mencapai lebih dari 80% dengan akurasi tertinggi mencapai 86.67% dan untuk pengujian dengan memperhatikan OOV mencapai rata-rata akurasi 67.74%. Jadi dapat disimpulkan bahwa Brill Tagger dapat digunakan untuk Bahasa Madura dengan tingkat akurasi yang baik.

Kata kunci: *part of speech, pos tagging, bahasa madura, brill tagger, tagset.*

POS TAGGING BAHASA MADURA WITH BRIL TAGGER ALGORITHM

Abstract

Bahasa Madura is regional language which is not only used on Madura Island but is also used in other areas such as in several regions in Jember, Pasuruan, and Probolinggo. Today, Bahasa Madura began to be abandoned, especially among young people. One reason is sense of pride and also quite difficult to learn Bahasa Madura because it has a variety of dialects and language levels. The reduced use of Bahasa Madura can lead to the extinction of Bahasa Madura as one of the regional languages in Indonesia. Therefore, there needs to be an effort to maintain Madurese Language. One of them is by conducting research on Madurese Language in the field of Natural Language Processing so that in the future learning about Madurese can be done through digital media. Part of Speech (POS) Tagging is the basis of text processing research, so the Madura Language POS Tagging application needs to be made for use in other Natural Language Processing research. This study uses Brill Tagger by using a corpus containing 10,535 words. POS Tagging with Brill Tagger Algorithm can provide the appropriate word class to word using lexical and contextual rule. The reason for using Brill Tagger is because it is the algorithm that has the best accuracy when implemented in English, Indonesian and several other languages. The experimental results with Brill Tagger show that the average accuracy without OOV (Out Of Vocabulary) obtained is 86.6% with the highest accuracy of 86.94% and the average accuracy for OOV words reached 67.22%. So it can be concluded that the Brill Tagger Algorithm can also be used for Bahasa Madura with a good degree of accuracy.

Keywords: *part of speech, pos tagging, bahasa madura, brill tagger, tagset.*

1. PENDAHULUAN

Bahasa Madura adalah bahasa daerah yang digunakan di Pulau Madura dan beberapa daerah lainnya seperti Jember, Pasuruan dan Probolinggo. Sebagai bahasa daerah, Bahasa Madura perlu dibina dan dikembangkan, terutama dalam hal peranannya sebagai sarana pengembangan kebudayaan daerah untuk mendukung kebudayaan nasional (halim,1976). Menurut Purwo (2000), Bahasa Madura menduduki peringkat keempat penutur terbanyak yang digunakan setelah Bahasa Jawa. Namun saat ini Bahasa Madura semakin banyak ditinggalkan oleh Masyarakat Madura khususnya di kalangan anak muda (Mulyadi, 2014). Ada banyak faktor yang menyebabkan semakin berkurangnya penggunaan dan pemahaman terhadap Bahasa Madura, antara lain karena (1) penggunaan Bahasa Indonesia sebagai bahasa utama pengantar pendidikan, (2) rasa malu saat menggunakan bahasa daerah, (3) sedikitnya penggunaan Bahasa Madura dalam media massa baik dalam bentuk tulisan maupun siaran berbahasa Madura, dan (4) sulitnya mempelajari Bahasa Madura yang memang memiliki ragam tutur/dialek dan cara penulisan yang unik (Sofyan, 2017).

Berkurangnya penggunaan Bahasa Madura dapat mengakibatkan punahnya Bahasa Madura sebagai salah satu bahasa daerah yang ada di Indonesia. Oleh karena itu, perlu adanya usaha untuk mempertahankan dan memelihara Bahasa Madura. Menurut Harimurti (2001), pemeliharaan Bahasa adalah usaha agar suatu bahasa tetap dipakai dan dihargai, terutama sebagai identitas kelompok dalam masyarakat bahasa yang bersangkutan melalui pengajaran, kesusasteraan, media massa dan lain-lain. Hal inilah yang melatarbelakangi penelitian tentang Bahasa Madura dengan menerapkan kemajuan teknologi informasi sehingga dapat meningkatkan eksistensi Bahasa Madura di kalangan masyarakat khususya masyarakat Madura.

Penelitian dalam bidang bahasa alami dan komputer biasa dikenal dengan sebutan natural language processing (NLP). Salah satu penelitian dalam bidang ini adalah part of speech tagging yang merupakan dasar untuk penelitian natural languange processing lainnya (Setyaningsih, 2017), seperti dalam word sense disambiguation, stemming (pencarian kata dasar), text summarization (peringkasan teks) dan question and answering. Part of speech (POS) biasa dikenal sebagai jenis kata dalam sebuah kalimat seperti kata kerja (verb), kata sifat (adjective), kata benda (noun) dan sebagainya (Manning & Schutze,1999.). POS tagging adalah proses memberi label pada setiap kata dalam kalimat dengan tag yang sesuai untuk kata tersebut (Christanti, Pragantha & Purnamasari, 2012)

Penelitian mengenai part of speech tagging di Indonesia sudah banyak dilakukan dengan menggunakan berbagai metode antara lain POS Tagging Bahasa Indonesia dengan HMM dan Rule Based (Kathryn & Agus, 2012), Probabilistic Part Of Speech Tagging for Bahasa Indonesia (Femphy, Mirna & Ruli, 2009) dengan menggunakan 37 tagset, Implementasi Brill Tagger untuk memberikan POS Tagging pada Dokumen Bahasa Indonesia (Christanti, Pragantha & Purnamasari, 2012) dan On Part of Speech Tagger for Indonesian Language (Yuwana, Yuliani & Pardede, 2017). Dari beberapa penelitian yang telah dilakukan, nilai akurasi tertinggi adalah dengan menggunakan Brill Tagger (Fahim, Naushad & Mumit, 2007). Brill Tagger diperkenalkan pertama kali oleh Eric Brill pada tahun 1992 (Brill, 1992). Proses Tagger merupakan transformation atau rules hasil belajar dari mendeteksi nilai error (Sriyati, 2016). Brill Tagger sendiri sudah diterapkan pada banyak bahasa, seperti : bahasa Inggris, Kadazan, dan Indonesia.

Part of speech merupakan dasar dalam pengembangan *text processing*. Karena itulah penulis melakukan penelitian part-of speech tagging pada Bahasa Madura dengan menggunakan Brill Tagger, sehingga dapat digunakan untuk pengembangan pengolahan teks dalam Bahasa Madura. Berdasarkan uraian tersebut maka ada beberapa tujuan yang ingin dicapai dalam penelitian ini yaitu :

1. Menetapkan daftar tagset standar untuk Bahasa Madura yang dapat digunakan sebagai dasar penelitian *text processing*.
2. Mengimplementasikan Brill Tagger untuk POS Tagging Bahasa Madura sehingga mempermudah proses pemberian label atau tag yang tepat pada kata.

2. METODE PENELITIAN

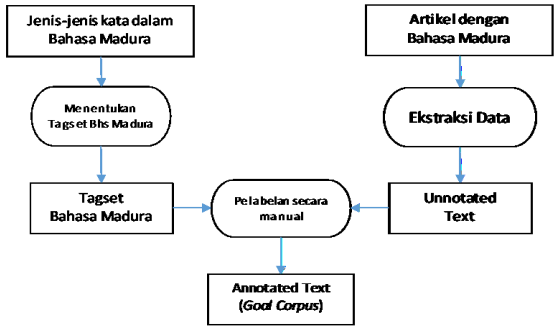
Penelitian ini merupakan lanjutan dari penelitian (Dewi & Ubaidi, 2018) yang lebih berfokus pada salah satu bagian dari Algoritma Brill Tagger yaitu pada tahap leksikal dengan akurasi tertinggi mencapai 87,43%. Selanjutnya pada penelitian ini dilakukan penambahan dan perbaikan corpus. Jika pada (Dewi & Ubaidi, 2018) tidak terlalu memperhatikan ketepatan urutan kata dalam kalimat maka pada penelitian ini susunan kalimat pada corpus diperbaiki sesuai dengan kaidah penulisan yang benar karena POS tagging yang akan dibuat tidak hanya pada tahap leksikal tapi juga pada tahap kontekstual. Penambahan jumlah *corpus* dapat meningkatkan jumlah kamus kata atau *lexicon* pada tahap leksikal, sedangkan ketepatan urutan kata penting karena sangat mempengaruhi pada *contextual rule* yang akan dihasilkan pada tahap kontekstual sehingga dapat meningkatkan nilai akurasi (Ayana, 2015).

Untuk membuat POS Tagging Bahasa Madura, diperlukan beberapa langkah yang harus dilakukan. Langkah pertama yaitu menambah daan memperbaiki corpus Bahasa Madura yang telah dibuat pada penelitian sebelumnya. Corpus kemudian digunakan

pada proses *learner*. Proses *learner* pada penelitian ini meliputi *lexical learner* dan *contextual learner*. Hasil dari proses *learner* ini yang akan digunakan untuk proses tagging.

2.1. Pengumpulan Data

Proses POS tagging dalam penelitian ini, dimulai dari penyusunan data set. Gambar 1 merupakan blok diagram penyusunan tagset bahasa madura dan proses pengambilan data *training* yang akan dijadikan *goal corpus*.

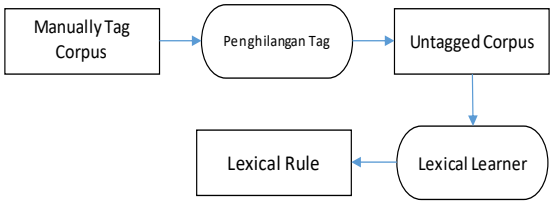


Gambar 1. Blog Diagram Penyusunan Data Set

Proses penyusunan data set diawali dengan menentukan tagset (kelas kata) standar untuk Bahasa Madura. Kemudian mengambil artikel berbahasa madura yang setelah dilakukan ekstraksi data disebut sebagai *unannotated text*. Berdasarkan tagset yang telah ditentukan, *unannotated text* ini kemudian diberi tag secara manual sehingga menjadi *annotated text (goal corpus)*. Pada penelitian ini untuk bagian penyusunan data corpus hanya dilakukan dengan perbaikan dan sedikit penambahan data corpus yang telah dibuat sebelumnya (Dewi & Ubaidi, 2018).

2.2. Lexical Learner

Setelah corpus terbentuk, proses dilanjutkan dengan *lexical learner*. *Lexical learner* merupakan proses pembelajaran untuk menghasilkan *lexical rule.* Dalam *lexical learner,* rule yang dihasilkan digunakan untuk melabeli kata dengan memperhatikan perubahan bentuk kata karena adanya imbuhan baik itu awalan, akhiran maupun awalan dan akhiran. Selain itu akan dilakukan juga pengecekan kata yang bersebelahan dan berapa frekuensi pasangan kata tersebut muncul.



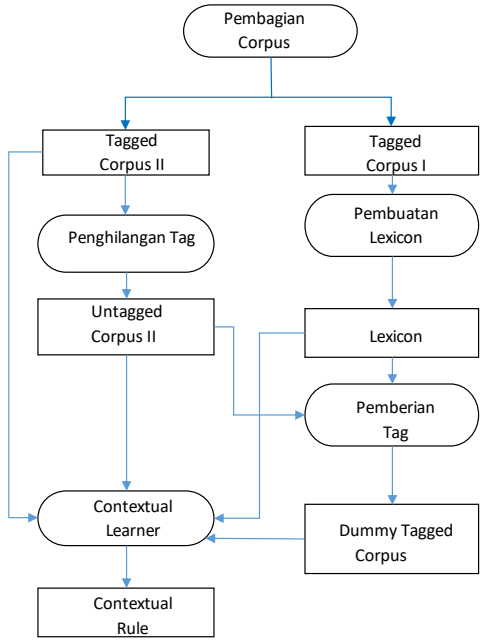
Gambar 2. Blog Diagram Lexical Learner

Gambar 2. menjelaskan tentang proses training pada *lexical learner* (Setyaningsih, 2017). Pada proses training dibutuhkan *manually tag corpus* yang merupakan corpus yang diberi tag secara manual. Manually tag corpus selanjutnya dihilangkan tagnya dan disebut *untagged corpus*. *Untagged corpus* kemudian dibandingakan dengan *manual tag corpus* sesuai dengan template *lexical rule* untuk menghasilkan *lexical rule*.

2.3. Contextual Learner

Contextual learner merupakan proses untuk menghasilkan *contextual rule*. *Contextual rule* adalah *rule* yang memperhatikan keberadaan tag disekitar kata yang sedang dicek atau dicari labelnya.

Pada dasarnya *contextual learner* digunakan untuk membandingkan *goal corpus (tagged corpus II)* dengan tag hasil *initial tagging* berdasarkan leksikon (Chaer, 2007) dan *lexical rule* (*dummy corpus*). Berikut blog diagram *contextual leaner* (Setyaningsih, 2017) dapat dilihat pada Gambar 3.



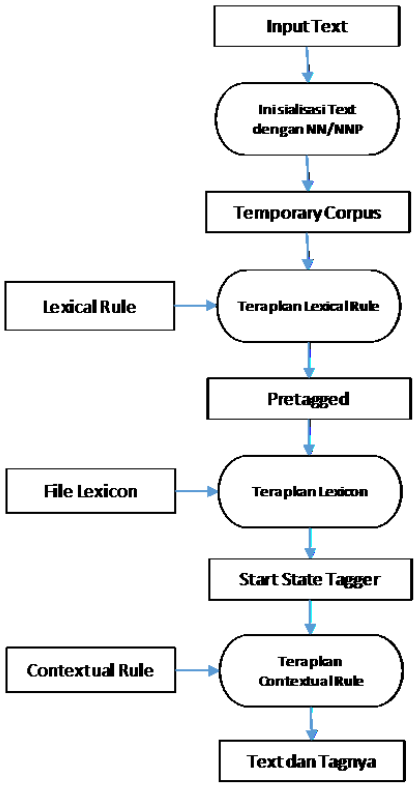
Gambar 3. Blog Diagram Contextual Learner

2.4. Pelabelan Kata (Tagging)

Pada penelitian ini, tagging dilakukan selain memperhatikan aturan leksikal juga aturan kontekstual (*contextual rule*). Proses pelabelan dimulai dengan menginputkan teks atau kalimat yang nantinya akan diberi label. Input teks awalnya diberi label awal berupa NN *(Common Noun)* dan NNP *(Proper Common Noun)* melalui proses inisialisasi teks. Hasil proses inisialisasi ini kemudian disebut sebagai *temporary corpus*.

Selanjutnya melalui proses *pretagged*, masing-masing rule dalam *lexical rule* dikenakan pada setiap kata dalam *temporary corpus*. Setelah itu setiap kata

akan dicari dalam leksikon. Leksikon berisi daftar kata dan tagnya yang merupakan sebagian dari data corpus yang digunakan. Jika kata ditemukan dalam leksikon, kelas kata akan diubah sesuai kelas kata dalam leksikon. Kata yang tidak ditemukan dalam leksikon kemudian akan dicek dengan *contextual rule*. Proses pelabelan teks dapat dilihat pada Gambar 4.



Gambar 4. Blog Diagram Pelabelan Kata

3. HASIL DAN PEMBAHASAN

3.1. Penelitian Terdahulu

Pada penelitian sebelumnya (Dewi & Ubaidi, 2018), corpus yang digunakan berjumlah 10.443 kata yang merupakan kumpulan artikel dan cerita berbahasa Madura. Proses training pada penelitian ini hanya dilakukan sampai pada tahap leksikal untuk menghasilkan *lexical rule* yang kemudian digunakan untuk proses *tagging*. Adapun *lexical rule* yang dihasilkan yaitu untuk *threshold* 10 menghasilkan 48 *rule*, *threshold* 20 hingga 40 menghasilkan 32 *rule*, sedangkan untuk *threshold* 50 menghasilkan 13 rule. Data uji yang digunakan pada proses tagging adalah data yang sama dengan data corpus. Rata-rata akurasi yang dicapai pada penelitian ini yaitu mencapai lebih dari 80% dengan akurasi tertinggi mencapai 87,43%.

3.2. Pengumpulan Data

Corpus Bahasa Madura yang digunakan berjumlah 10.535 kata yang merupakan kumpulan artikel dan cerita berbahasa Madura. Proses penyusunan corpus dilakukan dengan melibatkan sejumlah ahli dalam Bahasa Madura sehingga struktur dan susunan kata pada kalimat yang digunakan sesuai dengan Tata Bahasa Madura. Corpus dibuat dengan menggunakan kumpulan kalimat yang kemudian diberi kelas kata secara manual dengan menggunakan Tagset Bahasa Madura (Dewi & Ubaidi, 2018) yang ditunjukkan pada Tabel 1.

Tabel 1. Contoh Lexical Rule

No	Tagset	Simbol
1	Verba Transitif	VBT
2	Verba Intrasitif	VBI
3	Adjective	JJ
4	Adverb	RB
5	Common Noun	NN
6	Proper Common Noun	NNP
7	Genetive Common Noun	NNG
8	Personal Pronoun	PRP
9	Locative Pronoun	PRL
10	Primary Numeral	CDP
11	Collective Numeral	CDC
12	Distributive Numeral	CDD
13	Irregular Numeral	CDI
14	Kata bantu bilangan	CDB
15	WH Pronoun	WPRP
16	WH Adverb	WRB
17	Determiner	DT
18	Article	AR
19	Preposition	IN
20	Coordinate Conjunction	CC
21	Subordinate Conjunction	SC
22	Particle	RP
23	Interjection	UH
24	Positive Modal	MD
25	Negative Modal	NEG
26	Symbol	Sym
27	Sentence Terminator	ST
28	Comma	,
29	Ellipsis	...
30	Colon	:
31	Semi Colon	;
32	Open Paranthesis	OP
33	Close Paranthesis	CP
34	Quotation	QT
35	Dash	DASH
36	Slash	GM

Pada dasarnya struktur Bahasa dalam Bahasa Madura sama dengan Bahasa Indonesia, sehingga penentuan kelas katanya juga tidak jauh berbeda. Hanya saja ada beberapa kelas kata yang dipecah seperti jika dalam Bahasa Indonesia (Arawinda et al, 2014) kata kerja cukup diberi kelas kata verb (VB), maka dalam penelitian ini dibagi menjadi *verb transitif* (VBT) dan *verb intrasitif* (VBI).

3.3. Lexical Learner

Threshold pada *lexical learner* digunakan sebagai syarat berhentinya proses *learner*. Proses pembelajaran akan berhenti jika sudah diperoleh nilai terbaik (*bestscore*) yang diperoleh lebih kecil dari threshold. Tabel 2 merupakan contoh hasil uji coba dengan menggunakan variasi nilai threshold.

Hasil uji coba untuk threshold 10 menghasilkan 54 *rule*, threshold 20 hingga 40 menghasilkan 33 *rule*, sedangkan untuk threshold 50 menghasilkan 13 *rule*.

Hal ini menunjukkan bahwa nilai threshold berbanding terbalik dengan jumlah *rule* yang dihasilkan.

Tabel 2. Contoh Lexical Rule

Threshold	Jumlah	Contoh Lexical Rule
10	54	an redeletesuf NN
		an rehassuf NN
		pa haspref NN
		ng deletepref VBT
		è haspref VBT
		a haspref VBT
		na redeletesuf NN
		0 char CDP
		VBT a fhassuf NN
		ma addpref JJ
		JJ deletereant NN
		JJ ma fhaspref NN
		m addpref NN
		JJ a fchar NN
20	33	an redeletesuf NN
		an rehassuf NN
		pa haspref NN
		ng deletepref VBT
		è haspref VBT
		na redeletesuf NN
		0 char CDP
50	13	an redeletesuf NN
		an rehassuf NN
		pa haspref NN
		ng deletepref VBT
		è haspref VBT
		na redeletesuf NN

Setelah dilakukan penambahan dan perbaikan data corpus, terdapat perbedaan jumlah dan aturan leksikal yang dihasilkan. Pada penelitian sebelumnya untuk threshold 10 menghasilkan 48 *rule*, threshold 20 hingga 40 menghasilkan 32 *rule*, sedangkan untuk threshold 50 juga menghasilkan 13 *rule*.

3.4. Contextual Learner

Threshold pada contextual learner dengan Brill Tagger berfungsi syarat berhentinya proses learning. Berikut hasil uji coba dengan menggunakan variasi threshold, dimana masing-masing percobaan pada nilai threshold.

Tabel 3. Contoh Contextual Rule

T	Jumlah Rule	Contoh Contextual Rule
2	48	NN PRP CURWD kita
		NN VBT PREV1OR2WD ta’
		NN VBT SURROUNDTAG SC NN
		NN IN CURWD È
		NN NNP NEXTWD Madhurá
		NN SC CURWD jhá’
		NN VBT PREVTAG MD
		NN JJ PREVTAG RB
		NN IN CURWD dhá’ka
		NN IN CURWD akadhi
		NN CC CURWD nangèng
		NN CDI CURWD sabágián
		VBT NNG CURWD èssèna
		NN NNP SURROUNDTAG , ,
		NN SC CURWD saèngghána
		NN SC CURWD amarghá
		NN IN CURWD Kalabán
		NNP SC CURWD Sè
		NNP VBI CURWD Bádá
		NNP NEG CURWD Ta’
		NN VBT PREVTAG CP
		NN VBT PREVWD carana
		NN NNG SURROUNDTAG NN NEG
		NN NNG CURWD bhádhánna
		NN NNG CURWD asalla
		NN JJ PREVWD sè
		NN SC CURWD Saamponna
		NNP NN NEXTTAG DT
		NNP NN PREV1OR2TAG OP
		NNP NN CURWD Taon
		NN CDP PREVWD taon
		NN DT CURWD sadhájána
		NN DT CURWD Ka’dinto
		NN WP RBIGRAM pasèra sè
		NNP RB NEXTTAG JJ
		NN NNP NEXT1OR2WD tolèsanna
3	33	NN PRP CURWD kita
		NN VBT PREV1OR2WD ta’
		NN VBT SURROUNDTAG SC NN
		NN IN CURWD È
		NN NNP NEXTWD Madhurá
		NN SC CURWD jhá’
		NN VBT PREVTAG MD
		NN JJ PREVTAG RB
		NN IN CURWD dhá’ka
		NN IN CURWD akadhi
		NN CC CURWD nangèng
		NN CDI CURWD sabágián
		VBT NNG CURWD èssèna
		NN NNP SURROUNDTAG , ,
		NN VBT PREVWD kaangghuy
		NN VBT CURWD nombuwághi
		NN SC CURWD saèngghána
4	24	NN PRP CURWD kita
		NN VBT PREV1OR2WD ta’
		NN VBT SURROUNDTAG SC NN
		NN IN CURWD È
		NN NNP NEXTWD Madhurá
		NN SC CURWD jhá’
		NN VBT PREVTAG MD
		NN JJ PREVTAG RB
		NN IN CURWD dhá’ka

Setelah dilakukan beberapa kali perubahan threshold didapatkan jumlah *contextual rule* dengan cukup bervariasi tergantung threshold yang diberikan. Semakin kecil nilai threshold maka semakin banyak *contextual rule* yang didapatkan dan juga sebaliknya.

3.5. Pelabelan Kata (Tagging)

Dalam proses *tagging* dilakukan perhitungan untuk mengetahui nilai akurasi dari Brill Tagger untuk POS Tagging Bahasa Madura. Proses pelabelan kata dilakukan dengan menggunakan dua data yang berbeda yaitu data yang sama dengan data training dan data baru yaitu data yang tidak digunakan dalam proses training. Pada bagian berikut akan dibahas akurasi pada tahap leksikal dan kontekstual dengan beberapa perubahan nilai threshold.

Uji coba pertama menggunakan data yang sama dengan data yang digunakan dalam proses *learner* (data corpus). Tabel 4 berikut menunjukkan contoh hasil pelabelan dengan menggunakan data yang sama.

Adapun potongan inputan yang dijadikan kalimat uji coba yaitu "*Maskè la dháddhi sèttong kabunga'an jhá' sampè ageppa' dhádhá, tapè dháddhiá conto toladán sè saè mongghu dhá'ka sana' barajana"*.

Tabel 4. Contoh Hasil Tagging Menggunakan Data Corpus

Manually Tag Corpus	Hasil Tahap Lexical	Hasil Tahap Contextual
Maskè/SC	Maskè/SC	Maskè/SC
la/RB	la/RB	la/RB
dháddhi/VBT	dháddhi/VBT	dháddhi/VBT
sèttong/CDP	sèttong/CDP	sèttong/CDP
kabunga'an/NN	kabunga'an/NN	kabunga'an/NN
jhá'/NEG	jhá'/SC	jhá'/SC
sampè/IN	sampè/VBT	sampè'/VBI
ageppa'/VBT	ageppa'/VBT	ageppa'/VBT
dhádhá/NN	dhádhá/NN	dhádhá/NN
,/,	,/,	,/,
tapè/CC	tapè/CC	tapè/CC
dháddhiá/VBT	dháddhiá/VBT	dháddhiyá/VBT
conto/NN	conto/NN	conto/NN
toladán/JJ	toladán/NN	toladán/NN
sè/SC	sè/SC	sè/SC
saè/JJ	saè/JJ	saè/JJ
mongghu/IN	mongghu/IN	mongghu/IN
dhá'ka/IN	dhá'ka/NN	dhá'ka/IN
sana'/NN	sana'/NN	sana'/NN
barajana/NNG	barajana/NNG	barajana/NNG

Contoh proses tagging karena *lexical rule* yaitu rule "a haspref VBT" yang artinya jika tag awal NN maka ubah tag menjadi VBT jika kata yang akan dilabeli berawalan "a". Rule ini berhasil dikenai pada kata *ageppa'* (memukul) sehingga mendapatkan tag yang benar yaitu VBT. Rule ini diperoleh dari hasil learner dengan threshold 10 sampai 40. Untuk threshold 50, *rule* yang dihasilkan tidak diperoleh *rule* ini sehingga kata *ageppa'* masih memiliki tagset yang salah. *Contextual rule* dapat mengubah tag sebuah kata menjadi benar. Seperti kata dhá'ka (ke) yang awalnya mendapatkan tag yang salah yaitu NN (salah) berubah tagnya menjadi IN (benar) karena adanya rule hasil Brill Tagger "NN IN CURWD dhá'ka" yang memiliki arti "ubah tag menjadi IN jika katanya *dhá'ka* dan memiliki tag awal NN".

Sedangkan untuk nilai rata-rata akurasi yang dicapai dapat dilihat pada tabel 5. Threshold yang digunakan dalam tahap leksikal adalah yang memiliki nilai akurasi tertinggi (T=10).

Tabel 5. Rata-rata Akurasi Menggunakan Data Corpus

T	Jumlah Kata	Hasil Benar Lexical Rule	Hasil Benar Contextual Rule	Rata-rata Akurasi
2	541	504 Kata (93,16%)	506	93.53%
3	541		506	93.53%
4	541		506	93.53%

Akurasi pada tahap leksikal mengalami kenaikan dibandingkan hasil penelitian sebelumnya yang hanya mencapai akurasi tertinggi sebesar 87,43% menjadi 93,16%. Hal ini menunjukkan bahwa penambahan dan perbaikan corpus dapat meningkatkan nilai akurasi pada tahapan leksikal. Penambahan jumlah corpus dapat meningkatkan jumlah data dalam file leksikon, sedangkan perbaikan kata dalam corpus memperngaruhi ketepatan rule yang dihasilkan.

Selanjutnya untuk uji coba kedua menggunakan data baru yang tidak digunakan dalam proses learner (data corpus). Tabel 6 berikut menunjukkan contoh hasil tagging dengan menggunakan data yang baru.

Adapun potongan inputan yang dijadikan kalimat uji coba yaitu "*Maskè la dháddhi sèttong kabunga'an jhá' sampè ageppa' dhádhá, tapè dháddhiá conto toladán sè saè mongghu dhá'ka sana' barajana"*.

Tabel 6. Contoh Hasil Tagging Menggunakan Data Baru

Manually Tag Corpus	Hasil Tahap Lexical	Hasil Tahap Contextual
Mèlè/VBT	Mèlè/NNP	Mèlè/NNP
sapè/NN	sapè/NN	sapè/NNP
kaangghuy/IN	kaangghuy/IN	kaangghuy/IN
ghápanèka/DT	ghápanèka/DT	ghápanèka/DT
tanto/MD	tanto/MD	tanto/MD
bisaos/RB	bisaos/RB	bisaos/RB
dhá'/IN	dhá'/IN	dhá'/IN
sè/SC	sè/SC	sè/SC
ampon/RB	ampon/RB	ampon/RB
pèlak/MD	pèlak/NN	pèlak/JJ
mèlè/VBT	mèlè/NN	mèlè/VBT
sapè/NN	sapè/NN	sapè/NN
sè/SC	sè/SC	sè/SC
bhághus/JJ	bhághus/JJ	bhághus/JJ
./ST	./ST	./ST
Bágiyán/NN	Bágiyán/NN	Bágiyán/NN
sè/SC	sè/SC	sè/SC
mennang/JJ	mennang/NN	mennang/JJ
èkèrèm/VBT	èkèrèm/VBT	èkèrèm/VBT
ka/IN	ka/IN	ka/IN
Kerrap/NNP	Kerrap/NNP	Kerrap/NNP
Gubeng/NNP	Gubeng/NNP	Gubeng/NNP
./ST	./ST	./ST
È/IN	È/IN	È/IN
mosèm/NN	mosèm/NN	mosèm/NN
nèmor/NN	nèmor/VBT	nèmor/VBT
,/,	,/,	,/,
biyasana/JJ	biyasana/RB	biyasana/RB
teppa'/JJ	teppa'/NN	teppa'/JJ
ka/IN	ka/IN	ka/IN
Bulán/NNP	Bulán/NN	Bulán/NN
Oktober/NNP	Oktober/NNP	Oktober/NNP
./ST	./ST	./ST

Kata yang berhasil diberi tag yang benar pada tahap lexical yaitu kata *èkèrèm* (dikirim). Adanya rule "è haspref VBT" yang artinya, ubah tag menjadi VBT jika kata yang akan dilabeli berawalan "è". Untuk tahap kontekstual, kata *mennang* memperoleh rule yang benar karena rule "NN JJ PREVWD sè" yang artinya, jika tag awal adalah NN dan terletak setelah kata sè maka ubah tag menjadi JJ. Namun terkadang menyebabkan kesalahan tag karena tag lain yang salah seperti pada kata *sapè* yang karena NN "NN NNP PREVTAG NNP" yang artinya, jika tag awal adalah NN dan tag sebelumnya adalah NNP maka ubag tag menjadi NNP, menyebabkan kata *sapè* yang sebenarnya sudah memiliki tag yang benar (NN) diubah tagnya menjadi NNP.

Tabel 7. Rata-rata Akurasi Menggunakan Data Baru

T	Jumlah Kata	Hasil Benar Lexical Rule	Hasil Benar Contextual Rule	Rata-rata Akurasi
2	585	504 (85.81%)	507	86.67%
3	585		505	86.32%
4	585		505	86.32%

Dari tabel 7 di atas dapat dilihat bahwa pelabelan pada tahap kontekstual menghasilkan akurasi yang meningkat dari 85.81% menjadi 86.67% dengan menggunakan data yang baru. Semakin kecil nilai threshold, akurasi yang diperoleh cenderung semakin tinggi. Hal ini karena semakin banyaknya rule yang diperoleh dan diterapkan. Namun terkadang rule yang ada bisa mengakibatkan tag menjadi salah sehingga dapat menurunkan nilai akurasi.

Selanjutnya dilakukan pengujian dengan memperhatikan OOV *(Out of Vocabulari)*. Dalam uji coba diketahui dari 585 kata uji, jumlah kata yang digunakan dalam data latih (*Knownword*) adalah 342 kata dan jumlah kata yang belum pernah muncul atau ada dalam latih (*unKnownword*) adalah 243 kata. Setelah dilakukan uji coba didapat bahwa ada 507 kata yang berhasil diberi tag dengan benar yang terdiri dari 333 kata adalah *knownword* dan 174 kata adalah *unknownword*. Hasil uji coba dengan memperhatikan OOV dapat dilihat pada tabel 8 berikut.

Tabel 8. Rata-rata Akurasi dengan memperhatikan OOV

Overall Acc	Known Word Acc	Unknown Word Acc	Akurasi dengan OOV
$=\frac{507}{585}$	$=\frac{333}{342}$	$=\frac{174}{243}$	$\frac{86{,}67\%}{(97{,}36\%/71{,}60\%)}$
$= 86{,}67\%$	$= 97{,}36\%$	$= 71{,}60\%$	$= 67.74\%$

4. KESIMPULAN

Tagset (Kelas Kata) Bahasa Madura yang dapat dibentuk dari penelitian ini adalah 36 tagset. Tagset digunakan untuk membuat manual tag yang selanjutnya diolah untuk menghasilkan *lexical rule* melalui *lexical learner* dan *contextual rule* melalui *contetual learner*.

Nilai threshold pada *lexical learner* dan *contextual learner* mempengaruhi jumlah *rule* yang diperoleh dalam proses *learner*. Semakin rendah nilai threshold maka semakin banyak *rule* yang diperoleh dan begitu juga sebaliknya semakin tinggi nilai threshold maka semakin sedikit jumlah *rule* yang diperoleh. Hasil percobaan menunjukkan bahwa semakin banyak rule yang diperoleh maka nilai akurasi semakin tinggi.

Setelah dilakukan penambahan dan perbaikan data corpus yang digunakan pada penelitian sebelumnya, akurasi pada tahap leksikal meningkat yaitu dari 87,43% menjadi 93,16% dengan menggunakan data yang sama dengan data corpus. Untuk Penerapan Brill Tagger secara keseluruhan pada POS Tagging Bahasa Madura mencapai akurasi di atas 80% dengan akurasi tertinggi mencapai 86.67% jika tidak memperhatikan keberadaan OOV dan mencapai rata-rata akurasi 67.74% jika memperhatikan keberadaan OOV.

Kesimpulan ini menunjukan bahwa perbaikan corpus dengan memperhatikan ketepatan urutan kata dalam Bahasa Madura ternyata meningkatkan akurasi pelabelan kata, sehingga dapat digunakan dalam pengelompokan kelas kata untuk mendukung pembelajaran Bahasa Madura dalam rangka melestarikan Bahasa Madura.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih pada Direktorat Riset dan Pengabdian kepada Masyarakat khususnya Direktorat Jenderal Penguatan Riset dan Pengembangan karena telah memberikan kontribusi berupa dana penelitian sehingga kami dapat melaksanakan penelitian ini dengan baik. Terima kasih juga kami sampaikan kepada semua pihak yang secara tidak langsung membantu pelaksanaan penelitian ini.

DAFTAR PUSTAKA

AYANA, A.G. 2015. Improving Brill's Tagger Lexical and Transformation Rule for Afaan Oromo Language. PeerJ PrePrints, pp.1-11.

BRILL, E., 1992. A simple rule-based part of speech tagger. Proc. third Conf. Appl. Nat. Lang. Process, pp. 152.

CHAER, A. 2007. Linguistik Umum. Jakarta: Rineka Cipta.

CHRISTANTI, V., J. PRAGANTHA, E. PURNAMASARI. 2012. Implementasi Brill Tagger untuk memberikan POS-Tagging pada Dokumen Bahasa Indonesia. Jurnal Teknik dan Ilmu Komputer, 1(3), pp. 301–315.

DEWI, N.P., UBAIDI, 2018. Lexical Rule dan Pengaruh Penggunaan Lexicon Pada Pos Tagging Bahasa Madura. Jurnal Matrik, 18(1) pp.69-70.

DINAKARAMANI, A., RASHEL, F., LUTHFI,A., MANURUNG, R. 2014. Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus. International Conference on Asian Language Processing (IALP), 20-22 Oktober 2014,pp. 66-69.

HALIM, A. 1976. Politik Bahasa Nasional 1 dan 2. Jakarta: Aneka Ilmu.

HASAN, F.M., UZZAMAN, N., KHAN, M. 2007. Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla. Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp.121-126.

KRIDALAKSANA, H. 2001. Kamus Linguistik, Jakarta: Gramedia.

MANNING, C. D., HINRICH S. 1999. Foundation of Statistical Natural Language Processing. Cambridge: MIT Press Textbook on statistical and probabilistic methods in NLP.

MEGYESI, B. 1998. Brill's Rule-Based PoS Tagger for Hungarian. Master's Degree Thesis in Computational Linguistics. Department of Linguistics, Stockholm University, Sweden.

MULYADI. 2014. Pemakaian Bahasa Madura Di Kalangan Remaja. Okara, Vol.2, pp.45-68.

PISCELDO, F., ADRIANI, M., MANURUNG, R. 2009. Probabilistic Part Of Speech Tagging for Bahasa Indonesia. Third International MALINDO Workshop.

PURWO, B.K. 2000. Bangkitnya Kebhinekaan Dunia Linguistik dan Pendidikan. Jakarta: Mega Media Abadi.

SETYANINGSIH, E.R. 2017. Penetapan Tagset dan Modifikasi Brill Tagger untuk Part-of Speech Bahasa Indonesia. Dinamika Teknologi, 9(1), pp.37-42.

SOFYAN, A. 2017. Tata Bahasa Bahasa Madura. Sidoarjo: Bahasa Surabaya.

SRIYATI, N.P.M. 2016. Part-Of-Speech Tagging Untuk Dokumen Bahasa Bali Menggunakan Algoritma Brill Tagger: Fakultas Matematika dan Ilmu Pengetahuan Alam. Tugas Akhir. Universitas Udayana.

WIDHIYANTI, K., HARJOKO, A. 2012. POS Tagging Bahasa Indonesia Dengan HMM dan Rule Based. Jurnal Informatika, 8(2), pp.151-167.

YUWANA, R.S., YULIANI, A.R., PARDEDE,H.F. 2017. On Part of Speech Tagger for Indonesian Language. International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 1-2 Nopember 2017, pp. 369-372.