DISERTASI

KERANGKA KERJA TOKENISASI BERDASARKAN STRUKTUR KALIMAT BAHASA INDONESIA

JOHANNES PETRUS
NIM. 03043621924006

FAKULTAS TEKNIK
PROGRAM STUDI DOKTOR ILMU TEKNIK
UNIVERSITAS SRIWIJAYA
2023

DISERTASI

KERANGKA KERJA TOKENISASI BERDASARKAN STRUKTUR KALIMAT BAHASA INDONESIA

JOHANNES PETRUS
NIM. 03043621924006

FAKULTAS TEKNIK
PROGRAM STUDI DOKTOR ILMU TEKNIK
UNIVERSITAS SRIWIJAYA
2023

HALAMAN PENGESAHAN
KERANGKA KERJA TOKENISASI BERDASARKAN STRUKTUR KALIMAT BAHASA INDONESIA

DISERTASI

Diajukan untuk melengkapi salah satu syarat
Memperoleh gelar Doktor dalam bidang ilmu Teknik Informatika

oleh
Johannes Petrus
03043621924006

Palembang, 28 Juli 2023

Promotor

Dr. Ermatita, M.Kom.
NIP. 196709132006042001

Ko-Promotor 1

Dr. Ir. Sukemi, M.T.
NIP. 196612032006041001

Ko-Promotor 2

Prof. Dr. Erwin, S.Si.,M.Si.
NIP. 197101291994121001

Ketua Program Studi

Prof. Dr. Ir. Nukman, M.T.
NIP. 195903211987031001

Mengetahui,
Dekan Fakultas Teknik

Prof. Dr. Eng. Ir. Joni Arliansyah, M.T.
NIP. 196706151995121002

HALAMAN PERSETUJUAN

Karya tulis ilmiah berupa Laporan Disertasi ini dengan judul "Kerangka Kerja Tokenisasi Berdasarkan Struktur Kalimat Bahasa Indonesia" telah dipertahankan dihadapan Tim Penguji Karya Tulis Ilmiah Program Studi Doktor Ilmu Teknik Fakultas Teknik Universitas Sriwijaya pada tanggal 4 Juli 2023.

Palembang, 4 Juli 2023

Tim Penguji Karya Tulis Ilmiah berupa Laporan Disertasi

Ketua

Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T. ( )
NIP. 197502112003121002

Anggota

1. **Dr. Ade Silvia Handayani, S.T., M.T.** ( )
   NIP. 197609302000032002

2. **Dr. Bambang Tutuko, M.T.** (2/8/2023)
   NIP. 196001121989031002
   Bambang Tutuko

3. **Dr. Firdaus, S.T.,M.Kom.** ( )
   NIP. 197801212008121003

Mengetahui,
Dekan Fakultas Teknik

Prof. Dr. Eng. Ir. Joni Arliansyah, M.T.
NIP. 196706151995121002

Koordinator Program Studi

Prof. Dr. Ir. Nukman, M.T.
NIP. 195903211987031001

KATA PENGANTAR

Salam Sejahtera.

Puji Syukur kepada Tuhan Yang Maha Esa, atas berkat dan anugerahNya, penulis dapat menyelesaikan Disertasi yang berjudul **Kerangka kerja Tokenisasi berdasarkan Struktur Kalimat Bahasa Indonesia**.

Penyelesaian Disertasi ini tidak terlepas dari peran serta, bimbingan dan dukungan dari banyak pihak, untuk itu penulis ingin menyampaikan ucapan terima kasih kepada :

1. Bapak Prof. Dr. Eng. Ir. Joni Arliansyah, M.T., selaku Dekan Fakultas Teknik Universitas Sriwijaya.
2. Bapak Prof. Dr. Ir. Nukman, M.T., selaku koordinator Program Studi Ilmu Teknik Program Doktor, Fakultas Teknik, Universitas Sriwijaya.
3. Bapak Prof. Ir. Subriyer Nasir, MS., Ph.D., selaku orang pertama yang memperkenalkan Program Doktor Fakultas Teknik Universitas Sriwijaya kepada penulis.
4. Bapak Dr. Bhakti Yudho Suprapto, S.T., M.T., selaku Wakil Dekan Fakultas Teknik Universitas Sriwijaya dan sekaligus Ketua Tim Penguji.
5. Ibu Dr. Ermatita, M.Kom., selaku Promotor.
6. Bapak Dr. Ir. Sukemi, M.T. selaku ko-Promotor pertama.
7. Bapak Prof. Dr. Erwin, S.Si., M.Si., selaku ko-Promotor kedua.
8. Ibu. Dr. Ade Silvia Handayani, S.T., M.T., bapak Prof. Dr. Bambang Tutuko, M.T., bapak Dr. Firdaus, S.T., M.Kom, yang telah memberikan masukan dan saran yang berharga pada saat ujian Disertasi.
9. Ibu Dian Palupi Rini, S.Si., M.Kom.,Ph.D., yang telah memberikan masukan dan saran yang berharga pada saat Seminar Proposal.
10. Bapak Alexander Kurniawan, selaku Pembina Yayasan Multi Data Palembang atas dukungannya.
11. Istri dan Ananda yang selalu mendampingi dengan sabar selama proses kuliah dan penyelesaian Disertasi ini.
12. Tim komunitas *Doctoral Research* yang banyak memberikan saran dan masukan.
13. Ibu Evi Susantri Sinaga, S.Pd., yang menjadi mitra dalam penyempurnaan dataset.
14. Teman dan sahabat dari Universitas MDP yang terus memberikan semangat, dukungan dan bantuan yang luar biasa.
15. Semua pihak yang telah membantu penyelesaian Disertasi ini yang tidak dapat penulis sebutkan satu per-satu.

Semoga Tuhan Yang Maha Esa membalas semua kebaikan dari bapak dan ibu semua.

Palembang, 28 Juli 2023

Penulis

HALAMAN PERNYATAAN INTEGRITAS

Saya yang bertanda tangan dibawah ini :

N a m a	:	**Johannes Petrus**
N.I.M.	:	03043621924006
Judul	:	Kerangka Kerja Tokenisasi berdasarkan Struktur Kalimat Bahasa Indonesia

Menyatakan bahwa Disertasi saya merupakan hasil karya saya sendiri didampingi tim promotor dan ko-promotor dan bukan hasil jiplakan / plagiat. Apabila ditemukan unsur penjiplakan / plagiat dalam Disertasi ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai aturan yang berlaku.

Demikian pernyataan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.

Palembang, 28 Juli 2023

Johannes Petrus

NIM. 03043621924006

RINGKASAN

KERANGKA KERJA TOKENISASI BERDASARKAN STRUKTUR KALIMAT BAHASA INDONESIA

Karya Tulis Ilmiah berupa Disertasi, Juli 2023

Johannes Petrus, M.T.I.; dibimbing oleh Dr. Ermatita, M.Kom,
Dr. Ir. Sukemi, M.T., dan Prof. Dr. Erwin, S.Si.,M.Si.

Program Studi Doktor Ilmu Teknik, Fakultas Teknik, Universitas Sriwijaya.

Kerangka Kerja Tokenisasi berdasarkan Struktur Kalimat Bahasa Indonesia.

Penelitian ini bertujuan untuk membangun sebuah kerangka kerja tokenisasi berdasarkan struktur kalimat bahasa Indonesia. Tokenisasi akan menghasilkan token baik berupa kata tunggal maupun multi kata, yang berbeda dengan konsep tokenisasi umum yang hanya menghasilkan token kata tunggal saja. Untuk menghasilkan token seperti diatas, Penelitian Disertasi ini menggunakan metode ekstraksi struktur kalimat yang menghasilkan fungsi-fungsi kalimat sebagai sebuah token. Metode ini merupakan hal yang baru karena sejauh yang penulis ketahui proses ekspresi multi kata (*multi word expression*) menggunakan metode statistik, linguistik, kamus dan jaringan neural.
Hasil ekstraksi struktur kalimat berupa unsur fungsi kalimat seperti Subjek, Predikat, Objek, Pelengkap dan Keterangan. Sebuah kalimat minimal terdiri dari Subjek dan Predikat. Masing-masing fungsi kalimat dapat berupa sebuah kata atau gabungan beberapa kata. Gabungan beberapa kata tersebut dapat menjadi token multi kata
Penelitian ini menerapkan pembelajaran mesin untuk melakukan ekstraksi struktur kalimat, dengan terlebih dahulu membangun sebuah dataset struktur kalimat bahasa Indonesia. Ekstraksi struktur kalimat dalam penelitian ini hanya dilakukan terhadap kalimat tunggal dan berjenis kalimat aktif. Dalam percobaan mengekstrak struktur dari 100 kalimat dan membandingkan token yang diprediksi dengan token yang seharusnya, diperoleh nilai *Precision* sebesar 0,92 dan nilai *Recall* sebesar 0,86.

Kata Kunci : Kerangka Kerja, Multi Kata, Segmentasi Kalimat, Struktur Kalimat, Token.

SUMMARY

A TOKENIZATION FRAMEWORK BASED ON INDONESIAN SENTENCE STRUCTURE

Scientific papers in the form of a Dissertation, July 2023

Johannes Petrus, M.T.I.; supervised by Dr. Ermatita, M.Kom,
Dr. Ir. Sukemi, M.T., and Prof. Dr. Erwin, S.Si.,M.Si.

Doctor of Engineering Study Program, Faculty of Engineering, Sriwijaya University.

A Tokenization Framework Based On Indonesian Sentence Structure

This research aims to build a tokenization framework based on Indonesian sentence structure. Tokenization will produce tokens in the form of single words or multi words, which is different from the general tokenization concept which only produces single word tokens. To generate tokens as above, the method used in this research is sentence structure extraction which returns the sentence's function as a token. This method is something new because as far as the author knows, the multi word expression process uses statistical, linguistics, dictionaries and neural network methods.

The results of sentence structure extraction are in the form of sentence function elements such as Subject, Predicate, Object, Complement and Adverbs. A sentence at least consists of a Subject and a Predicate. Each sentence function can be a single word or a combination of several words. The combination of several words can be a multi-word token.

This research applies machine learning to perform sentence structure extraction, by first building an Indonesian sentence structure dataset. The extraction of sentence structure in this research is only done on single sentences and active sentences. In the experiment of extracting the structure of 100 sentences and comparing the predicted tokens with the supposed tokens, a Precision value of 0.92 and a Recall value of 0.86 were obtained.

Keywords: Framework, Multi-word, Sentence Segmentation, Sentence Structure, Token.

PUBLIKASI

Publikasi Utama

Judul Artikel #1	:	*A Novel Approach: Tokenization Framework based on Sentence Structure in Indonesian Language*
DOI	:	10.14569/IJACSA.2023.0140264
Nama Jurnal	:	*International Journal of Advanced Computer Science and Applications (IJACSA)*
	:	Volume 14 Issue 2, (Maret 2023)
ISSN	:	2158-107X (Print) / 2156-5570 (Online)
Negara	:	*United Kingdom* (UK)
Terindeks	:	Scopus (Elsevier)
	:	*Web of Science* (*Clarivate / Thomson Reuters*)
SJR	:	0,26

Judul Artikel #2	:	*An adaptable sentence segmentation based on Indonesian rules*
DOI	:	10.11591/ijai.v12.i3.pp1491-1499
Nama Jurnal	:	*IAES International Journal of Artificial Intelligence (IJ-AI)*
	:	Volume 12 Issue 3, (September 2023)
ISSN/e-ISSN	:	2089-4872/2252-8938
Negara	:	Indonesia
Terindeks	:	Scopus (Elsevier)
	:	*Web of Science* (*Clarivate / Thomson Reuters*)
SJR	:	0,35

Publikasi Pendahuluan

Nama Konferensi	:	*2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*
Judul	:	*Soft and Hard Clustering for Abstract Scientific Paper in Indonesian*
Tahun	:	2019
DOI	:	10.1109/ICIMCIS48181.2019.8985231
Penerbit / Index	:	IEEE Xplore / Scopus
Penghargaan	:	*Best Paper*

DAFTAR ISI

DAFTAR TABEL

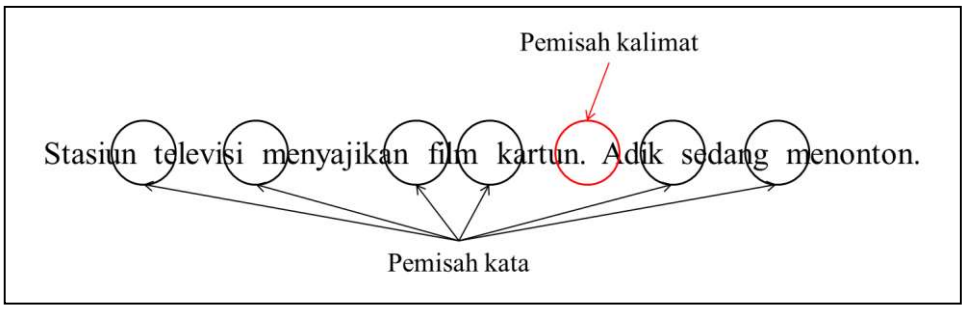
DAFTAR GAMBAR


UNIVERSITAS
SRIWIJAYA
ILMU ALAT PENGABDIAN

BAB I.
PENDAHULUAN

1.1. Latar Belakang

Tokenisasi adalah proses memecah sebuah string menjadi unit-unit terkecilnya yang dinamakan sebagai token. Unit terkecil sebuah paragraf adalah kalimat dan unit terkecil sebuah kalimat adalah kata. Kalimat akan terpisah dengan kalimat lain dan kata akan terpisah dengan kata yang lain karena ada pemisahnya atau ada batasnya. Batas kalimat ditandai dengan penggunaan salah satu tanda baca titik, tanya atau seru serta diikuti dengan spasi dan huruf besar pada kata berikutnya. Sementara batas atau pemisah untuk kata adalah spasi. Spasi dapat berada dibelakang kata, didepan kata atau didepan dan dibelakang sebuah kata. Gambar 1.1 berikut ini menunjukkan batas pemisah kalimat dan kata.



Gambar 1. 1. Batas pemisah kalimat ditandai dengan tanda baca titik dan batas pemisah kata ditandai dengan spasi.

Menentukan batas kata lebih mudah dari batas kalimat. Menentukan batas kalimat menghadapi kendala yang berhubungan dengan tanda baca, singkatan, dan karakter dalam tanda kurung [1]. Menentukan tanda baca sebagai akhir kalimat, menjadi hal yang krusial. Tanda titik selain sebagai tanda akhir kalimat juga memiliki banyak peran lain hingga menimbulkan ambiguitas.

Pada Tabel 1.1. tokenisasi kalimat pada paragraf pertama dapat dilakukan dengan benar namun pada paragraf ke-2 terdapat singkatan “Bpk.” dengan tanda titik diakhirnya dan sesuai dengan ciri sebuah akhir kalimat namun sebenarnya bukan akhir kalimat. Kondisi yang membingungkan / ambigu ini membuat proses tokenisasi kalimat menjadi keliru.

Tabel 1. 1. Tokenisasi kalimat, memisahkan setiap kalimat yang ada dalam sebuah paragraf menjadi kalimat-kalimat individual.

Paragraf ke-1	Stasiun televisi menyajikan film kartun. Adik sedang menonton.
Token kalimat ke-1	Stasiun televisi menyajikan film kartun.
Token kalimat ke-2	Adik sedang menonton.
Paragraf ke-2	Bpk. Presiden memimpin rapat kabinet. Menteri mendengarkan arahan.
Token kalimat ke-1	Bpk.
Token kalimat ke-2	Presiden memimpin rapat kabinet.
Token kalimat ke-3	Menteri mendengarkan arahan.

Tokenisasi merupakan proses yang fundamental pada hampir seluruh aplikasi pemrosesan bahasa alami. Tokenisasi perlu dilakukan agar setiap token dapat diubah menjadi data vektor untuk proses analitik teks yang lebih baik. Pemrosesan bahasa alami sangat bergantung pada hasil tokenisasi.

Pendekatan standard tokenisasi adalah terbentuknya token kata tunggal, yaitu token yang berupa semua kata-kata individual yang dipisahkan oleh spasi [2]–[4], seperti contoh pada tabel 1.2. dibawah ini.

Tabel 1. 2. Tokenisasi kata tunggal, setiap kata yang diakhiri spasi atau berada diantara spasi atau diawali dengan spasi, akan menjadi token sendiri.

Kalimat Masukan	Hasil tokenisasi kata tunggal				
Stasiun televisi menyajikan film kartun	Stasiun	televisi	menyajikan	film	kartun

Token kata tunggal telah diimplementasikan dalam banyak penelitian diantaranya oleh [5]–[8] untuk bidang *Information Retrieval*, [9]–[11] untuk bidang *document Clustering,* [12]–[14] untuk bidang *Information Extraction*, [15]–[20] untuk bidang analisa sentimen (*sentiment analysis*), serta [21] untuk bidang *Classification*.

Sejumlah publikasi menyatakan token dapat berupa kata tunggal, gabungan kata (multi kata) bahkan kalimat [10]–[13][22]. Token multi kata atau gabungan kata terdiri lebih dari 1 (satu) kata yang secara bersama-sama menjadi sebuah token, seperti ditunjukkan pada Tabel 1.3 berikut ini.

Tabel 1. 3. Token multi kata, token yang terdiri dari kata yang bergabung dengan kata lain disebelahnya agar memiliki arti yang benar.

Kalimat Masukan	Hasil tokenisasi kata tunggal		
Stasiun televisi menyajikan film kartun	Stasiun televisi	menyajikan	film kartun

Pada Tabel 1.3. diatas terdapat dua kata “stasiun televisi” dan “film kartun” yang bergabung dan memiliki arti yang benar sesuai konteks kalimat yaitu sebagai sebuah lembaga penyiaran dan sebagai sebuah produk jenis film. Apabila dipisahkan menjadi “stasiun” dan “televisi” maka arti semula sebagai sebuah lembaga penyiaran dapat bermakna menjadi tempat kereta berhenti dan alat media informasi.

Penelitian tentang token gabungan kata telah ada, berupa penelitian tentang ekspresi multi kata (*multiword expression / MWE*) yang bertujuan guna menentukan pasangan kata seperti yang dilakukan oleh [23]–[30]. Penelitian tersebut umumnya dilakukan untuk dokumen berbahasa Inggris atau bahasa lainnya termasuk bahasa yang tidak mengenal spasi seperti bahasa Mandarin, Jepang atau Thai.

Penelitian serupa untuk teks berbahasa Indonesia masih sedikit, namun telah ada yang melakukannya seperti oleh [23][31], [32]. Peneliti [23] melakukan pembentukan gabungan kata berupa frasa maksimal 2 kata tanpa preposisi (kata depan) atau konjungsi (kata penghubung), peneliti [31] menggunakan teknik statistik untuk menemukan pasangan kata yang terbatas hanya untuk pasangan 2 kata, dan peneliti [32] menggunakan metode kamus dan linguistik dengan penggunaan *POS tag noun, verb* dan *adjective* saja.

Beberapa metode yang digunakan untuk memperoleh ekspresi multi kata yaitu Statistik, Linguistik (*rule-based*), Kamus dan Pembelajaran Mesin serta kombinasi dari metode-metode tersebut. Metode Statistik dan Linguistik merupakan metode yang sangat dominan digunakan yaitu dengan menghitung frekwensi kemunculan bersama dua kata dan penggunaan pola-pola kombinasi jenis kata.

Penelitian yang menggunakan metode Statistik dilakukan oleh [31] melalui proses ekstraksi dan menemukan pasangan kata dari dokumen teks berbahasa Indonesia. Penelitian ini terbatas hanya untuk pasangan 2 (dua) kata atau frasa dan hasil ekstraksi diteliti pengaruhnya terhadap kinerja *clustering*. Perhitungan Statistik dilakukan dengan memberikan *threshold* minimal. Hasilnya menunjukkan bahwa penggunaan frasa saja memiliki kinerja yang rendah dari pada campuran kata dan frasa. Sementara itu penelitian oleh [33] dengan metode uji statistik akan menetapkan pasangan kata sebagai ekspresi multi kata apabila lebih sering muncul dalam web *corpus*.

Penelitian dengan metode Linguistik dikerjakan oleh [34] dengan membentuk gabungan kata dalam bentuk *name entity* dan kata majemuk dengan konstruksi V+V, N+V, Adj+V. Peneliti [35] melakukan ekstraksi multi kata dalam bentuk frasa kata benda dengan menggunakan pola (*Noun Preposition Noun*), (*Noun Preposition Noun + Noun*) dan (*Noun Preposition Noun Preposition Noun*). Akurasi Hasil ekstraksi rata-rata 97%.

Sementara itu peneliti yang menggunakan gabungan metode Statistik dan Linguistik (*hybrid*) seperti [27] melakukan ektraksi multi kata dengan metode Statistik (DC, PMI, MDC) serta menggunakan struktur pola Linguistik untuk n-gram menggunakan 3 pola *Compound Noun, Noun Noun* dan *Verb noun,* sementara untuk bigram menggunakan 12 pola (A-Adv), (A-N), (A-P), (Adv-A), (N-A), (N-P), (N-V), (P-N), (V-Adv), (V-Par), (V-P), (V-V). Peneliti [29] mengidentifikasi istilah bidang kimia sebagai ekspresi multi kata. Peneliti [36] menggunakan kombinasi metode Statistik dan *Rule-based* untuk mengekstrak ekspresi multi kata bilingual bahasa Mandarin dan Inggris. Penelitian oleh [37] menggunakan *Association measure* dan linguistik untuk memberikan skor pada kandidat multi kata. Peneliti [38] dengan membangun kamus terminologi dan morfologi kata majemuk bidang *agricultural engineering* dan menghitung frekwensi kemunculan kandidat dalam korpus bahasa Serbia. Kebanyakan MWT memiliki struktur sintaksis AN (kata sifat yang diikuti kata benda) dan NNgi (Kata benda yang diikuti oleh dua kata sifat / kata benda). Peneliti [39] mengidentifikasi kandidat frasa yang terdiri dari kata-kata *closed-class* dan memeriksa kemunculan bersama kolokasinya. Frasa harus melebihi ambang batas frekwensi minimum. Peneliti [40] melakukan ekstraksi MWE dengan metode koefisien dari *Pearson Chi Square*; koefisien *Dice*; *Pointwise –Mutual Information* (PMI) dan metode Linguistik menggunakan POS (*part of speech*). Peneliti [41] melakukan ekstraksi frasa kata kerja berdasarkan *POS tagged pattern based extraction* yang sesuai dan pengukuran skor asosiasi seperti *Maximum likelihood Estimator* (MLE), *dice coefficient* (DC) dan *t-score (*TS). Sedangkan peneliti [42], mengidentifikasi ekspresi multi kata (MWE) menggunakan *Bayesian Networks* mengekspresikan fitur Linguistik yang saling bergantung (*interdependen*) untuk keperluan klasifikasi.

Metode Pembelajaran Mesin dimanfaatkan oleh [25] yang menggunakan jaringan saraf Bi-LSTM, untuk memodelkan urutan informasi (*sequence information*), *phrase head word expansion* untuk mengetahui kata yang lebih sering muncul dalam sebuah frasa serta *correlation degree* untuk menghitung tingkat probabilitas kedekatan dua POS yang bersebelahan dan *K-Means clustering* untuk mengidentifikasi tiga bentuk MWE (kata benda majemuk, konstruksi kata kerja dan idiom). Peneliti [30] menggunakan model *supervised* dimana kandidat ekspresi multi kata diekstrak menggunakan *Support Vector Machine* (SVM) hingga 7 gram.

Metode kamus digunakan oleh [24] dengan dua sumber yaitu set korpora tertulis dan kamus istilah khusus dan idiom yang dikumpulkan dari internet. Penelitian untuk bahasa Turki ini selain menggunakan kamus juga menggunakan beberapa metode lain yaitu Statistik untuk

menghitung frekwensi kemunculan bersama yang tinggi, Linguistik melalui pola-pola POS, kamus idiom dan kamus istilah-istilah untuk domain khusus.

Penelitian-penelitian diatas, secara umum penekanannya untuk memastikan kombinasi 2 kata atau lebih adalah ekspresi multi kata atau bukan, sementara tokenisasi lebih fokus kepada bagaimana memisahkan kata dengan kata tetangganya yang menggunakan karakter spasi sebagai pemisahnya dengan hasil token kata tunggal. Token kata tunggal dapat mengakibatkan makna kata menyimpang jauh dari konteks sebenarnya [31]. Beberapa kata yang seharusnya bergabung harus tetap bersatu sebagai token multi kata. Setiap unsur dari token multi kata dituliskan terpisah dengan spasi, namun spasi tidak dapat digunakan sebagai pemisah. Hal ini menandakan tokenisasi tidak dapat dipandang sebagai masalah yang simpel yaitu hanya menemukan spasi untuk mengetahui batas sebuah kata karena pada kenyataannya spasi tidak selalu menjadi pemisah kata.

Token multi kata dapat mewakili isi dokumen dengan lebih baik [43]. Dua kata atau lebih yang bergabung menjadi satu sebagai sebuah token, haruslah berada pada satu fungsi kalimat yang sama atau kata-kata yang berada pada fungsi kalimat yang berbeda tidak mungkin dapat bergabung. Fungsi kalimat adalah Subjek (S), Predikat (P), Objek (O), Pelengkap (E) dan Keterangan (K). Dengan demikan menunjukkan bahwa fungsi kalimat juga mempunyai peran penting dalam menghasilkan token multi kata. Gabungan fungsi kalimat membentuk struktur kalimat.

Setiap fungsi kalimat harus berada pada kalimat yang sama, sehingga sangat penting dapat memisahkan kalimat individual dengan benar dari paragrafnya. Kalimat yang berdiri sendiri akan mudah untuk diekstrak struktur kalimatnya. Hasil ekstraksi struktur kalimat adalah diperolehnya unsur-unsur fungsi kalimat yang ada pada kalimat dan setiap fungsi kalimat tersebut dapat berisi hanya satu kata atau gabungan kata yang kemudian dapat menjadi token. Metode Pembelajaran Mesin dapat digunakan untuk melakukan ekstraksi struktur kalimat dengan terlebih dahulu mempersiapkan *dataset.*

Penelitian Disertasi ini bertujuan untuk melakukan tokenisasi kalimat Bahasa Indonesia. Token yang dihasilkan akan berupa token kata tunggal dan token multi kata. Metode yang digunakan adalah melalui ekstraksi atau penguraian struktur kalimat, sedangkan tokenisasi atau segmentasi kalimat menggunakan metode berbasis aturan linguistik. Hasil penelitian ini dibuat dalam bentuk sebuah kerangka kerja.

1.2. Permasalahan

Tokenisasi berfungsi untuk memisahkan bagian kalimat dalam dokumen berdasarkan kata per-kata individual dan menggunakan spasi sebagai satu-satunya parameter pemisah kata. Dalam bahasa Indonesia dan banyak bahasa negara lain terdapat kata-kata yang tidak boleh dipisahkan karena memisahkan kata-kata tersebut dapat memberikan arti yang berbeda atau bahkan menghilangkan maknanya.

Setelah melakukan tinjauan pustaka secara sistematis maka diperoleh permasalahan dalam proses tokenisasi adalah sebagai berikut :

1. Tanda titik berperan penting untuk menentukan batas kalimat dalam proses tokenisasi kalimat, namun tanda baca ini sering menimbulkan kebingungan / ambigu karena tanda titik juga memiliki fungsi selain sebagai tanda batas kalimat.
2. Tokenisasi adalah aksi yang sangat penting dilakukan sebelum data teks diproses lebih lanjut. Tokenisasi yang umum dikenal saat ini adalah tokenisasi kata per-kata dengan spasi sebagai tanda pemisah antar kata, tetapi ada kata-kata yang tidak boleh dipisahkan (harus tetap berpasangan).
3. Penelitian tentang ekspresi multi kata yang menggunakan metode Statistik, Linguistik (*Rule based*), Kamus dan Pembelajaran Mesin / jaringan neural, mampu untuk menemukan pasangan kata sebagai satu kesatuan, tetapi tidak memperhatikan bahwa kata-kata yang bergabung harus berada pada fungsi kalimat yang sama sesuai dengan kaidah tata bahasa.
4. Tokenisasi yang menggunakan spasi sebagai pemisah hanya menghasilkan token kata tunggal, disisi lain aplikasi ekspresi multi kata ditujukan untuk menemukan pasangan kata. Masing-masing proses berdiri sendiri dan belum ditemukan metode tokenisasi teks yang menghasilkan token kata tunggal dan multi kata (pasangan kata) sekaligus.
5. Komputer mampu memprediksi struktur kalimat dengan mudah apabila komputer telah mempelajari sekumpulan data (*dataset*) yang berisikan pola struktur kalimat, namun ketersediaan *dataset* demikian belum ada sama sekali secara publik.

1.3. Kontribusi dan *Novelty*

Sebagai hasil dari studi empiris serta eksperimen yang dilakukan, berikut merupakan kontribusi dari penelitian Disertasi ini :

1. Tokenisasi atau segmentasi kalimat tidak hanya memperhatikan penggunaan tanda baca akhir kalimat seperti tanda titik yang masih menimbulkan ambiguitas, tetapi harus memperhatikan fitur-fitur dari kata lain yang ada disebelah kata yang bertanda baca (token kandidat). Penelitian ini melakukan segmentasi kalimat berbasis aturan (*rule-based*) dan memberikan kontribusi dalam bentuk daftar 34 aturan untuk menetapkan status token kandidat sebagai akhir kalimat atau bukan. Aturan-aturan tersebut ditopang dengan 27 fitur. Bagian penelitian ini menghasilkan karya ilmiah pertama yang telah dipublikasikan pada jurnal internasional bereputasi.
2. Penelitian ini memberikan kontribusi dalam bentuk sebuah pendekatan menemukan pasangan kata berdasarkan struktur kalimat. Ekstraksi struktur kalimat ini menjamin diperolehnya pasangan kata atau token multi kata yang berada pada fungsi kalimat yang sama.
3. Penelitian ini memberikan kontribusi dalam bentuk terciptanya sebuah kerangka kerja proses tokenisasi yang baru. Kerangka kerja ini terdiri dari 2 (dua) tahap, yaitu tahap pertama adalah tokenisasi kalimat dan tahap kedua adalah penguraian struktur kalimat. Kerangka kerja tokenisasi ini mampu melakukan segmentasi kalimat serta menghasilkan token kata tunggal dan token multi kata secara bersamaan. Kerangka kerja tokenisasi ini dapat diadaptasi untuk bahasa negara lain yang mengenal spasi dengan sedikit penyesuaian. Bagian penelitian ini menghasilkan karya ilmiah kedua yang telah dipublikasikan pada jurnal internasional bereputasi.
4. Penelitian ini memberikan kontribusi dengan menghasilkan sebuah *dataset* yang berisi kalimat bahasa Indonesia dengan pola struktur untuk kalimat tunggal dan jenis kalimat aktif. *Dataset* semacam ini belum pernah ditemukan dalam portal penyedia *dataset* yang bersifat publik.

Berdasarkan kontribusi yang dihasilkan maka dapat diketahui keterbaruan dari penelitian ini, yaitu:

1. Dalam proses segmentasi kalimat, pengolahan 27 fitur pada token yang berada disebelah kiri dan kanan token kandidat (token yang disertai dengan tanda baca akhir kalimat) merupakan usulan pembaruan untuk mengatasi kelemahannya yang ada pada penelitian sebelumnya seperti tidak dapat mengidentifikasi singkatan nama tengah seseorang (contoh: Nila F. Moeloek).

2. Tokenisasi dengan metode ekstraksi struktur kalimat merupakan invensi baru yang efektif untuk menghasilkan token kata tunggal dan token multi kata atau kombinasi keduanya sekaligus, dimana penelitian sebelumnya hanya untuk token kata tunggal saja atau ekspresi multi kata saja dengan menggunakan metode Linguistik, Statistik, Kamus dan Pembelajaran Mesin atau kombinasi metode.
3. Pembangunan kerangka kerja tokenisasi berdasarkan struktur kalimat merupakan hal baru yang belum pernah dipublikasikan oleh peneliti lain sebelumnya.

1.4. Tujuan Penelitian

Dalam rangka menghasilkan suatu sistem tokenisasi dengan pendekatan yang baru ini, terdapat beberapa tujuan yang ingin dicapai meliputi :

1. Untuk membuat suatu sistem segmentasi kalimat berbasis aturan yang mampu mengenali status sebuah token kandidat dengan mendefinisikan sekumpulan aturan dan fitur.
2. Untuk menghasilkan suatu proses tokenisasi yang menghasilkan gabungan kata sebagai sebuah token atau disebut token multi kata, selain token kata tunggal.
3. Untuk melakukan ekstraksi struktur kalimat yang mampu mengelompokkan kata atau gabungan kata kedalam fungsi kalimat yang sesuai.
4. Untuk merancang sebuah kerangka kerja proses tokenisasi yang lengkap.
5. Untuk membuat sebuah *dataset* yang berisi kumpulan kalimat dengan pola struktur kalimat tertentu dan diimplementasikan pada proses pembelajaran mesin.

1.5. Manfaat Penelitian

Penelitian ini memberikan banyak manfaat untuk pengembangan ilmu pengetahuan terutama dalam dunia penelitian, seperti :

1. Bagi peneliti, penelitian ini akan menjawab pertanyaan tentang proses tokenisasi yang menghasilkan token kata tunggal dan token multi kata sekaligus.
2. Bagi Penelitian masa depan, penelitian ini akan menjadi salah satu rujukan dan inspirasi dalam mengembangkan penelitian yang sejenis.

1.6. Batasan Penelitian

Penelitian ini dibatasi hanya untuk penulisan teks dalam Bahasa Indonesia yang formal yang umum sesuai aturan tata bahasa Indonesia tanpa membatasi pada bidang tertentu atau domain tertentu. Bahasa gaul atau bahasa *slank* tidak dibahas. Penetapan Bahasa Indonesia sebagai obyek sentral pada penelitian ini didasari oleh karena Bahasa Indonesia menggunakan spasi sebagai pembatas kata, juga belum tersedianya dataset untuk bahasa secara global serta perbedaan dalam tata cara penulisan struktur kalimat. Bahasa negara lain juga menggunakan spasi sebagai pembatas kata, tetapi ada bahasa dari negara tertentu yang tidak mengenal spasi seperti bahasa Jepang, bahasa Mandarin, bahasa Korea atau bahasa Thailand. Struktur kalimat bahasa Inggris dipengaruhi oleh waktu penggunaannya (*past, present, future*).

Pembahasan dalam penelitian ini termasuk proses pembentukan *dataset* struktur kalimat dalam bahasa Indonesia. Kumpulan data pada *dataset* dibatasi hanya pada kalimat-kalimat sederhana yang hanya memiliki satu Predikat dan bersifat sebagai kalimat aktif yang positif. Pembangunan *dataset* struktur kalimat dengan proporsi maksimal 3 (tiga) kata untuk Subjek, Predikat dan Objek, sementara untuk Pelengkap dan Keterangan masing-masing maksimal 5 (lima) kata.

1.7. Sistematika Laporan

Laporan Disertasi ini disajikan secara komprehensif dan berurutan yang terdiri dari 7 (tujuh) bab yang dapat dijelaskan sebagai berikut :

BAB I. PENDAHULUAN

Bab ini berisikan uraian mengenai proses tokenisasi yang menjadi alasan penelitian Disertasi ini dilakukan. Uraian tersebut tersesbut didasari dari fenomena, hasil penelitian dari peneliti sebelumnya dan permasalahan yang perlu dicarikan penyelesaiannya sebagai tujuan dari penelitian ini. Bab ini juga menyajikan permasalahan yang timbul dan manfaat yang ingin diperoleh selain pernyataan kontribusi dan *novelty* dari penelitian ini.

BAB II. LANDASAN TEORI

Bab ini berisi teori-teori yang terkait dengan penulisan Disertasi ini serta menguraikan hasil-hasil penelitian yang telah dilakukan oleh peneliti sebelumnya yang juga berkenaan dengan penelitian ini. Bab ini merupakan rujukan dasar untuk membahas objek yang akan diteliti.

BAB III. METODOLOGI PENELITIAN

Bab ini berisi metodologi yang menggambarkan langkah-langkah sistematis yang dilakukan oleh peneliti guna mencapai tujuan penelitian. Langkah-langkah sistematis tersebut diarahkan untuk menghasilkan suatu pengembangan atau penciptaan baru dari metode sebelumnya yang sesuai dengan topik penelitian ini.

BAB IV. SEGMENTASI KALIMAT

Bab ini berisi uraian dalam melakukan segmentasi kalimat dari sebuah teks string. Dalam Bab ini juga disertakan rancangan model yang digunakan (*Rule-based*), persiapan data, mendeteksi batas kalimat, menyatukan kata-kata menjadi kalimat tunggal serta menguji tingkat keberhasilan dalam melakukan segmentasi kalimat.

BAB V. EKSTRAKSI STRUKTUR KALIMAT

Bab ini berisi uraian tentang ektraksi atau pemisahan setiap kata dari sebuah kalimat kedalam fungsi kalimatnya masing-masing. Proses ekstraksi didahului dengan membangun dataset, memberikan label data, melatih dataset dan pengujian model untuk melakukan ekstraksi struktur kalimat.

BAB VI. KERANGKA KERJA TOKENISASI KALIMAT BAHASA INDONESIA

Bab ini menggambarkan kerangka kerja tokenisasi sebagai sebuah usulan baru dari penelitian ini. Secara spesifik, bab ini menjelaskan tentang tokenisasi berdasarkan struktur kalimat, perbandingan token yang dihasilkan berdasarkan spasi dan berdasarkan struktur kalimat, termasuk juga melakukan pengujian kemiripan kalimat berdasarkan token kata tunggal maupun token multi kata.

BAB VII. KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan yang secara singkat menjelaskan pokok-pokok hasil penelitian dan saran untuk pengembangan pada masa yang akan datang.


UNIVERSITAS
SRIWIJAYA
ILMU ALAT PENGABDIAN

DAFTAR PUSTAKA

[1] S. J. Putra, M. N. Gunawan, I. Khalil, and T. Mantoro, "Sentence boundary disambiguation for Indonesian language," *ACM Int. Conf. Proceeding Ser.*, no. September 2018, pp. 587–590, 2017, doi: 10.1145/3151759.3156474.

[2] S. Vijayarani and R. Janani, "Text Mining: open Source Tokenization Tools – An Analysis," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.

[3] H. Lane, C. Howard, and H. M. Hapke, *Natural Language Processing in Action*. Shelter Island, NY 11964: Manning Publications Co, 2019.

[4] G. Wilcock, *Introduction to linguistic annotation and text analytics*, vol. 2, no. 1. 2009.

[5] V. Singh and B. Saini, "An Effective Tokenization Algorithm for Information Retrieval Systems," no. April 2015, pp. 109–119, 2014, doi: 10.5121/csit.2014.4910.

[6] S.Gowri, D. G. S. A. Mala, and Divya.G, "Text Preprocessing for the improvement of Information Retrieval in Digital Textual Analysis," *Int. Conf. Math. Sci.*, no. August 2016, 2014, [Online]. Available: https://www.researchgate.net/publication/306538095.

[7] W. B. Trihanto, R. Arifudin, and M. A. Muslim, "Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Na?ve Bayes Classifier," *Sci. J. Informatics*, vol. 4, no. 2, pp. 179–190, 2017, doi: 10.15294/sji.v4i2.11876.

[8] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, and A. Valencia, "Information retrieval and text mining technologies for chemistry," *Chem. Rev.*, vol. 117, no. 12, pp. 7673–7761, 2017, doi: 10.1021/acs.chemrev.6b00851.

[9] N. Garg, R. G.-I. J. of E. Research, and U. 2016, "Clustering Techniques for Text Mining: A Review," *Int. J. Eng. Res.*, vol. 5, no. 4, pp. 241–243, 2016, [Online]. Available: http://www.ijer.in/ijer/publication/v5s4/IJER_2016_404.pdf.

[10] A. T. Hermawan, Gunawan, and A. Halim, "Document ndex Graph Untuk Phrase Matching Dalam Pengelompokan Dokumen Web Berita Berbahasa Indonesia," *J. Ilm. Teknol. dan Rekayasa - Din. Teknol.*, vol. 5, no. 1, pp. 43–48, 2012.

[11] U. Rahardja, T. Hariguna, and W. M. Baihaqi, "Opinion mining on e-commerce data using sentiment analysis and k-medoid clustering," *Proc. - 2019 12th Int. Conf. Ubi-Media Comput. Ubi-Media 2019*, pp. 168–170, 2019, doi: 10.1109/Ubi-Media.2019.00040.

[12] J. Joseph and J. R. Jeba, "Information Extraction using Tokenization and Clustering Methods," *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 3690–3692, 2019, doi: 10.35940/ijrte.d7943.118419.

[13] M. Kumar, A. Garg, A. Munjal, and A. Tanwar, "Twitter based Information Extraction," *Int. J. New*

Technol. Res., vol. 3, no. 3, pp. 52–55, 2017, [Online]. Available: www.ijntr.org.

[14] S. P. Panda, V. Behera, A. Pradhan, and A. Mohanty, "A Rule-based Information Extraction System," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9, pp. 1613–1617, 2019, doi: 10.35940/ijitee.i8156.078919.

[15] H. Juwiantho, E. I. Setiawan, J. Santoso, and M. H. Purnomo, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2Vec Menggunakan Deep Convolutional Neural Network," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 181–188, 2020, doi: 10.25126/jtiik.202071758.

[16] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," *Proc. - 2016 6th Int. Annu. Eng. Semin. Ina. 2016*, pp. 28–31, 2017, doi: 10.1109/INAES.2016.7821901.

[17] H. Sudira, A. L. Diar, and Y. Ruldeviyani, "Instagram Sentiment Analysis with Naive Bayes and KNN: Exploring Customer Satisfaction of Digital Payment Services in Indonesia," *2019 Int. Work. Big Data Inf. Secur. IWBIS 2019*, pp. 21–26, 2019, doi: 10.1109/IWBIS.2019.8935700.

[18] R. I. Permatasari, M. A. Fauzi, P. P. Adikara, and E. D. L. Sari, "Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes," *3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc.*, pp. 92–95, 2018, doi: 10.1109/SIET.2018.8693195.

[19] Kusrini and M. Mashuri, "Sentiment analysis in twitter using lexicon based and polarity multiplication," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 365–368, 2019, doi: 10.1109/ICAIIT.2019.8834477.

[20] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," *Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017*, vol. 2018-Janua, pp. 266–269, 2017, doi: 10.1109/ICITACEE.2017.8257715.

[21] V. Chunwijitra and C. Wutiwiwatchai, "Classification-based spoken text selection for LVCSR language modeling," *Eurasip J. Audio, Speech, Music Process.*, vol. 2017, no. 1, 2017, doi: 10.1186/s13636-017-0121-5.

[22] S. A. Fahad, "Design and Develop Semantic Textual Document Clustering Model," *J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 26–39, 2017, doi: 10.15640/jcsit.v5n2a4.

[23] D. Gunawan, A. Amalia, and I. Charisma, "Automatic extraction of multiword expression candidates for Indonesian language," *Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016*, no. November, pp. 304–309, 2017, doi: 10.1109/ICCSCE.2016.7893589.

[24] S. K. Metin and M. Taze, "A procedure to build multiword expression data set," *2nd Int. Conf. Comput. Commun. Syst. ICCCS 2017*, pp. 46–49, 2017, doi: 10.1109/CCOMS.2017.8075264.

[25] Y. Liang, H. Tan, H. Li, Z. Wang, and W. Gui, "A language-independent hybrid approach for multi-

word expression extraction," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 3273–3279, 2017, doi: 10.1109/IJCNN.2017.7966266.

[26] M. Attia, A. Toral, L. Tounsi, P. Pecina, and J. Van Genabith, "Automatic Extraction of Arabic Multiword Expressions," *Proc. Work. Multiword Expressions from Theory to Appl. (MWE 2010)*, no. August, pp. 18–26, 2010.

[27] S. Agrawal, R. Sanyal, and S. Sanyal, "Statistics and linguistic rules in multiword extraction: A comparative analysis," *Int. J. Reason. Intell. Syst.*, vol. 6, no. 1–2, pp. 59–70, 2014, doi: 10.1504/IJRIS.2014.063954.

[28] S. Agrawal, R. Sanyal, and S. Sanyal, "Hybrid method for automatic extraction of multiword expressions," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 33–38, 2018, doi: 10.14419/ijet.v7i2.6.10063.

[29] L. Huang and C. Ling, "Representing Multiword Chemical Terms through Phrase-Level Preprocessing and Word Embedding," *ACS Omega*, vol. 4, no. 20, pp. 18510–18519, 2019, doi: 10.1021/acsomega.9b02060.

[30] M. Farahmand and R. Martins, "A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features," *Proc. 10th Work. Multiword Expressions (MWE 2014)*, pp. 10–16, 2014.

[31] A. Hamzah, A. Susanto, F.Soesianto, and J. E. Istyanto, "Comparison of Word and Phrase Features in Clustering performance of Indonesian language text documents (Perbandingan Feature Kata dan Frasa dalam kinerja Clustering dokumen teks berbahasa Indonesia)," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2007, no. SNATI, p. B-53-B-58.

[32] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," *Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014*, pp. 70–73, 2014, doi: 10.1109/IALP.2014.6973521.

[33] T. Tanabe, M. Takahashi, and K. Shudo, "A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing," *Comput. Speech Lang.*, vol. 28, no. 6, pp. 1317–1339, 2014, doi: 10.1016/j.csl.2013.09.001.

[34] S. Pal, S. K. Naskar, P. Pecina, S. Bandyopadhyay, and A. Way, "Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation," *Proc. Work. Multiword Expressions from Theory to Appl.*, no. August, pp. 45–53, 2010.

[35] P. Thanawala and J. Pareek, "MwTExt: automatic extraction of multi-word terms to generate compound concepts within ontology," *Int. J. Inf. Technol.*, vol. 10, no. 3, pp. 303–311, 2018, doi: 10.1007/s41870-018-0111-6.

[36] J. Duan, M. Zhang, W. Jingzhong, and Y. Xu, "A hybrid framework to extract bilingual multiword expression from free text," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 314–320, 2011, doi:

10.1016/j.eswa.2010.06.067.

[37] W. Zhang, T. Yoshida, X. Tang, and T. B. Ho, “Improving effectiveness of mutual information for substantival multiword expression extraction,” *Expert Syst. Appl.*, vol. 36, no. 8, pp. 10919–10930, 2009, doi: 10.1016/j.eswa.2009.02.026.

[38] V. Pajić, S. Vujičić Stanković, R. Stanković, and M. Pajić, “Semi-automatic extraction of multiword terms from domain-specific corpora,” *Electron. Libr.*, vol. 36, no. 3, pp. 550–567, 2018, doi: 10.1108/EL-06-2017-0128.

[39] B. Vincent, “Investigating academic phraseology through combinations of very frequent words: A methodological exploration,” *J. English Acad. Purp.*, vol. 12, no. 1, pp. 44–56, 2013, doi: 10.1016/j.jeap.2012.11.007.

[40] E. M. da Silva and R. R. Souza, “Information Retrieval System Using Multiwords Expressions (Mwe) As Descriptors,” *J. Inf. Syst. Technol. Manag.*, vol. 9, no. 2, pp. 213–234, 2012, doi: 10.4301/s1807-17752012000200002.

[41] P. Sanjanaashree, M. Anand Kumar, and K. P. Soman, “Dependency based multiword expression extraction towards NLP applications,” *ACM Int. Conf. Proceeding Ser.*, vol. 10-11-Octo, 2014, doi: 10.1145/2660859.2660928.

[42] Y. Tsvetkov and ShulyWintner, “Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources,” *Comput. Linguist.*, 2013, doi: 10.1162/COLI a 00177.

[43] N. J. Perdana, “Implementasi Algoritma Google Latent Semantic Distance untuk Ekstraksi Rangkaian Kata Kunci Artikel Jurnal Ilmiah,” *Comput. J. Comput. Sci. Inf. Syst.*, vol. 2, pp. 186–195, 2018.

[44] K. Lim and J. Park, “Real-world sentence boundary detection using multitask learning: A case study on French,” *Nat. Lang. Eng.*, pp. 1–21, 2022, doi: 10.1017/S1351324922000134.

[45] S. Raharjo, R. Wardoyo, and A. E. Putra, “Rule Based Sentence Segmentation of Indonesia Language,” *J. Eng. Appl. Sci.*, vol. 13, pp. 8986–8992, 2018.

[46] C. N. Purwanto, A. T. Hermawan, J. Santoso, and Gunawan, “Distributed Training for Multilingual Combined Tokenizer using Deep Learning Model and Simple Communication Protocol,” *2019 1st Int. Conf. Cybern. Intell. Syst. ICORIS 2019*, vol. 1, no. August, pp. 110–113, 2019, doi: 10.1109/ICORIS.2019.8874898.

[47] J. Santoso, E. I. Setiawan, C. N. Purwanto, and F. Kurniawan, “Indonesian Sentence Boundary Detection using Deep Learning Approaches,” *Knowl. Eng. Data Sci.*, vol. 4, no. 1, p. 38, 2021, doi: 10.17977/um018v4i12021p38-48.

[48] C. E. González-Gallardo, E. L. Pontes, F. Sadat, and J. M. Torres-Moreno, “Automated Sentence Boundary Detection in Modern Standard Arabic Transcripts using Deep Neural Networks,”

Procedia Comput. Sci., vol. 142, pp. 339–346, 2018, doi: 10.1016/j.procs.2018.10.485.

[49] A. Michaud, O. Adams, T. A. Cohn, G. Neubig, and S. Guillaume, "Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit," *Lang. Doc. Conserv.*, vol. 12, pp. 393–429, 2018.

[50] J. Savelka, V. R. Walker, M. Grabmair, and K. D. Ashley, "Sentence boundary detection in adjudicatory decisions in the United States," *TAL Trait. Autom. des Langues*, vol. 58, no. 2, pp. 21–45, 2017.

[51] M. S. U. Miah *et al.*, "Sentence Boundary Extraction from Scientific Literature of Electric Double Layer Capacitor Domain: Tools and Techniques," *Appl. Sci.*, vol. 12, no. 3, pp. 1–19, 2022, doi: 10.3390/app12031352.

[52] D. F. Wong, L. S. Chao, and X. Zeng, "I sentenizer-: Multilingual sentence boundary detection model," *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/196574.

[53] C. G. Chithra and E. Ramaraj, "Heuristic Sentence Boundary Detection and Classification," *Int. J. Emerg. Technol.*, vol. 7, no. 2, pp. 199–206, 2016.

[54] N. Zampieri, I. Illina, and D. Fohr, "Multiword Expression Features for Automatic Hate Speech Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12801 LNCS, pp. 156–164, 2021, doi: 10.1007/978-3-030-80599-9_14.

[55] D. Premasiri and T. Ranasinghe, "BERT (s) to Detect Multiword Expressions BERT (s) to Detect Multiword Expressions," *Eur. 2022 Int. Conf. 'Computational Corpus-based Phraseol.*, no. August, 2022, doi: 10.48550/arXiv.2208.07832.

[56] F. Bu, X. Y. Zhu, and M. Li, "A new multiword expression metric and its applications," *J. Comput. Sci. Technol.*, vol. 26, no. 1, pp. 3–13, 2011, doi: 10.1007/s11390-011-9410-0.

[57] E. Wehrli, V. Seretan, and L. Nerima, "Sentence Analysis and Collocation Identification," *Proc. Work. Multiword Expr. from Theory to Appl. (MWE 2010)*, pp. 27–35, 2010.

[58] J. N. Arwidarasti, I. Alfina, and A. A. Krisnadhi, "Adjusting Indonesian Multiword Expression Annotation to the Penn Treebank Format," *2020 Int. Conf. Asian Lang. Process. IALP 2020*, pp. 75–80, 2020, doi: 10.1109/IALP51396.2020.9310479.

[59] N. Loukachevitch and E. Parkhomenko, *Recognition of multiword expressions using word embeddings*, vol. 934. Springer International Publishing, 2018.

[60] K. Steyer and A. Brunner, "Contexts, Patterns, Interrelations - New Ways of Presenting Multi-word Expressions," no. April, pp. 82–88, 2015, doi: 10.3115/v1/w14-0814.

[61] X. Peng, Z. Yi, X. Y. Wei, D. Z. Peng, and Y. S. Sang, "Free-gram phrase identification for modeling Chinese text," *Inf. Process. Lett.*, vol. 113, no. 4, pp. 137–144, 2013, doi: 10.1016/j.ipl.2012.11.005.

[62] A. Hautli and S. Sulger, "Extracting and classifying Urdu multiword expressions," *ACL HLT 2011*

- *49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Student Sess.*, no. June, pp. 24–29, 2011.

[63] M. A. Finlayson and N. Kulkarni, "Detecting Multi-Word Expressions improves Word Sense Disambiguation," *Proc. 8th Work. Multiword Expressions from Parsing Gener. to Real World (MWE 2011)*, no. June, pp. 20–24, 2011, [Online]. Available: http://www.aclweb.org/anthology/W11-0805.

[64] J. Legrand and R. Collobert, "Phrase Representations for Multiword Expressions," *Proc. 12th Work. Multiword Expressions*, no. 2011, pp. 67–71, 2016, doi: 10.18653/v1/w16-1810.

[65] M. Piasecki and K. Kanclerz, "Non-Contextual vs Contextual Word Embeddings in Multiword Expressions Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13501 LNAI, pp. 193–206, 2022, doi: 10.1007/978-3-031-16014-1_16.

[66] K. Sriraghav, S. Jayanthi, N. Vidya, and V. S. F. Enigo, "ScrAnViz-A tool to scrap, analyze and visualize unstructured-data using attribute-based opinion mining algorithm," *2017 Innov. Power Adv. Comput. Technol. i-PACT 2017*, vol. 2017-Janua, pp. 1–5, 2017, doi: 10.1109/IPACT.2017.8244916.

[67] Accenture.com, "Unstructured data is key to insights," *https://www.accenture.com/*, 2020. https://www.accenture.com/us-en/services/applied-intelligence/unstructured-data-solutions?gclid=Cj0KCQjwocPnBRDFARIsAJJcf94LtjxbmNiUVZHGrPikVe3nOPeqSfk-6j3fCTdDEmpNOQTQB_l4ywMaAjffEALw_wcB (accessed May 14, 2020).

[68] V. Gupta and A. Gosain, "A comprehensive review of unstructured data management approaches in data warehouse," *Proc. - 2013 Int. Symp. Comput. Bus. Intell. ISCBI 2013*, pp. 64–67, 2013, doi: 10.1109/ISCBI.2013.20.

[69] V. A. Ingle, "Processing of unstructured data for information extraction," *3rd Nirma Univ. Int. Conf. Eng. NUiCONE 2012*, pp. 6–8, 2012, doi: 10.1109/NUICONE.2012.6493202.

[70] R. A. Kambau and Z. A. Hasibuan, "Unified concept-based multimedia information retrieval technique," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2017-Decem, no. May, 2017, doi: 10.1109/EECSI.2017.8239086.

[71] B. Marr, "What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone," *forbes.com*, 2019. https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/?sh=4d68f81815f6 (accessed Mar. 20, 2020).

[72] M. F. Abdullah and K. Ahmad, "The mapping process of unstructured data to structured data," *Int. Conf. Res. Innov. Inf. Syst. ICRIIS*, vol. 2013, pp. 151–155, 2013, doi:

10.1109/ICRIIS.2013.6716700.

[73] N. Veeranjaneyulu, M. N. Bhat, and A. Raghunath, “Approaches for Managing and Analyzing Unstructured Data,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 01, pp. 19–24, 2014.

[74] A. Kulkarni and A. Shivananda, *Natural Language Processing Recipes*, Unlocking. Bangalore, India: APress, 2019.

[75] R. K. Lomotey and R. Deters, “Topics and terms mining in unstructured data stores,” *Proc. - 16th IEEE Int. Conf. Comput. Sci. Eng. CSE 2013*, pp. 854–861, 2013, doi: 10.1109/CSE.2013.129.

[76] T. King, “80 Percent of Your Data Will Be Unstructured in Five Years,” *https://solutionsreview.com/*, 2019. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/ (accessed Mar. 07, 2020).

[77] A. C. Eberendu, “Unstructured Data: an overview of the data of Big Data,” *Int. J. Comput. Trends Technol.*, vol. 38, no. 1, pp. 46–50, 2016, doi: 10.14445/22312803/ijctt-v38p109.

[78] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[79] S. Geetha and G. S. Anandha Mala, “Effectual extraction of data relations from unstructured data,” *IET Conf. Publ.*, vol. 2012, no. 624 CP, pp. 58–61, 2012, doi: 10.1049/cp.2012.2190.

[80] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, “Text Mining: Techniques, Applications and Issues,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 414–418, 2016, doi: 10.14569/ijacsa.2016.071153.

[81] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, “The application of text mining methods in innovation research: current state, evolution patterns, and development priorities,” 2020.

[82] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” *arXiv*, no. July, 2017.

[83] M. Anandarajan, C. Hill, and T. Nolan, *Practical Text Analytics - Maximizing the Value of Text Data*, vol. 2. Switzerland: Springer, 2019.

[84] D. Haryalesmana, “ID-Stopwords,” 2016. https://github.com/masdevid/ID-Stopwords (accessed Jun. 17, 2020).

[85] F. Rahutomo and A. R. T. H. Ririd, “Evaluasi Daftar Stopword Bahasa Indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 1, p. 41, 2019, doi: 10.25126/jtiik.2019611226.

[86] Linguamatics, “What is Text Mining, Text Analytics and Natural Language Processing?,” *Linguamatics*. https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing (accessed Feb. 20, 2021).

[87] A. Srivastav and S. Prajapat, “Text similarity algorithms to determine Indian penal code sections for offence report,” *IAES Int. J. Artif. Intell.*, vol. 11, no. 1, p. 34, 2022, doi: 10.11591/ijai.v11.i1.pp34-40.

[88] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning*. USA: O'Reilly Media, 2012.

[89] Y. S. Kalambe, D. Pratiba, and P. Shah, "Big Data Mining Tools for Unstructured Data: a Review," *Ijitr*, vol. 3, no. 2, pp. 2012–2017, 2015, [Online]. Available: http://www.ijitr.com/index.php/ojs/article/view/610.

[90] K. F. Wong, W. Li, R. Xu, and Z. S. Zhang, *Introduction to Chinese natural language processing*, vol. 2, no. 1. 2010.

[91] Kementrian Pendidikan dan Kebudayaan, *Kalimat*, Seri Penyu. 2015.

[92] P. B. D. P. N. Pusat, *Kamus Bahasa Indonesia*. Jakarta, 2008.

[93] C. Ramisch, *Multiword Expressions Acquisition*, A Generic. France: Springer, 2015.

[94] M. Rodrigues and A. Teixeira, *Advanced Applications of Natural Language Processing for Performing Information Extraction*. Springer, 2015.

[95] Farhanaaz and V. Sanju, "An exploration on lexical analysis," *Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016*, pp. 253–258, 2016, doi: 10.1109/ICEEOT.2016.7755127.

[96] A. Zheng and A. Casari, *Feature engineering for machine learning, Principles and Techniques for Data Scientist*, no. September. Portugal: Springer, 2018.

[97] T. P. P. B. I. Indonesia, *Pedoman Umum Ejaan Bahasa Indonesia*, Fourth. Jakarta: Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan, 2016.

[98] A. Singh, B. P. Singh, A. K. Poddar, and A. Singh, "Sentence boundary detection for Hindi–English social media text," *Adv. Intell. Syst. Comput.*, vol. 709, pp. 207–215, 2018, doi: 10.1007/978-981-10-8633-5_22.

[99] N. Wanjari, G. M. Dhopavkar, and N. B. Zungre, "Sentence Boundary Detection for Marathi Language," *Phys. Procedia*, vol. 78, no. December 2015, pp. 550–555, 2016, doi: 10.1016/j.procs.2016.02.101.

[100] K. Sirts and K. Peekman, "Evaluating sentence segmentation and word tokenization systems on estonian web texts," *Front. Artif. Intell. Appl.*, vol. 328, pp. 174–181, 2020, doi: 10.3233/faia200620.

[101] D. Tuggener and A. Aghaebrahimian, "The sentence end and punctuation prediction in NLG text (SEPP-NLG) shared task 2021," *CEUR Workshop Proc.*, vol. 2957, no. 2016, 2021.

[102] D. Griffis, C. Shivade, E. Fosler-lussier, and A. M. Lai, "A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain Department of Computer Science and Engineering , Department of Biomedical Informatics , The Ohio State University , Columbus , OH . National Institutes of Hea," *AMIA Summits Transl. Sci. Proc. 2016*, pp. 88–97, 2016.

[103] S. ElBasha, A. Elhawil, and N. Drawil, "Multilingual Sentiment Analysis to Support Business

Decision-making via Machine learning models," *Third.Leabz.Org.Ly*, 2021, [Online]. Available: https://third.leabz.org.ly/wp-content/uploads/2022/05/Multilingual-Sentiment-Analysis-to-Support-Business-Decision.pdf.

[104] C. Ramisch, "A generic and open framework for multiword expressions treatment: from acquisition to applications," *Nouv. thse*, no. September, pp. 61–66, 2012, [Online]. Available: https://www.lume.ufrgs.br/bitstream/handle/10183/65777/000870122.pdf?sequence=1%5Cnhttp://dl.acm.org/citation.cfm?id=2390342%5Cnpapers2://publication/uuid/DA5C2EA2-EC30-466E-88E2-CA052238D681%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitl.

[105] M. A. Khak, "Idiom Dalam Bahasa Indonesia: Struktur Dan Maknai," *Widyaparwa*, vol. 39, no. 2, pp. 141–154, 2011, [Online]. Available: http://widyaparwa.kemdikbud.go.id/index.php/widyaparwa/article/view/36.

[106] H. Kridalaksana, *Kamus Linguistik*. Jakarta: PT Gramedia, 1982.

[107] S. S. T. W. Sasangka, *Kalimat*, Seri Penyu. Jakarta: Pusat Pembinaan dan Pemasyarakatan Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan, 2014.

[108] P. Hanks, *Lexical Analysis - Norms and Exploitations*. London, England: The MIT Press, 2013.

[109] M. Schonlau, N. Guenther, and I. Sucholutsky, "Text mining with n-gram variables," *Stata J.*, vol. 17, no. 4, pp. 866–881, 2017, doi: 10.1177/1536867X1801700406.

[110] C. C. Aggarwal and C. Zhai, *Mining Text Data*. New York, USA: Springer New York, 2012.

[111] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.

[112] L. Deng and Y. Liu, *Deep learning in natural language processing*. 2018.

[113] S. Senhadji and R. A. S. Ahmed, "Fake news detection using naïve Bayes and long short term memory algorithms," *IAES Int. J. Artif. Intell.*, vol. 11, no. 2, pp. 746–752, 2022, doi: 10.11591/ijai.v11.i2.pp746-752.

[114] H. Liu, A. Gegov, and M. Cocea, *Rule Based Systems for Big Data*, vol. 13. 2016.

[115] A. M. Aubaid and A. Mishra, "A rule-based approach to embedding techniques for text document classification," *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10114009.

[116] N. Sadvilkar and M. Neumann, "PySBD: Pragmatic Sentence Boundary Disambiguation," in *Proceeding of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020, pp. 110–114, doi: 10.18653/v1/2020.nlposs-1.15.

[117] F. Tesselaar, "Sentence Boundary Detection (SBD).," 2021. https://knod.github.io/sbd/.

[118] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[119] R. Karsi, M. Zaim, and J. El Alami, “Assessing naive bayes and support vector machine performance in sentiment classification on a big data platform,” *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, pp. 990–996, 2021, doi: 10.11591/IJAI.V10.I4.PP990-996.

[120] A. M. Aubaid and A. Mishra, “Text classification using word embedding in Rule-based methodologies: A systematic mapping,” *TEM J.*, vol. 7, no. 4, pp. 902–914, 2018, doi: 10.18421/TEM74-31.

[121] A. Raut and R. K. Pandey, “Sentiment analysis using optimized feature sets in different twitter dataset domains,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11, pp. 3035–3039, 2019, doi: 10.35940/ijitee.K2195.0981119.

[122] P. N. Bali, “Perancangan Penganalisis Struktur Kalimat Bahasa Indonesia Dengan Menggunakan Constraint-Based Formalism,” *Lontar Komput.*, vol. 5, no. 2, 2015.

[123] D. Jia-li and Y. Ping-fang, “A computational linguistic approach to natural language processing with applications to garden path sentences analysis,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 9, 2012, doi: 10.14569/ijacsa.2012.030909.

[124] D. Anggraini, A. B. Mutiara, T. M. Kusuma, and L. Wulandari, “Algorithm for simple sentence identification in Bahasa Indonesia,” *Proc. 3rd Int. Conf. Informatics Comput. ICIC 2018*, no. 100, pp. 1–6, 2018, doi: 10.1109/IAC.2018.8780552.

[125] H. Ye *et al.*, “Contrastive Triple Extraction with Generative Transformer,” *35th AAAI Conf. Artif. Intell. AAAI 2021*, vol. 16, pp. 14257–14265, 2021, doi: 10.1609/aaai.v35i16.17677.

[126] E. Yulianti, A. Kurnia, M. Adriani, and Y. S. Duto, “Normalisation of Indonesian-English Code-Mixed Text and its Effect on Emotion Classification,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 674–685, 2021, doi: 10.14569/IJACSA.2021.0121177.

[127] G. Venugopal-Wairagade, J. R. Saini, and D. Pramod, “Novel language resources for hindi: An aesthetics text corpus and a comprehensive stop lemma list,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 233–239, 2020, doi: 10.14569/ijacsa.2020.0110130.

[128] J. K. Raulji, J. R. Saini, K. Pal, and K. Kotecha, “A Novel Framework for Sanskrit-Gujarati Symbolic Machine Translation System,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 374–380, 2022, doi: 10.14569/IJACSA.2022.0130444.

[129] R. Dollah, C. Y. Sheng, N. Zakaria, M. S. Othman, and A. W. Rasib, “Deep learning classification of biomedical text using convolutional neural network,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 512–517, 2019, doi: 10.14569/ijacsa.2019.0100867.

[130] J. Camacho-Collados and M. T. Pilehvar, “On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis,” *EMNLP 2018 - 2018 EMNLP Work. BlackboxNLP Anal. Interpret. Neural Networks NLP, Proc. 1st Work.*, pp.

40–46, 2018, doi: 10.18653/v1/w18-5406.

[131] H. A. Putranto, O. Setyawati, and W. Wijono, “Effect of Phrase Detection with POS-Tagger on Sentiment Classification Accuracy using SVM,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 5, no. 4, pp. 252–259, 2016, doi: 10.22146/jnteti.v5i4.271.