Mia Kamayani - Perkembangan Part-of-Speech Tagger Bahasa Indonesia

by Mia Kamayani By Lutfan

Submission date: 26-Oct-2023 09:05AM (UTC+0700)
Submission ID: 2207493294
File name: 20191_jlk.pdf (226.87K)
Word count: 2998
Character count: 18179

Jurnal Linguistik Komputasional
JLK
Vol 2, No 2
September
2019

Perkembangan *Part-of-Speech Tagger* Bahasa Indonesia

Mia Kamayani[#1]

[#]*Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Prof. Dr. HAMKA*
Jl. Tanah Merdeka No.6, Pasar Rebo, Jakarta
[1]mia.kamayani@uhamka.ac.id

***Abstrak*—Tujuan dari artikel ini adalah membuat kajian literatur terhadap metode pelabelan *part-of-speech* (POS tagger) untuk Bahasa Indonesia yang telah dilakukan selama 11 tahun terakhir (sejak tahun 2008). Artikel ini dapat menjadi roadmap POS tagger Bahasa Indonesia dan juga dasar pertimbangan untuk pengembangan selanjutnya agar menggunakan dataset dan tagset yang standar sebagai *benchmark* metode. Terdapat 15 publikasi yang dibahas, pembahasan meliputi dataset, tagset dan metode yang digunakan untuk POS tag Bahasa Indonesia. Dataset yang paling banyak digunakan dan paling mungkin menjadi corpus standar adalah IDN Tagged Corpus terdiri dari lebih dari 250.000 token. Tagset Bahasa Indonesia hingga saat ini belum terstandarisasi dengan jumlah label bervariasi dari 16 tag hingga 37 tag. Metode yang paling banyak dikembangkan dan berpotensi menjadi *state-of-the-art* adalah neural network, dengan varian metode biLSTM dan CRF dan sejauh ini memberikan skor F1 dan akurasi tertinggi (>96%).**

***Kata kunci*— part-of-speech, POS tagger, Bahasa Indonesia, dataset, tagset**

I. PENDAHULUAN

Part-of-speech (POS) tagger adalah upaya untuk memberikan label pada kata di dalam teks sesuai dengan kelas katanya dan juga kemungkinan fitur morfologinya. Pos tag memberikan informasi mengenai definisi dan konteks kata. Pos tag sangat penting sebagai proses awal dalam task NLP.

Tulisan ini meninjau publikasi primer selama 11 tahun dari tahun 2008 sampai tahun 2019 mengenai metode yang dikembangkan untuk *POS tagger* Bahasa Indonesia dan juga aplikasinya. Terdapat 15 (lima belas) publikasi yang memberi pengaruh signifikan terhadap perkembangan pos tagger Bahasa Indonesia. Penelitian terdahulu mengenai pos tagger ini dikelompokkan berdasarkan dataset, tagset dan pendekatan. *State-of-the-art* pos tagger Bahasa Indonesia didapatkan dari perbandingan hasil evaluasi pendekatan yang digunakan. Perbandingan antar penelitian dilakukan jika dataset dan tagset yang digunakan sama. Tulisan ini berupaya untuk mendapatkan roadmap pos tagger Bahasa Indonesia agar penelitian selanjutnya dapat menggunakan dataset dan tagset yang sama untuk mengevaluasi metode yang dipakai. Sehingga penelitian di bidang pos tagger Bahasa Indonesia dapat dilakukan secara berkelanjutan.

Sistematika penulisan akan disusun berdasarkan dataset, tagset dan pendekatan yang digunakan. Dataset adalah data yang digunakan dalam *training* dan/atau *testing*, tagset adalah himpunan label *part-of-speech* yang digunakan pada proses *tagging*.

Pendekatan pos tag yang digunakan dikelompokkan berdasarkan jenisnya, kemudian disusun secara kronologis. Tulisan ini membagi jenis pendekatan menjadi empat kelompok pendekatan yaitu rule based, probabilistik, HMM dan neural network. Walaupun HMM dan neural network termasuk dalam probabilistik, namun karena jumlah penelitian untuk kedua metode tersebut cukup signifikan maka dibuat kelompok sendiri.

Identifikasi metode *state-of-the-art* untuk pos tagger Indonesia cukup sulit karena hampir setiap penelitian menggunakan dataset dan tagset yang berbeda, sehingga sulit dibandingkan. Namun diharapkan tulisan ini dapat memberikan informasi kepada peneliti mana dataset dan tagset yang mungkin layak digunakan sebagai standar untuk pemrosesan pos tag Bahasa Indonesia.

II. DATASET

Dataset yang digunakan untuk *training* maupun pengujian pos tag Indonesia menggunakan pelabelan manual. Dataset yang digunakan antara lain diambil dari corpus yang memang disediakan secara publik, namun ada pula yang membangun corpus sendiri. Lihat ringkasan perbandingan corpus pada Tabel I.

A. Dataset yang Tersedia

Dataset yang tersedia secara publik yang digunakan sebagai data latih dan data uji antara lain Pan Localization (https://www.panl10n.net/) dan IDENTIC [1]. Nurwidyantoro, 2012 [2] menggunakan 1 juta kata dari PAN Localization untuk pengujian. Wicaksono, 2010 [3] memodifikasi 15.000 token dari PAN Localization Bahasa Indonesia.

Dinakamarani, 2014 [4] mengambil 10.000 kalimat (262.330 token) pertama dari dataset IDENTIC dan melakukan pelabelan manual untuk data latih. Rashel,

2014 [5] juga mengambil 10.000 kalimat dari dataset IDENTIC (250.000 token), dataset ini disebut IDN Tagged Corpus dan dapat diakses online (https://github.com/famrashel/idn-tagged-corpus).

TABEL I
PERBANDINGAN CORPUS

Corpus	Jml Token	Sumber	Ket
Pis09	14.165	Artikel koran	
Wic10	15.000	Pan Localization Bahasa Indonesia	
Lar11	994.545	Pan Localization, subtitle film dan artikel lain (IDENTIC)	Paralel corpus
Din14	262.330	IDENTIC	
Ras14	250.000	IDENTIC	
Fu18	355.000	Situs berita online	

B. *Dataset yang Dibuat Sendiri*

Pisceldo, 2009 [6] membangun corpus yang terdiri dari 14.165 kata yang diambil dari artikel koran. Larasati, 2011 [7] menggunakan 45.011 kalimat (994.545 kata) Bahasa Indonesia gabungan dari PAN Localization, subtitle film dan artikel lain yang disebut dengan corpus IDENTIC [1]. Sihui Fu, 2018 [8] membangun dataset yang cukup besar yaitu 355.000 token yang diambil dari beberapa situs berita online Indonesia.

III. TAGSET

Tagset yang dikembangkan untuk pos tagger Indonesia berdasarkan sumber yang dibahas disini ada enam. Tagset UD v2.2 adalah tagset yang dikembangkan di project Universal Dependencies (lihat http://www.universaldependencies.org). Tagset INACL merupakan konvensi pelabelan POS Indonesia yang dilakukan oleh tim peneliti INACL dan ini merupakan penelitian yang sedang berlangsung. Lihat ringkasan perbandingan di Tabel II.

Permasalahan pada tagset Bahasa Indonesia adalah belum adanya standar tata Bahasa [8], walaupun sudah merujuk pada KBBI, seringkali fungsi kata dipengaruhi konteks kalimat dan akan berbeda dari definisi di KBBI. Misal kata 'sudah' di KBBI diberi label *adverb* tapi Din14 melabelinya sebagai *modal verb*. Antar penelitian pun bisa jadi memberikan label berbeda untuk kata yang sama, misalnya kata 'sekarang' diberi label adverb oleh Pis09, tapi Din14 memberi label noun.

TABEL II
PERBANDINGAN TAGSET

Tagset	Jml	Sumber	Keterangan
Pis09	37 dan 25	Tata Bahasa Baku Indonesia (Alwi, 2003)	
Wic10	35	Modifikasi dari tagset [6], [9]	
Lar11	17	Adaptasi dari [10] dan tagset PENN Treebank	
Ras14	-	KBBI v3	Dibagi 2 kelompok: Closed class dan mwe class
Din14	23	Modifikasi dari 25 tagset [6]	
Fu18	29	Adaptasi dari beberapa penelitian	
UDv2.2	16	Stanford Dependencies	
INACL	23		Dibagi 2 kelompok kata: konten dan fungsi

IV. PENDEKATAN YANG DIGUNAKAN

Ringkasan perbandingan empat kelompok pendekatan pos tagger Indonesia dapat dilihat pada Tabel III.

A. *Rule-Based*

Penelitian awal untuk analisis morfologi kata bahasa Indonesia dilakukan oleh Pisceldo dkk pada tahun 2008 [10]. Penelitian ini bertujuan melakukan analisis morfologi untuk proses afiksasi yang kompleks, di dalamnya didefinisikan tagset morfologi sejumlah 17, namun tag untuk part-of-speech sendiri hanya ada 3 yaitu verb, noun dan adj. Penelitian ini menggunakan 2 level pendekatan morfologi yaitu morfotaksis dan morfonemik, kemudian dimodelkan sebagai jaringan *finite state transducer* dan diimplementasikan menggunakan xfst dan lexc. Penelitian ini merupakan langkah awal analisis morfologi untuk Bahasa Indonesia.

Penelitian tentang analisis morfologi dikembangkan oleh Larasati pada tahun 2011 [11] dengan penggunaan fitur morfologi yang lebih lengkap (klitika, alternasi numerik dan morfem partikel tambahan) dan tagset yang lebih detail yaitu sejumlah 17. Penelitian ini menghasilkan tool MorphInd dan banyak digunakan oleh penelitian lain termasuk oleh proyek Universal Dependencies untuk Bahasa Indonesia.

POS tagger untuk Bahasa Indonesia dengan pendekatan *rule-based* telah dilakukan di tahun 2008 oleh Sari dkk dengan mengadaptasi Brill Tagger [9]. Penelitian ini menggunakan tagset yang terbatas dan dilatih pada corpus dalam jumlah kecil yang telah dianotasi manual. Akurasi dari pos tagger ini adalah 88%.

Setelah lama tidak mengalami perkembangan, rule-based digunakan kembali pada tahun 2014 [5], dengan

Korespondensi : Mia Kamayani

penambahan fitur *multi word expression* (MWE) dan *name entity recognition* (NER). Penelitian ini menggunakan beberapa resource seperti kamus MWE, kamus *closed-class word*, dan MorphInd [11] untuk pos tag pada *open-class word* dan disambiguasi. Penambahan fitur menghasilkan akurasi 79%, lebih baik daripada tanpa fitur MWE dan NER (70%). Corpus yang digunakan adalah 10.000 kalimat, 250.000 token (diambil dari IDENTIC corpus) yang dilabeli manual dengan jumlah tagset 23. Penelitian ini menghasilkan pos tagger yang dapat diakses online (http://bahasa.cs.ui.ac.id/postag/tagger).

B. Probabilistik

Penelitian *POS tagger* bahasa Indonesia di tahun 2009 oleh Pisceldo dengan membandingkan 2 metode probabilistik *Conditional Random Fields* (CRF) dan *Maximum Entropy* (ME) [6], CRF dan ME merupakan pendekatan probabilistik yang memberikan performa baik untuk pos tag bahasa Inggris (akurasi di atas 96%). Tagset yang digunakan ada 2 jenis yaitu 37 dan 25, corpus yang digunakan ada 2 jenis yaitu (1) artikel koran sebanyak 14.165 kata dan (2) sebagian dari terjemahan Penn Treebank (dari proyek Pan Localization) sebanyak 26.348 kata. Penelitian ini menunjukkan bahwa ME memberikan performa lebih baik dari CRF. Akurasi tertinggi dari ME mencapai 97.57%, sedangkan CRF 91.15%

Pengembangan dari penelitian di atas adalah penggunaan MapReduce pada ME untuk paralelisasi proses [2]. Hasilnya MapReduce efektif mempercepat proses tagging dan baik digunakan untuk jumlah corpus yang besar dan tidak disarankan untuk jumlah corpus sedikit.

ME menunjukkan hasil terbaik dibandingkan dengan 5 metode popular lainnya menurut Yuwana pada tahun 2017 [12]. Metode lain yang dibandingkan yaitu unigram, HMM, trigram, Brill, Naïve Bayes. Keenam metode ini diujikan pada set corpus yang digunakan oleh [3], dengan hasil akurasi ME tertinggi 88.43% dan terendah Brill 74.59%. Temuan dari penelitian ini adalah unigram memberikan akurasi terbaik kedua setelah ME.

TABEL III
PERBANDINGAN PENDEKATAN

Kelompok	Penelitian	Sub Metode
Rule based	Pis08	Finite state
	Sari08	Brill Tagger
	Lar11	Finite state
	Ras14	MWE+NER
Probabilistik	Pis09	Conditional Random Fields (CRF) dan Maximum Entropy (ME)
	Nur12	Paralelisasi ME dengan MapReduce
	Yuw17	ME
HMM	Wic10	Affix tree+lexicon
	Wid12	HMM+rule based
	Mul17	HMM+MorphInd
NN	Abka16	1 hidden layer (word embedding)
	Man18	2 input layer (word embedding, fitur morfologis dan kapitalisasi), 1 merge layer, 2 hidden layer
	Yuw18	2 hidden layer
	Kur18	biLSTM+CRF
	Fu18	biLSTM+CRF

C. Hidden Markov Model (HMM)

HMM termasuk pendekatan probabilistik, metode ini merupakan metode yang cukup banyak digunakan untuk pos tag di bahasa lain. Wicaksono pada tahun 2010 mengembangkan POS tagger Bahasa Indonesia dengan HMM [3]. Highlight dari penelitian ini adalah penggunaan affix tree dan penambahan lexicon memberikan performa yang sangat baik yaitu akurasi 96.5% (dengan tingkat *out-of-vocabulary*/OOV 15%). Corpus yang digunakan adalah 15.000 token dari Pan Localization yang telah dimodifikasi. Tagset yang digunakan adalah modifikasi 35 tagset yang digunakan oleh [6]. Penelitian oleh Wicaksono menjadi dasar pengembangan penelitian pos tagger HMM Bahasa Indonesia dan digunakan pula untuk task NLP Bahasa Indonesia lainnya yaitu mesin translasi [13], identifikasi opini dari teks [14], perangkat NLP [15] dan pengembangan alat belajar bahasa [16].

Pengembangan pos tagger HMM selanjutnya yaitu mengombinasikan HMM dengan metode lain yaitu HMM + rule-based oleh Widhiyanti dkk. di tahun 2012 [17], HMM + analisis morfologi oleh Muljono dkk di tahun 2017 [18], dan Ramadhanti di tahun 2019 [19]. Hasil dari ketiga penelitian ini menyimpulkan bahwa kombinasi metode HMM dengan metode lain memberikan performa yang lebih baik daripada HMM saja.

Penggabungan HMM dan analisis morfologi memberikan akurasi 99.14% [19]. Analisis morfologi yang digunakan adalah menggunakan afiks terhadap pembentukan kelas kata, misal imbuhan ‘mem-‘ dan

'ber' adalah termasuk kelas kata verba, sehingga informasi ini akan mengurangi tingkat OOV. Training data menggunakan IDN Tagged Corpus [5] dan testing menggunakan artikel berita online dengan 6676 token dan 30% OOV. HMM tanpa analisis morfologi memberikan akurasi 97.54%.

D. Neural Network

Pada tahun 2016 mulai dikembangkan pos tagger menggunakan arsitektur neural network oleh Abka dkk[20]. Penelitian Abka menggunakan 1 hidden layer dengan *word embedding* (CBOW, skip-gram dan GloVe). Hasil penelitiannya yaiu skor F1 80% dan akurasi >93%. Data *training* menggunakan corpus Wikipedia Bahasa Indonesia yang tidak dilabeli, data uji menggunakan 250.000 token yang secara manual ditag.

Di tahun 2018, Manik dkk menggunakan fitur morfologis dan kapitalisasi [21] untuk meningkatkan performa pos tagger dengan word embedding yang sudah ada. Arsitektur meural network yang digunakan adalah dua layer masukan, satu layer gabungan dan dua layer tersembunyi. Layer masukan pertama menggunakan *word embedding* (CBOW dan skip-gram) dan layer masukan kedua menggunakan fitur morfologis dan kapitalisasi. Hasilnya memberikan peningkatan yaitu skor F1 94% dan akurasi tertinggi 95% pada data uji yang sama dengan [20].

Teknologi deep neural network merupakan teknologi *state-of-the-art* di beberapa permasalahan NLP, termasuk pos tag [22]. Yuwana dkk [23] melakukan evaluasi model pada arsitektur neural network untuk pos tagger Indonesia. Hasilnya adalah model dengan dua hidden layer memberikan hasil lebih baik dari jumlah hidden layer yang lebih banyak.

Neural network terbukti lebih baik dibandingkan dengan rule-based dan CRF, penelitian Kurniawan di tahun 2018 menunjukkan metode bidirectional LSTM dan CRF unggul dengan skor F1 97.47%, sedangkan rule based memberikan skor F1 85.77% dan CRF memberikan skor F1 96.22% terhadap IDN Tagged Corpus. Penelitian ini mengklaim metodenya merupakan *state-of-the-art* untuk dataset IDN Tagged Corpus (lihat Tabel IV).

Penelitian lain menggunakan metode yang sama (bi-LSTM dan CRF) menghasilkan akurasi yang juga baik yaitu 95.68% [8]. Penelitian ini menggunakan dataset yang dibangun sendiri yaitu 355.000 token yang diambil dari artikel berita online dan dilabeli secara manual.

TABEL IV
F1 POS TAGGER NN UNTUK IDN TAGGED CORPUS

Metode	F1
NN 1-layer	80%
NN 2-layer	94%
biLSTM+CRF	97.47%
CRF	96.22%
Rule base	85.77%

V. KESIMPULAN

Corpus yang tersedia untuk pos tag berupa token yang sudah dilabeli pos Indonesia secara manual antara lain Pisceldo (14.165 token), Wicaksono (15.000 token), IDN Tagged Corpus (250.000 token), dan Sihui (355.000 token). Sejauh ini IDN Tagged Corpus merupakan corpus yang paling banyak digunakan dan tersedia online.

Tagset Indonesia saat ini belum terstandarkan, namun ada upaya untuk menstandarkan tagset dengan project Universal Dependencies dan juga konvensi tagset yang dibuat oleh INACL.

Pendekatan yang cukup menjanjikan untuk pos tagger Indonesia ke depannya yaitu penggunaan arsitektur neural network dengan menggunakan *bidirectional* LSTM dan CRF, terdapat dua penelitian dengan metode yang sama di tahun 2018 dengan hasil yang sangat baik yaitu skor F1 97.47% [24] dan akurasi 95.68% [8].

UCAPAN TERIMA KASIH

Terima kasih kepada Bu Ayu Purwarianti dan tim dari INACL yang telah membangun tagset Bahasa Indonesia dan membagikan tagset konvensi INACL untuk keperluan penulisan artikel ini. Penelitian ini didanai oleh Lemlitbang Universitas Muhammadiyah Prof. Dr. HAMKA (UHAMKA).

REFERENSI

[1] S. D. Larasati, "IDENTIC Corpus: Morphologically Enriched Indonesian - English Parallel Corpus," in *LREC*, 2012, pp. 902–906.

[2] A. Nurwidyantoro and E. Winarko, "Parallelization of Maximum Entropy POS Tagging for Bahasa Indonesia with MapReduce," *Int. J. Comput. Sci. Issues*, vol. 9, no. 4, pp. 1–6, 2012.

[3] A. F. Wicaksono and P. Ayu, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," in *Prooceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop*, 2010, no. August, pp. 1–7.

[4] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian Part of Speech Tagset and Manually Tagged Indonesian Corpus," in *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014*, 2014, pp. 66–69.

[5] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian Rule-Based part-of-speech Tagger," in *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014*, 2014, pp. 70–73.

[6] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic Part Of Speech Tagging for Bahasa Indonesia," in *Third International MALINDO Workshop*, 2009, no. May, pp. 1–6.

[7] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus," in *International Workshop on Systems and Frameworks for Computational Morphology*, 2011, pp. 119–129.

[8] S. Fu, N. Lin, G. Zhu, and S. Jiang, "Towards Indonesian Part-of-Speech Tagging : Corpus and Models," in *Proceedings of LREC 2018 Workshop on Belt and Road LRE*, 2018, vol. 1, pp. 2–7.

[9] S. Sari, H. Hayurani, M. Adriani, and S. Bressan, "Developing part of speech tagger for bahasa indonesia using brill tagger," in *The International Second MALINDO Workshop*, 2008.

[10] F. Pisceldo, R. Mahendra, R. Manurung, and I. W. Arka, "A Two-Level Morphological Analyser for the Indonesian Language," in *Proceedings of the 2008 Australasian Language Technology Association Workshop (ALTA 2008)*, 2008, vol. 6, pp. 142–150.

[11] S. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (morphind): Towards an indonesian corpus," *Int. Work. Syst.*, 2011.

[12] R. S. Yuwana, A. R. Yuliani, and H. F. Pardede, "On Part of Speech Tagger for Indonesian Language," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017, pp. 369–372.

[13] H. Sujaini, K. Kuspriyanto, A. Akhmad Arman, and A. Purwarianti, "A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 12, no. 3, p. 581, 2014.

[14] A. Hamzah and N. Widyastuti, "Document Subjectivity and Target Detection in Opinion Mining using HMM POS-Tagger," in *2015 International Conference on Information \& Communication Technology and Systems (ICTS)*, 2015, pp. 83–88.

[15] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 2016, pp. 1–5.

[16] Muljono, U. Afini, C. Supriyanto, and R. A. Nugroho, "The Development of Indonesian POS Tagging System for Computer-Aided Independent Language Learning," *Int. J. Emerg. Technol. Learn.*, vol. 12, no. 11, pp. 138–150, 2017.

[17] K. Widhiyanti and A. Harjoko, "POS Tagging for Bahasa Indonesia dengan HMM dan Rule Based," *INFORMATIKA*, vol. 8, no. 2, pp. 151–167, 2012.

[18] Muljono, U. Afini, and C. Supriyanto, "Morphology Analysis for Hidden Markov Model based Indonesian part-of-speech Tagger," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, 2017, pp. 237–240.

[19] F. Ramadhanti, Y. Wibisono, and R. A. Sukamto, "Analisis Morfologi untuk Menangani Out-of-Vocabulary Words pada Part-of-Speech Tagger Bahasa Indonesia Menggunakan Hidden Markov Model," *J. Linguist. Komputasional*, vol. 2, no. 1, p. 6, 2019.

[20] A. F. Abka, "Evaluating the use of word embeddings for part-of-speech tagging in Bahasa Indonesia," in *2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2016, pp. 209–214.

[21] L. P. Manik, A. Ferti Syafiandini, H. F. Mustika, A. Fatchuttamam Abka, and Y. Rianto, "Evaluating the Morphological and Capitalization Features for Word Embedding-Based POS Tagger in Bahasa Indonesia," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2018, pp. 49–53.

[22] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," *arXiv Prepr. arXiv1603.01354*, 2016.

[23] R. S. Yuwana, E. Suryawati, and H. F. Pardede, "On Empirical Evaluation of Deep Architectures for Indonesian POS Tagging Problem," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2018, pp. 204–208.

[24] K. Kurniawan and A. F. Aji, "Toward a Standardized and More Accurate Indonesian Part-of-Speech Tagging," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 303–307.

Mia Kamayani - Perkembangan Part-of-Speech Tagger Bahasa Indonesia

ORIGINALITY REPORT

9%	7%	4%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Telkom University Student Paper	1%
2	openlibrarypublications.telkomuniversity.ac.id Internet Source	1%
3	repository.uhamka.ac.id Internet Source	1%
4	docplayer.info Internet Source	1%
5	Erika Lukman. "Evaluasi aspek teknis terhadap kegiatan penangkapan ikan kakap merah (Lutjanus sp) dan pengembangannya di sekitar perairan Sinjai Teluk Bone", Agrikan: Jurnal Agribisnis Perikanan, 2013 Publication	1%
6	Submitted to Udayana University Student Paper	<1%
7	journal.nurulfikri.ac.id Internet Source	<1%

8	repository.upi.edu Internet Source	<1 %
9	123dok.com Internet Source	<1 %
10	Ayu Faradillah, Windia Hadi, Asih Miatun, Hikmatul Khusna. "PELATIHAN Pelatihan Pembelajaran Matematika yang Efektif melalui Metode Hypnoteaching", Jurnal SOLMA, 2018 Publication	<1 %
11	journals.ums.ac.id Internet Source	<1 %
12	jtiik.ub.ac.id Internet Source	<1 %
13	nanopdf.com Internet Source	<1 %
14	repository.unika.ac.id Internet Source	<1 %
15	jurnal.untan.ac.id Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		