Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
JLK
Vol 6, No 1
Maret 2023

Ekstraksi Kata Kunci pada Bahasa Indonesia Menggunakan Metode YAKE

Novi Yusliani[*1], Gerald Plakasa[*2], Abdiansah[*3], Mastura Diana Marieska[*4], Danny Matthew Saputra[*5]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya
Jln. Raya Palembang – Prabumulih Km.32 Indralaya, Ogan Ilir, Sumatera Selatan

[1]novi_yusliani@unsri.ac.id
[2]geraldplakasa12@gmail.com
[3]abdiansah@unsri.ac.id
[4]mastura.diana@ilkom.unsri.ac.id
[5]danny.saputra@gmail.com

***Abstrak*— Peneliti, Mahasiswa, dan Juga Dosen biasanya melakukan penelitian untuk menghasilkan publikasi hasil penelitiannya. Saat ini pertumbuhan publikasi ilmiah terus meningkat. ketika publikasi akan di berikan ke reviewer maka publikasi yang kirimkan harus sesuai dengan bidang yang diampu oleh *reviewer* tersebut. Salah satu cara untuk mengetahui inti dari sebuah publikasi ilmiah yaitu dengan melakukan ekstraksi kata kuncinya. Metode yang digunakan untuk ekstraksi kata kunci salah satunya yaitu YAKE (*Yet Another Keyword Extraction*). Penelitian ini menggunakan *dataset* 100 publikasi ilmiah dari *website* jtiik, jatisi, dan jepin dengan topik Ilmu Komputer. Berdasarkan penelitian yang telah dilakukan, konfigurasi pada parameter *Levenshtein Distance* memiliki pengaruh terhadap hasil kata kuncinya. Evaluasi dari penelitian ini menghasilkan nilai *f-measure* sebesar 54,1% dan nilai akurasi sebesar 97,05% dengan parameter *Levenshtein Distance* < 2.**

***Kata kunci*— Ekstraksi Kata Kunci, Yake, Levenshtein Distance**

I. PENDAHULUAN

Saat ini publikasi hasil penelitian terus berkembang di Indonesia agar ter-indeks secara nasional dan Internasional [1]. Akibatnya, banyak publikasi hasil penelitian berbahasa Indonesia yang akan diberikan ke *reviewer*. Ketika publikasi hasil penelitian diberikan ke *reviewer*, bidang ilmu yang dimiliki *reviewer* dan publikasinya harus sesuai [1].

Kata kunci dari sebuah publikasi hasil penelitian dapat menjadi salah satu rujukan bidang ilmu untuk *reviewer*. Untuk mendapatkan kata kunci yang sesuai dengan publikasi yang dibuat dapat dilakukan proses ekstraksi kata kunci atau dapat di definiskan sebuah proses otomatis untuk menggambil term dari sebuah teks [9].

Algoritma yang digunakan untuk melakukan ekstraksi akta kunci terdapat beberapa pendekatan, yaitu *Statistic*, *Graph*, *Machine Learning*, dan lainnya. *TextRank* merupakan algoritma yang melakukan ekstraksi kata kunci yang pengaplikasiannya berdasarkan pendekatan *Graph* [5]. Terdapat Algoritma lain seperti TF-IDF. *SingleRank*, RAKE, dan YAKE. Berdasarkan dari hasil penelitian yang telah dilakukan terbukti bahwa YAKE mendapatkan hasil yang lebih baik [3]. Berdasarkan hal tersebut penelitian ini akan melakukan Ekstraksi Kata kunci Pada Bahasa Indonesia Menggunakan Metode YAKE.

II. LANDASAN TEORI

A. *Keyphrase Extraction*

Keyphrase Extraction atau dalam bahasa Indonesia Ekstraksi Kata Kunci adalah proses pengambilan kata kunci yang digunakan untuk mendapatkan teks unit terkecil yang mempresentasikan makna atau informasi maksud dari teks yang dilakukan ekstraksi tersebut [5]. Ekstraksi kata kunci adalah proses yang sangat penting untuk mengambil informasi utama dari sebuah teks dengan berdasarkan pendekatan *Machine Learning* atau *Statistic* [4].

B. *YAKE*

YAKE (*Yet Another Keyword Extraction*) merupakan metode Unsupervised Term Extraction yang mana tidak bergantung pada bahasanya [10]. Beberapa komponen utama dari metode didalam YAKE adalah *Text Pre-Processing*, *Feature Extraction*, *Individual Term Weighting*, dan *Candidate Keywords Generation* [3]. Pertama yang dipertimbangkan ketika melakukan pemrosesan teks adalah penghapusan karakter spesial dan melakukan *lowercase* untuk keseluruhan teks, agar teks yang digunakan menjadi lebih bersih dan juga dilakukan proses tokenisasi. Selanjutnya yang dilakukan adalah *Feature Extraction*, pada bagian ini akan mengambil beberapa fitur penting untuk melihat karakteristik dari setiap tokennya, berikut ini fitur yang dipertimbangkan pada YAKE:

Submitted 30-11-2022; Revised 23-03-2023; Accepted 25-03-2023

1) *Casing (W_{case}):* Pada fitur ini bagian penting yang dipertimbangkan yaitu huruf kapital dan akronim untuk mendapatkan bobotnya didapatkan dengan menggunakan persamaan (1).

$$W_{Case} = \frac{\max(TF(U(w)),\ TF(A(w)))}{log_2(TF(w))} \quad (1)$$

Keterangan:

TF(U(w)) : berapa kali kata w yang awalannya kapital
TF(A(w)) : berapa kali kata w yang akronim
TF(w) : berapa kali kata w muncul

2) *Word Position ($W_{Position}$):* Fitur ini akan mempertimbangkan kata w muncul di kaliamt mana. Bobotnya didapatkan dengan persamaan berikut.

$$W_{Position} = log_2(log_2(2 + Median(Sen_w))) \quad (2)$$

Keterangan:

Sen_w : posisi kemunculan kata w di beberapa kalimat

3) *Word Frequency (W_{Freq}):* Pada fitur ini yang akan dipertimbangkan adalah tingkat kemunculan sebuah kata dengan bobotnya didapatkan pada persamaan berikut.

$$W_{Freq} = \frac{TF(w)}{MeanTF + 1 * \sigma} \quad (3)$$

Keterangan:

σ : Simpangan baku
TF(w) : berapa kali kata w muncul

4) *Word Relatedness to Context (W_{Rel}):* Fitur ini akan mempertimbangkan kata yang memiliki karakteristik yang mirip dengan kata yang biasa masuk ke dalam *stopword*, untuk mendapatkan bobotnya dapat menggunakan persamaan berikut.

$$W_{Rel} = \left(0.5 + \left(\left(WL * \frac{TF(w)}{MaxTF}\right) + PL\right)\right) + \left(0.5 + \left(\left(WR * \frac{TF(w)}{MaxTF}\right) + PR\right)\right) \quad (4)$$

Keterangan:

WL : rasio antara frekuensi kata berbeda disisi kiri dan frekuensi kata yang muncul di sisi kiri
WR : rasio antara frekuensi kata berbeda disisi Kanan dan frekuensi kata yang muncul di sisi Kanan
PL : rasio antara frekuensi kata berbeda disisi kiri dan MaxTF
PR : rasio antara frekuensi kata berbeda disisi Kanan dan MaxTF

5) *Word DifSentence ($W_{DifSentence}$):* Pada Fitur berikut ini hal yang dipertimbangkan adalah menghitung kata yang muncul di banyak kalimat, bobotnya didapatkan dari persamaan berikut.

$$W_{DifSentence} = \frac{SF(w)}{\#Sentences} \quad (5)$$

Keterangan:

SF(W) : tingkat kalimat yang terdapat kata w

Selanjutnya, hal yang dilakukan adalah *Individual Term Weighting* yaitu melakukan perhitungan untuk semua bobot yang sudah di dapatkan pada masing-masing kata menggunakan persamaan berikut.

$$S(w) = \frac{W_{Rel} * W_{Position}}{W_{Case} + \frac{W_{Freq}}{W_{Rel}} + \frac{W_{DifSentence}}{W_{Rel}}} \quad (6)$$

Terakhir, yaitu bagian *Candidate Keyword List Generation* yaitu melakukan pengambilan calon kata kunci dengan melakukan sliding window lalu menghitung bobot final dengan persamaan berikut.

$$S(kw) = \frac{\prod_{w \in kw} S(w)}{TF(kw) * (1 + \sum_{w \in kw} S(w))} \quad (7)$$

setelah mendapatkan bobot masing-masing calon kata kunci, dilakukan pengurangan kata kunci menggunakan jarak *Levenshtein Distance*.

C. Levenshtein Distance

Levenshtein Distance merupakan ukuran lingustik komputasi untuk mengukur jarak dua *string* atau kata berdasarkan urutan karakternya [2]. Ukuran ini mengembalikan nilai berupa biaya minimum yang dibutuhkan untuk mengubah string awal ke dalam string target berdasarkan jarak. *Levenshtein Distance* dilakukan dengan menghitung jumlah minimum perubahan kata menjadi kata lain dengan melakukan penghapusan, penyisipan, dan pergantian [7]. [11] menjelaskan langkah-langkah yang digunakan dalam penelitian ini untuk mendapatkan nilai *levenshtein distance*.

D. Confusion Matrix

Confusion Matrix adalah pengukuran perbandingan dari hasil klasifikasi sebuah obyek dan nilai aslinya [6]. Perbandingan yang dilakukan berdasarkan inspirasi dari domain *Information Retrieval* yaitu *Precision*, *Recall*, dan *F-Score* [8].

1) *Precision:* rasio dari obyek yang dianggap benar dan jumlah total obyek dari sistem. Persamaan (8) menunjukkan bahwa untuk mendapatkan nilai *precision* dibutuhkan nilai TP (*true positive*) dan nilai FP (*false positive*).

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

2) *Recall:* Recall adalah rasio dari obyek yang dianggap benar dan jumlah obyek yang ada pada *dataset*. Nilai *recall* berdasarkan persamaan (9) diperoleh dari nilai TP dibagi dengan hasil penjumlahan TP dan FN (*false negative*).

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

3) *F-Score: F-Score* adalah pengukuran yang mempertimbangkan antara nilai *Precision* dan *Recall*, yang ditunjukkan pada persamaan (10).

$$F - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (10)$$

III. METODE PENELITIAN

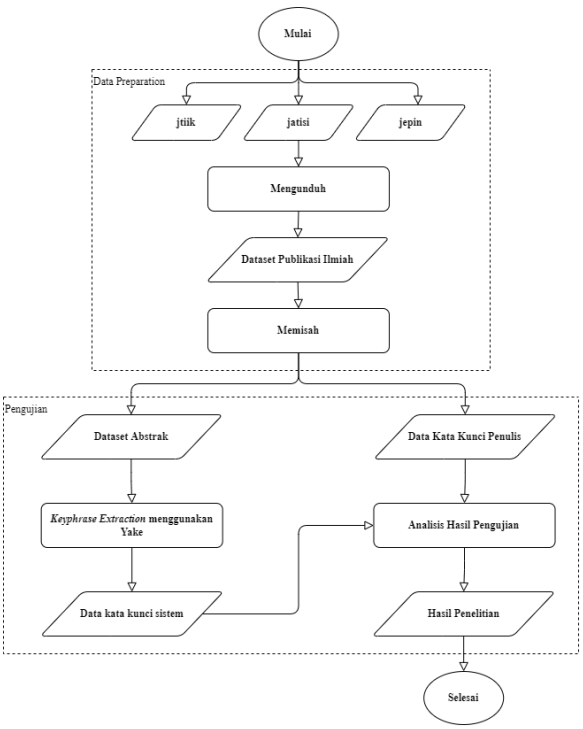
A. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah 100 data Publikasi Ilmiah dari website website jtiik, jatisi, jepin yang mana hanya terbatas pada Ilmu Komputer topiknya.

B. Struktur Dataset

Dataset yang digunakan untuk penelitian ini adalah data abstrak dan Judul dari masing-masing 100 data jurnal. Di ambil juga kata kunci dari penulis untuk nanti digunakan pada proses evaluasi atau pengujian.

C. Kerangka Kerja



Gambar 1. Kerangka Kerja

Penelitian ini akan melewati setiap tahapan proses pada kerangka kerja yang ditampilkan pada Gambar 1. Berdasarkan kerangka kerja pada Gambar 1, tahapan dalam penelitian ini adalah

1) *Data Preparation:* Berdasarkan dari tiga *website* sumber data, di unduh secara acak 100 publikasi Ilmiah Berbahasa Indonesia dan masing-masing abstrak, judul, dan kata kunci penulis di simpan kedalam .txt.

2) *Pengujian:* Pada Judul dan Abstrak yang dikumpulkan dilakukan proses ekstraksi kata kunci menggunakan YAKE dan kata kunci yang di hasilkan akan di analisis dengan kata kunci dari penulis.

D. Metode Pengembangan Perangkat Lunak

Penelitian ini menggunakan metode Pengembangan Perangkat Lunak RUP (Rational Unified Process). Empat fase utama pada RUP yaitu,

1) *Inseption:* Merupakan tahapan pertama yang didalamnya yaitu proses mengumpulkan data, menentukan ruang lingkup analisis, dan mendesain perangkat lunak.

2) *Elaboration:* Merupakan tahapan kedua yang mana prosesnya yaitu meranvang perangkat lunak dengan menentukan spesifikasi fitur, melakukan analisis dan desain Perangkat Lunak.

3) *Construction:* Merupakan tahapan ketiga dimana membuat aplikasi dan melakukan pengujian untuk aplikasinya.

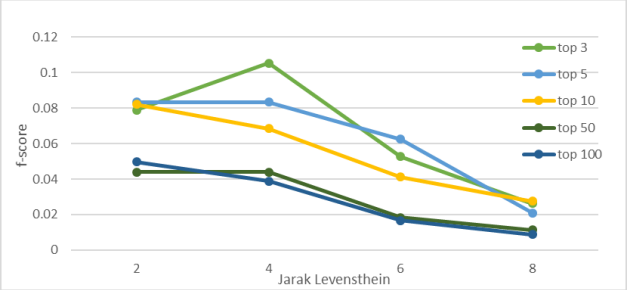
4) *Transition:* Merupakan tahapan keempat, proses didalamnya yaitu melakukan perbaikan dan penambahan data lalu membuat kesimpulan dari perangkat lunak yang telah dikembangkan.

IV. HASIL DAN PEMBAHASAN

Pengujian yang dilakukan pada penelitian ini akan menggunakan Similarity agar menampilkan hasil yang lebih jelas. Pengujian dilakukan beberapa tahapan sampai mendapatkan hasil yang baik.

A. Konfigurasi Parameter

Pengujian dilakukan pada 100 data Publikasi Ilmiah dengan mempertimbangkan beberapa jarak Levenshtein Distance seperti <2. <4. <6. Dan <8. Gambar 2 merupakan hasil visualisasi pengujian dengan masing-masing jarak Levenstehin Distance, masing-masing Top Kata Kunci (3,5,10,50,100) untuk sampel 10 dataset.

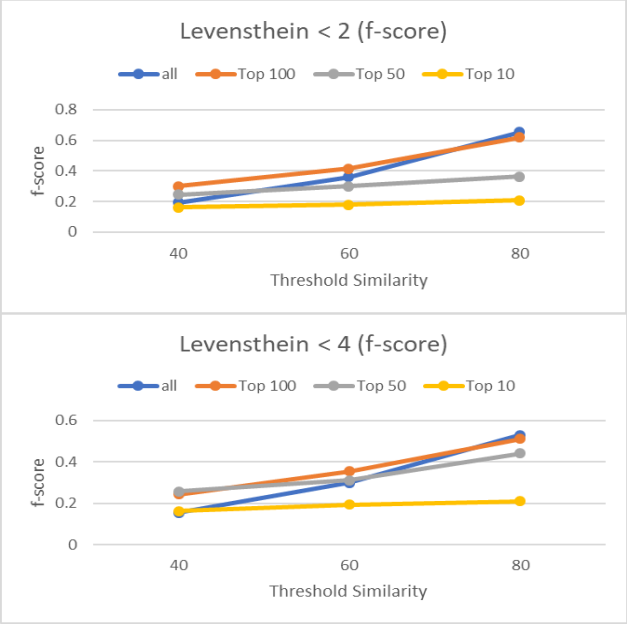


Gambar 2. Konfigurasi Parameter Levenshtein Distance

Dari hasil visualisasi Konfigurasi Parameter, terlihat nilai F-Score lebih baik pada Jarak Levenshtein <2 dan <4.

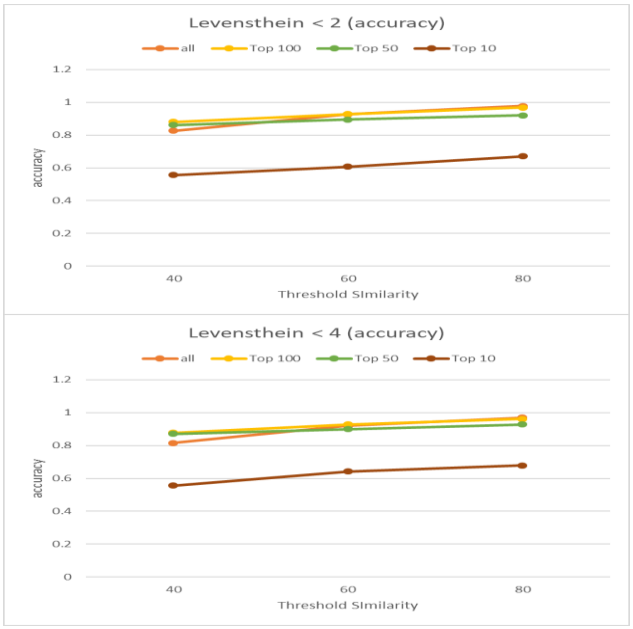
B. Hasil Pengujian Simlarity

Pengujian yang dilakukan pada penelitian ini akan menggunakan Cosine Similarity untuk melihat kemampuan kata kunci yang dihasilkan system. Gambar 3 merupakan hasil pengujian dengan jarak Levenshtein <2 dan < 4, lalu nilai Thereshold Similarity > 40, >60, dan > 80. Untuk kata kunci yang dihasilkan merupakan keseluruhan kata kunci (all), top 100, top 50, dan top 10. Menggunakan sampel 10 dataset.



Gambar 3. Nilai F-Score jarak Levenshtein <2 dan <4

Gambar 4 merupakan hasil perhitungan akurasi untuk ketentuan yang sama seperti sebelumnya.

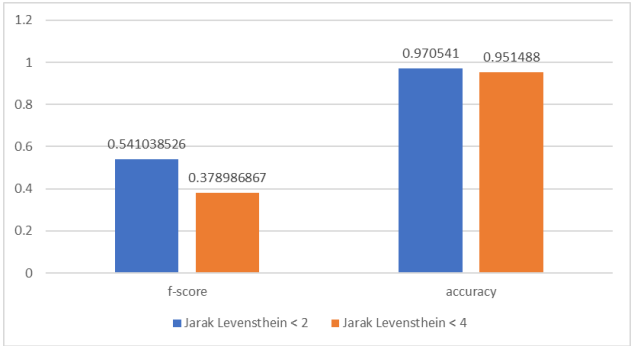


Gambar 4. Nilai Accuracy jarak Levenshtein < 2 dan < 4

Berdasarkan hasil visualisasi untuk nilai F-Score dan nilai akurasinya terlihat bahwa kata kunci yang dikeluarkan keseluruhan merupakan hasil yang lebih baik, disebabkan karena kata kunci yang menyerupai kata kunci penulis tidak selalu berada di peringkat atas. Terlihat juga pada visualiasinya Thereshold Similarity >80 juga membuat hasil lebih baik.

C. Hasil Pengujian Keseluruhan

Setelah mendapatkan konfigurasi yang baik, selanjutnya akan melakukan pengujian pada keseluruhan dataset, Gambar 5 merupakan perbandingan saat Jarak Levenshtein < 2 atau Jarak Levenshtein < 4.



Gambar 5. Perbandingan Jarak Levenshtein

Berdasarkan hasil perbandingan yang ditunjukkan pada Gambar 5 jarak *Levenshtein* <2 memiliki hasil yang lebih baik. Tabel 1 memperlihatkan contoh hasil keluaran kata kunci untuk satu *sample* di dalam *dataset*. Pada tabel 1, *generate keyphrase* adalah kumpulan semua kandidat *keyphrase* yang dihasilkan oleh sistem, *choosen keyphrase*

adalah kumpulan *keyphrase* yang dipilih oleh sistem berdasarkan kandidat *keyphrase* yang ada, dan *golden keyphrase* adalah *keyphrase* yang dihasilkan oleh penulis jurnal.

Tabel 1. Hasil Pengujian Kata Kunci satu Sampel

No.	*Generate Keyphrase*	*Choosen Keyphrase*	*Golden Keyphrase*
1	['bayes pada shooter', 'penyerangan pada non-player', 'satunya menggunakan metode', 'metode naïve bayes', 'npc menggunakan metode', 'npc', 'aksi', 'substansi', 'pembuatan', 'dikalahkan', 'khusus', 'pemain', 'dimiliki npc', 'mengatur', 'npc mampu', 'adaptif', 'menghasilkan', 'sesuai', 'bisanya', 'statis', 'behaviour', 'repetitif', 'menurunkan', 'bermain', 'diprediksi', 'tantangan', 'ai', 'mengatasi', 'teknik', 'learning', 'peneliti', 'diambil', 'dimiliki', 'amunisi', 'jarak', 'nyawa', 'parameter', 'pisau', 'tembak', 'dibagi', 'hasil', 'akurasi', 'otonom', 'dibanding', 'persentase', 'unggul', 'substansi penting', 'khusus agar npc', 'diambil oleh npc', 'menghasilkan behaviour', 'ai seperti', 'mengatasi masalah', 'banyak peneliti', 'teknik learning', 'akurasi 80', 'dengan akurasi', 'lebih unggul', 'bisanya menghasilkan behaviour', 'repetitif sehingga menurunkan', 'salah', 'naïve', 'tingkat', 'penerapan', 'variasi', 'kemenangan npc', 'satunya', 'keputusan', 'kondisi', 'granat', 'kemenangan', 'character', 'non-player', 'mudah', 'menurunkan tingkat', 'untuk mengatur', 'strategi', 'penelitian', 'dalam pembuatan', 'rulebase', 'shooter', 'adanya variasi', 'dilakukan penerapan', 'penyerangan', 'memberikan variasi', 'serangan', 'kemenangan npc dibanding', 'persentase kemenangan npc', 'sesuai kondisi npc', 'adaptif jika', 'tingkat tantangan bermain', 'menurunkan tingkat tantangan', 'otonom jika', 'penyerangan npc', 'mudah dikalahkan', 'persentase kemenangan', 'strategi khusus', 'tidak adaptif', 'menggunakan', 'akan mudah', 'mudah diprediksi', 'aksi penyerangan', 'metode', 'hasil penelitian', 'bayes', 'unggul dalam persentase', 'jumlah amunisi', 'statis dan', 'rulebase dapat', 'mengatur penyerangan npc', 'peneliti yang', 'mengatur penyerangan', 'parameter yang', 'serangan pisau', 'serangan tembak', 'game', 'penyerangan dibagi', 'diprediksi dan repetitif', 'menggunakan teknik', 'rulebase bisanya menghasilkan', 'serangan terhadap pemain', 'behaviour yang statis', 'bayes membuat npc', 'amunisi yang dimiliki', 'npc dibanding metode', 'dibagi menjadi serangan', 'salah satunya', 'penyerangan secara otonom', 'bayes juga', 'pembuatan game', 'tingkat kemenangan npc', 'terdapat kondisi', 'untuk keputusan', 'salah satu substansi', 'kondisi baru', 'tantangan bermain game', 'non-player character', 'npc tidak mudah', 'satunya adalah', 'melakukan penyerangan', 'variasi serangan', 'penelitian ini', 'keputusan penyerangan', 'naïve bayes', 'strategi penyerangan npc', 'metode naïve', 'npc adalah rulebase', 'penelitian menunjukkan penerapan', 'strategi penyerangan', 'aksi penyerangan non-player',	['aksi penyerangan', 'npc', 'metode naïve bayes', 'naïve bayes', 'rulebase']	['Aksi Penyerangan', 'NPC', 'Naïve Bayes', 'Rulebase']

	'keputusan serangan', 'penerapan metode', 'penyerangan diperlukan strategi', 'variasi serangan sesuai', 'bayes digunakan', 'keputusan penyerangan dibagi', 'serangan sesuai kondisi', 'penyerangan adalah nyawa', 'metode rulebase', 'rulebase sebesar 16', 'metode yang', 'shooter game', 'npc pada shooter', 'serangan yang diambil', 'bayes sebagai strategi', 'bayes sebesar 60', 'metode rulebase bisanya', 'dibanding metode rulebase', 'penyerangan non-player character', 'penerapan metode naïve', 'salah satu metode']		

Tabel 2 merupakan *Confusion Matrix* untuk keseluruhan *sample* yang ada di dalam *dataset*.

Tabel 2. Confution Matrix Hasil Akhir

No.	*Confusion Matrix*				*Precision*	*Recall*	*F-Score*	*Accuracy*
	TP	*TN*	*FP*	*FN*				
1	4	165	1	0	0.8	1	0.889	0.994
2	5	225	3	0	0.625	1	0.769	0.987
3	5	169	3	0	0.625	1	0.769	0.983
4	2	168	8	2	0.2	0.5	0.286	0.944
5	4	144	0	1	1	0.8	0.889	0.993
6	2	262	2	2	0.5	0.5	0.5	0.985
7	1	171	4	2	0.2	0.333	0.25	0.966
8	2	100	0	3	1	0.4	0.571	0.971
9	5	141	6	1	0.455	0.833	0.588	0.954
10	5	135	0	0	1	1	1	1
11	3	167	1	3	0.75	0.5	0.6	0.977
12	4	103	11	2	0.267	0.667	0.381	0.892
13	3	175	6	0	0.333	1	0.5	0.967
14	1	219	7	4	0.125	0.2	0.154	0.952
15	1	152	11	2	0.083	0.333	0.133	0.922
16	3	175	5	1	0.375	0.75	0.5	0.967
17	4	194	2	1	0.667	0.8	0.727	0.985
18	5	183	1	0	0.833	1	0.909	0.995
19	5	239	3	0	0.625	1	0.769	0.988
20	2	139	1	3	0.667	0.4	0.5	0.972
21	3	179	1	2	0.75	0.6	0.667	0.984
22	2	178	6	1	0.25	0.667	0.364	0.963
23	3	185	5	2	0.375	0.6	0.462	0.964
24	4	172	9	1	0.308	0.8	0.444	0.946
25	2	162	4	2	0.333	0.5	0.4	0.965
26	4	252	0	1	1	0.8	0.889	0.996
27	4	259	1	1	0.8	0.8	0.8	0.992
28	2	135	1	3	0.667	0.4	0.5	0.972
29	4	131	5	2	0.444	0.667	0.533	0.951
30	2	265	4	1	0.333	0.667	0.444	0.982
31	2	241	3	1	0.4	0.667	0.5	0.984
32	5	202	9	1	0.357	0.833	0.5	0.954
33	3	77	3	2	0.5	0.6	0.545	0.941
34	5	172	5	1	0.5	0.833	0.625	0.967
35	3	118	2	1	0.6	0.75	0.667	0.976
36	5	221	1	1	0.833	0.833	0.833	0.991
37	5	152	0	1	1	0.833	0.909	0.994
38	4	156	9	0	0.308	1	0.471	0.947
39	3	146	1	1	0.75	0.75	0.75	0.987
40	2	145	3	1	0.4	0.667	0.5	0.974
41	3	155	10	2	0.231	0.6	0.333	0.929
42	1	158	6	2	0.143	0.333	0.2	0.952
43	3	181	9	2	0.25	0.6	0.353	0.944
44	5	202	9	1	0.357	0.833	0.5	0.954
45	3	177	7	1	0.3	0.75	0.429	0.957
46	4	104	4	1	0.5	0.8	0.615	0.956
47	3	120	9	1	0.25	0.75	0.375	0.925
48	3	151	6	3	0.333	0.5	0.4	0.945
49	1	186	4	4	0.2	0.2	0.2	0.959
50	5	204	0	0	1	1	1	1
51	5	203	8	0	0.385	1	0.556	0.963
52	4	193	2	1	0.667	0.8	0.727	0.985
53	3	120	2	1	0.6	0.75	0.667	0.976
54	4	178	6	1	0.4	0.8	0.533	0.963
55	1	178	0	2	1	0.333	0.5	0.989
56	2	126	3	1	0.4	0.667	0.5	0.97
57	2	111	4	2	0.333	0.5	0.4	0.95
58	4	167	6	1	0.4	0.8	0.533	0.961
59	3	141	6	1	0.333	0.75	0.462	0.954
60	3	124	3	0	0.5	1	0.667	0.977
61	2	197	5	3	0.286	0.4	0.333	0.961
62	3	165	3	1	0.5	0.75	0.6	0.977
63	2	228	0	3	1	0.4	0.571	0.987
64	3	210	9	2	0.25	0.6	0.353	0.951
65	3	198	5	2	0.375	0.6	0.462	0.966
66	3	141	2	3	0.6	0.5	0.545	0.966
67	2	127	9	1	0.182	0.667	0.286	0.928
68	3	207	1	2	0.75	0.6	0.667	0.986
69	3	166	5	0	0.375	1	0.545	0.971
70	4	265	11	1	0.267	0.8	0.4	0.957
71	3	191	2	2	0.6	0.6	0.6	0.98
72	2	144	2	3	0.5	0.4	0.444	0.967
73	5	181	4	3	0.556	0.625	0.588	0.964
74	2	239	1	2	0.667	0.5	0.571	0.988
75	1	145	1	3	0.5	0.25	0.333	0.973
76	3	204	7	1	0.3	0.75	0.429	0.963
77	3	249	0	2	1	0.6	0.75	0.992
78	3	129	4	0	0.429	1	0.6	0.971
79	3	159	10	3	0.231	0.5	0.316	0.926
80	3	173	6	2	0.333	0.6	0.429	0.957
81	4	228	7	1	0.364	0.8	0.5	0.967
82	4	173	6	0	0.4	1	0.571	0.967
83	0	223	3	3	0	0	0	0.974
84	6	219	1	1	0.857	0.857	0.857	0.991
85	5	212	7	0	0.417	1	0.588	0.969
86	4	198	0	1	1	0.8	0.889	0.995
87	4	164	1	1	0.8	0.8	0.8	0.988
88	3	250	2	2	0.6	0.6	0.6	0.984
89	1	181	3	3	0.25	0.25	0.25	0.968
90	4	142	2	0	0.667	1	0.8	0.986
91	5	216	1	0	0.833	1	0.909	0.995
92	0	198	0	5	0	0	0	0.975
93	4	171	7	0	0.364	1	0.533	0.962
94	4	173	2	1	0.667	0.8	0.727	0.983
95	2	198	7	2	0.222	0.5	0.308	0.957
96	6	195	6	0	0.5	1	0.667	0.971
97	4	162	9	0	0.308	1	0.471	0.949

98	4	209	2	1	0.667	0.8	0.727	0.986
99	3	232	1	0	0.75	1	0.857	0.996
100	5	116	3	0	0.625	1	0.769	0.976
Hasil Keseluruhan					0.442	0.698	0.541	0.971

Terlihat nilai F-Score Lebih rendah daripada nilai akurasi, disebabkan karena pada keseluruhan dataset yang dikumpulkan kata kunci dari penulis banyak yang bentuk katanya tidak sesuai dengan abstrak dan judulnya itu sendiri. Tabel 3 merupakan hasil Confusion Matrix tanpa Similarity.

No.	*Confusion Matrix*				*Precision*	*Recall*	*F-Score*	*Accuracy*
	TP	*TN*	*FP*	*FN*				
1	4	0	166	0	0.024	1	0.046	0.024
2	5	0	228	0	0.021	1	0.042	0.021
3	5	0	172	0	0.028	1	0.055	0.028
4	2	0	176	2	0.011	0.5	0.022	0.011
5	4	0	144	1	0.027	0.8	0.052	0.027
6	2	0	264	2	0.008	0.5	0.015	0.007
7	1	0	175	2	0.006	0.333	0.011	0.006
8	2	0	100	3	0.02	0.4	0.037	0.019
9	5	0	147	1	0.033	0.833	0.063	0.033
10	5	0	135	0	0.036	1	0.069	0.036
11	3	0	168	3	0.018	0.5	0.034	0.017
12	4	0	114	2	0.034	0.667	0.065	0.033
13	3	0	181	0	0.016	1	0.032	0.016
14	1	0	226	4	0.004	0.2	0.009	0.004
15	1	0	163	2	0.006	0.333	0.012	0.006
16	3	0	180	1	0.016	0.75	0.032	0.016
17	4	0	196	1	0.02	0.8	0.039	0.02
18	5	0	184	0	0.026	1	0.052	0.026
19	5	0	242	0	0.02	1	0.04	0.02
20	2	0	140	3	0.014	0.4	0.027	0.014
21	3	0	180	2	0.016	0.6	0.032	0.016
22	2	0	184	1	0.011	0.667	0.021	0.011
23	3	0	190	2	0.016	0.6	0.03	0.015
24	4	0	181	1	0.022	0.8	0.042	0.022
25	2	0	166	2	0.012	0.5	0.023	0.012
26	4	0	252	1	0.016	0.8	0.031	0.016
27	4	0	260	1	0.015	0.8	0.03	0.015
28	2	0	136	3	0.014	0.4	0.028	0.014
29	4	0	136	2	0.029	0.667	0.055	0.028
30	2	0	269	1	0.007	0.667	0.015	0.007
31	2	0	244	1	0.008	0.667	0.016	0.008
32	5	0	211	1	0.023	0.833	0.045	0.023
33	3	0	80	2	0.036	0.6	0.068	0.035
34	5	0	177	1	0.027	0.833	0.053	0.027
35	3	0	120	1	0.024	0.75	0.047	0.024
36	5	0	222	1	0.022	0.833	0.043	0.022
37	5	0	152	1	0.032	0.833	0.061	0.032
38	4	0	165	0	0.024	1	0.046	0.024
39	3	0	147	1	0.02	0.75	0.039	0.02
40	2	0	148	1	0.013	0.667	0.026	0.013
41	3	0	165	2	0.018	0.6	0.035	0.018
42	1	0	164	2	0.006	0.333	0.012	0.006
43	3	0	190	2	0.016	0.6	0.03	0.015
44	5	0	211	1	0.023	0.833	0.045	0.023
45	3	0	184	1	0.016	0.75	0.031	0.016
46	4	0	108	1	0.036	0.8	0.068	0.035
47	3	0	129	1	0.023	0.75	0.044	0.023
48	3	0	157	3	0.019	0.5	0.036	0.018
49	1	0	190	4	0.005	0.2	0.01	0.005
50	5	0	204	0	0.024	1	0.047	0.024
51	5	0	211	0	0.023	1	0.045	0.023
52	4	0	195	1	0.02	0.8	0.039	0.02
53	3	0	122	1	0.024	0.75	0.047	0.024
54	4	0	184	1	0.021	0.8	0.041	0.021
55	1	0	178	2	0.006	0.333	0.011	0.006
56	2	0	129	1	0.015	0.667	0.03	0.015
57	2	0	115	2	0.017	0.5	0.033	0.017
58	4	0	173	1	0.023	0.8	0.044	0.022
59	3	0	147	1	0.02	0.75	0.039	0.02
60	3	0	127	0	0.023	1	0.045	0.023
61	2	0	202	3	0.01	0.4	0.019	0.01
62	3	0	168	1	0.018	0.75	0.034	0.017
63	2	0	228	3	0.009	0.4	0.017	0.009
64	3	0	219	2	0.014	0.6	0.026	0.013
65	3	0	203	2	0.015	0.6	0.028	0.014
66	3	0	143	3	0.021	0.5	0.039	0.02
67	2	0	136	1	0.014	0.667	0.028	0.014
68	3	0	208	2	0.014	0.6	0.028	0.014
69	3	0	171	0	0.017	1	0.034	0.017
70	4	0	276	1	0.014	0.8	0.028	0.014
71	3	0	193	2	0.015	0.6	0.03	0.015
72	2	0	146	3	0.014	0.4	0.026	0.013
73	5	0	185	3	0.026	0.625	0.051	0.026
74	2	0	240	2	0.008	0.5	0.016	0.008
75	1	0	146	3	0.007	0.25	0.013	0.007
76	3	0	211	1	0.014	0.75	0.028	0.014
77	3	0	249	2	0.012	0.6	0.023	0.012
78	3	0	133	0	0.022	1	0.043	0.022
79	3	0	169	3	0.017	0.5	0.034	0.017
80	3	0	179	2	0.016	0.6	0.032	0.016
81	4	0	235	1	0.017	0.8	0.033	0.017
82	4	0	179	0	0.022	1	0.043	0.022
83	0	0	226	3	0	0	0	0
84	6	0	220	1	0.027	0.857	0.052	0.026
85	5	0	219	0	0.022	1	0.044	0.022
86	4	0	198	1	0.02	0.8	0.039	0.02
87	4	0	165	1	0.024	0.8	0.046	0.024
88	3	0	252	2	0.012	0.6	0.023	0.012
89	1	0	184	3	0.005	0.25	0.011	0.005
90	4	0	144	0	0.027	1	0.053	0.027
91	5	0	217	0	0.023	1	0.044	0.023
92	0	0	198	5	0	0	0	0
93	4	0	178	0	0.022	1	0.043	0.022
94	4	0	175	1	0.022	0.8	0.043	0.022
95	2	0	205	2	0.01	0.5	0.019	0.01
96	6	0	201	0	0.029	1	0.056	0.029
97	4	0	171	0	0.023	1	0.045	0.023
98	4	0	211	1	0.019	0.8	0.036	0.019
99	3	0	233	0	0.013	1	0.025	0.013
100	5	0	119	0	0.04	1	0.078	0.04
Hasil Keseluruhan					0.017	0.698	0.034	0.017

Dari keseluruhan Confusion Matrix yang terlihat system yang dikembangkan telah berhasil menghasilkan kata kunci yang baik desuai dengan penulis, tetapi urutan dari rangkingnya masih berantakan sehingga harus menampilkan keseluruhan kata kunci terlebih dahulu.

V. KESIMPULAN

Dari Penelitian yang dilakukan dan pengujian-pengujian yang dilakukan, bahwa YAKE dapat melakukan ekstraksi kata kunci pada Bahasa Indonesia

dengan hasil akurasi sebesar 97.05% dan F-Score sebesar 54,1%. Konfigurasi yang dipakai yaitu Levenshtein Distance <2 dan Kata kunci yang dikeluarkan seluruh (all).

UCAPAN TERIMA KASIH/ ACKNOWLEDGMENT

Penelitian/ Publikasi artikel ini dibiayai oleh: Anggaran DIPA Badan Layanan Umum Universitas Sriwijaya Tahun Anggaran 2021. Nomor SP DIPA-023.17.2.677515/2021, tanggal 13 Desember 2021,. Nomor: 0019/UN9/SK.LP2M.PT/2022 tanggal 15 Juni 2022.

REFERENSI

[1] Arini, Ni Wayan Sri, Ida Bagus Putu Widja, and I. Komang Rinartha Yasa Negara. "Analisis Frekuensi Kata untuk Mengekstrak Kata Kunci dari Artikel Ilmiah Berbahasa Indonesia." Jurnal Eksplora Informatika 8.2 (2019): 80-84.

[2] Behara, Krishna NS, Ashish Bhaskar, and Edward Chung. "A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance." *Transportation Research Part C: Emerging Technologies* 111 (2020): 513-530.

[3] Campos, Ricardo, et al. "YAKE! collection-independent automatic keyword extractor." European Conference on Information Retrieval. Springer, Cham, 2018.

[4] Firdausillah, Fahri, and Erika Devi Udayanti. "Keyphrase Extraction on Covid-19 Tweets Based on Doc2Vec and YAKE." Journal of Applied Intelligent System 6.1 (2021): 23-31.

[5] Kurniawan, Ahmad. Aplikasi sistem ekstraksi kata kunci berbahasa indonesia menggunakan algoritma textrank studi kasus data wikipedia Indonesia. BS thesis. Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

[6] Lamasigi, Zulfrianto Yusrin. "DCT Untuk Ekstraksi Fitur Berbasis GLCM Pada Identifikasi Batik Menggunakan K-NN." *Jambura Journal of Electrical and Electronics Engineering* 3.1 (2021): 1-6.

[7] Lubis, Andre Hasudungan, Ali Ikhwan, and Phak Len Eh Kan. "Combination of levenshtein distance and rabin-karp to improve the accuracy of document equivalence level." International Journal of Engineering & Technology 7.2.27 (2018): 17-21.

[8] Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik. "Textual keyword extraction and summarization: State-of-the-art." *Information Processing & Management* 56.6 (2019): 102088.

[9] Perdana, Novario Jaya. "IMPLEMENTASI ALGORITMA GOOGLE LATENT SEMANTIC DISTANCE UNTUK EKSTRAKSI RANGKAIAN KATA KUNCI ARTIKEL JURNAL ILMIAH." Computatio: Journal of Computer Science and Information Systems 2.2 (2018): 186-195.

[10] Yunmar, Rajif Agung, Andika Setiawan, and Hartanto Tantriawan. "The Combination of YAKE and Language Processing for Unsupervised Term Extraction Ontology Learning." IOP Conference Series: Earth and Environmental Science. Vol. 537. No. 1. IOP Publishing, 2020.

[11] Md. Mosabbir Hossain, Md. Farhan Labib, Ahmed Sady Rifat, Amit Kumar Das, Monira Mukta. “Auto-correction of english to bengali transliteration system using levenshtein distance.” International Conference on Smart Computing & Communications (ICSCC) IEEE. Jun 28, 2019: 1-5.