

*KEYPHRASE EXTRACTION MENGGUNAKAN PRETRAINED  
LANGUAGE MODEL ROBERTA DAN POSITION AWARE GRAPH*

*Diajukan Sebagai Syarat Untuk Menyelesaikan  
Pendidikan Program Strata-1 Pada  
Jurusan Teknik Informatika*



Oleh:

M. Bintang Khadafi  
NIM: 09021282025070

**Jurusan Teknik Informatika**  
**FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA**  
**2024**

LEMBAR PENGESAHAN SKRIPSI

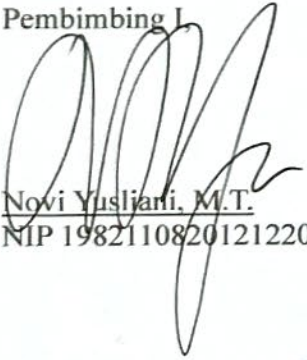
*KEYPHRASE EXTRACTION MENGGUNAKAN PRETRAINED  
LANGUAGE MODEL ROBERTA DAN POSITION AWARE GRAPH*


Oleh:

M. Bintang Khadafi  
NIM: 09021282025070

Palembang, 8 Maret 2024  
Pembimbing II,

Pembimbing I

  
Novi Yusliarti, M.T.  
NIP 198211082012122001

  
Desty Rodiah, S.Kom., M.T.  
NIP 198912212020122011

Mengetahui,  
Ketua Jurusan Teknik Informatika

  
Dr. M. Fachrurrozi, S.Si., M.T.  
NIP 198005222008121002

## TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

Pada hari Jumat tanggal 8 Maret 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : M. Bintang Khadafi


NIM : 09021282025070

Judul : *Keyphrase Extraction Menggunakan Pretrained Language Model RoBERTa dan Position Aware Graph*

dan dinyatakan **LULUS**.

1. Ketua

Dr. Abdiansah, S.Kom., M.CS.  
NIP 198410012009121005



2. Penguji I

Rizki Kurniati, M.T.  
NIP 199107122019032016



3. Pembimbing I

Novi Yusliani, M.T.  
NIP 198211082012122001



4. Pembimbing II

Desty Rodiah, S.Kom., M.T.  
NIP 198912212020122011



Mengetahui,  
Ketua Jurusan Teknik Informatika



Dr. M. Fachrurrozi, S.Si., M.T.  
NIP 198005222008121002



## HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : M. Bintang Khadafi  
NIM : 09021282025070  
Program Studi : Teknik Informatika  
Judul Skripsi : *Keyphrase Extraction Menggunakan Pretrained Language Model RoBERTa dan Position Aware Graph*

Hasil pengecekan *Software iThenticate/Turnitin*: 4%

Menyatakan bahwa laporan proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan dari pihak mana pun.



Palembang, 29 Februari 2024



M. Bintang Khadafi  
NIM 09021282025070

## **MOTTO DAN PERSEMBAHAN**

Motto:

“Karena sesungguhnya sesudah kesulitan itu ada kemudahan, sesungguhnya sesudah kesulitan itu ada kemudahan.”

- Al Insyirah: 5-6

Kupersembahkan Karya Tulis ini Kepada:

- Allah SWT
- Orang Tua
- Keluarga Besar
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

## ***ABSTRACT***

*Searching for the desired scientific journal from the many published journals is currently a challenge for researchers because it takes a lot of time. Therefore, the use of keyphrases or keywords that represent the content of scientific journals is needed as a solution to save search time. The keyphrase extraction process is carried out to obtain relevant keyphrases. One of the keyphrase extraction methods is Position Aware Graph combined with a pretrained language model called RoBERTa to represent language vectors. This research aims to create a keyphrase extraction system using Pretrained Language Model RoBERTa and Position Aware Graph. In the test, 100 datasets containing title and abstract data of scientific journals were used. The test results show that the best configuration is keyphrase extraction with the selection of 10 selected keyphrases with an average f1-score value of 0.1434.*

*Keywords: Keyphrase, Keyphrase Extraction, Position Aware Graph, Pretrained Language Model, RoBERTa, Scientific Journal.*

## ABSTRAK

Pencarian jurnal ilmiah yang diinginkan dari banyaknya jurnal yang dipublikasikan saat ini menjadi tantangan bagi para peneliti karena memakan waktu yang tidak sedikit. Oleh karena itu, penggunaan *keyphrase* atau kata kunci yang mewakili isi dari jurnal ilmiah diperlukan sebagai solusi untuk menghemat waktu pencarian. Proses mengekstraksi kata kunci atau *keyphrase extraction* dilakukan untuk mendapatkan *keyphrase* yang relevan. Salah satu metode *keyphrase extraction* adalah *Position Aware Graph* yang dikombinasikan dengan model bahasa yang telah dilatih bernama RoBERTa untuk merepresentasikan vektor bahasa. Penelitian ini bertujuan untuk membuat sistem *keyphrase extraction* menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*. Dalam pengujiannya, digunakan 100 *dataset* yang berisi data judul dan abstrak jurnal ilmiah. Hasil pengujian menunjukkan bahwa konfigurasi terbaik adalah *keyphrase extraction* dengan pemilihan 10 *keyphrase* terpilih dengan rata-rata nilai *f1-score* sebesar 0.1434.

Kata Kunci: Jurnal Ilmiah, *Keyphrase*, *Keyphrase Extraction*, *Position Aware Graph*, *Pretrained Language Model*, RoBERTa.

## KATA PENGANTAR

Puji syukur kepada Allah SWT atas rahmat dan nikmat Nya yang lebih diberikan kepada penulis sehingga dapat menyelesaikan skripsi ini dengan baik. Skripsi ini disusun sebagai salah satu syarat menyelesaikan Pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam menyelesaikan skripsi ini, penulis menerima bantuan, bimbingan dan dukungan dari banyak pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin menyampaikan terima kasih kepada:

1. Allah SWT atas rahmat dan nikmat-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua dan keluarga besar yang telah mendoakan, memberi semangat, memotivasi, dan nasihat untuk menyelesaikan skripsi ini.
3. Bapak Dr. M. Fachrurrozi, S.Si., M.T. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Ibu Mastura Diana Marieska, M.T. selaku Dosen Pembimbing Akademik yang telah memberikan banyak sekali bantuan dan arahan kepada penulis selama perkuliahan.
5. Ibu Novi Yusliani, M.T. selaku Dosen Pembimbing I dan Ibu Desty Rodiah, S.Kom., M.T. selaku Dosen Pembimbing II yang telah membimbing serta memberikan arahan kepada penulis selama proses pengerjaan skripsi.
6. Seluruh dosen program studi serta admin Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Saudara Sheva Satrian, M. Farhan Ghifari, Alif Toriq Alkausar, Fadhil Zahran Muwafa, Anwaripasha Akbar, dan Bayu Daru Pangestu sebagai teman yang selalu memotivasi penulis untuk semangat mengerjakan skripsi.
8. Seluruh Staf Administrasi dan Pegawai Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi tugas akhir penulis.



9. Seluruh teman-teman yang telah memberikan saran, motivasi, dan semangat kepada penulis.
10. Pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan dikarenakan kurangnya pengalaman dan pengetahuan penulis. Oleh karena itu penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian selanjutnya. Semoga tugas akhir ini dapat bermanfaat. Terima kasih.

Palembang, 29 Februari 2024

Penulis



M. Bintang Khadafi

## DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI .....	ii
TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI .....	iii
HALAMAN PERNYATAAN .....	iv
MOTTO DAN PERSEMBAHAN .....	v
<i>ABSTRACT</i> .....	vi
ABSTRAK .....	vii
KATA PENGANTAR .....	viii
DAFTAR ISI .....	x
DAFTAR TABEL .....	xiv
DAFTAR GAMBAR .....	xvii
BAB I PENDAHULUAN .....	I-1
1.1    Pendahuluan .....	I-1
1.2    Latar Belakang Masalah .....	I-1
1.3    Rumusan Masalah .....	I-3
1.4    Tujuan Penelitian .....	I-4
1.5    Manfaat Penelitian .....	I-4
1.6    Batasan Penelitian .....	I-4
1.7    Sistematika Penulisan .....	I-5
1.8    Kesimpulan .....	I-6
BAB II KAJIAN LITERATUR .....	II-1
2.1    Pendahuluan .....	II-1
2.2    Landasan Teori .....	II-1
2.2.1 <i>Keyphrase Extraction</i> .....	II-1
2.2.2 <i>Pretrained Language Model RoBERTa</i> .....	II-3
2.2.3 <i>Position Aware Graph</i> .....	II-5
2.2.4 <i>Confusion Matrix</i> .....	II-9

2.3	Penelitian Lain yang Relevan .....	II-11
2.4	Kesimpulan .....	II-12
BAB III METODOLOGI PENELITIAN.....		III-1
3.1	Pendahuluan.....	III-1
3.2	Pengumpulan Data.....	III-1
3.2.1	Jenis dan Sumber Data.....	III-1
3.2.2	Metode Pengumpulan Data.....	III-2
3.3	Tahapan Penelitian.....	III-3
3.3.1	Mengumpulkan Data.....	III-4
3.3.2	Menentukan Kerangka Kerja Penelitian .....	III-4
3.3.3	Menentukan Kriteria Pengujian .....	III-7
3.3.4	Menentukan Format Data Pengujian .....	III-7
3.3.5	Menentukan Alat Bantu Penelitian .....	III-8
3.3.6	Melakukan Pengujian Penelitian .....	III-9
3.3.7	Melakukan Analisis dan Menarik Kesimpulan Penelitian.....	III-9
3.4	Kesimpulan .....	III-11
BAB IV PENGEMBANGAN PERANGKAT LUNAK .....		IV-1
4.1	Pendahuluan.....	IV-1
4.2	Fase Insepsi.....	IV-1
4.2.1	Pemodelan Bisnis.....	IV-1
4.2.2	Kebutuhan Sistem .....	IV-2
4.2.3	Analisis dan Desain .....	IV-3
4.2.3.1	Analisis Kebutuhan Perangkat Lunak .....	IV-3
4.2.3.2	Analisis Data .....	IV-4
4.2.3.3	Analisis <i>Preprocessing</i> .....	IV-4
4.2.3.5	Analisis <i>Position Aware Graph Centrality Scoring</i> .....	IV-14
4.2.3.6	Analisis Peningkatan Kandidat <i>Keyphrase</i> .....	IV-16
4.2.3.7	Analisis Evaluasi <i>Confusion Matrix</i> .....	IV-17

4.2.3.8	Desain Perangkat Lunak.....	IV-17
4.3	Fase Elaborasi.....	IV-22
4.3.1	Pemodelan Bisnis.....	IV-22
4.3.1.1	Perancangan Data.....	IV-23
4.3.1.2	Perancangan <i>Interface</i> (Antarmuka).....	IV-23
4.3.2	Kebutuhan Sistem.....	IV-26
4.3.3	Analisis dan Perancangan.....	IV-27
4.3.3.1	Activity Diagram.....	IV-27
4.3.3.2	Sequence Diagram.....	IV-30
4.4	Fase Konstruksi.....	IV-31
4.4.1	Kebutuhan Sistem.....	IV-31
4.4.2	Implementasi.....	IV-33
4.4.2.1	Implementasi <i>Class</i> .....	IV-33
4.4.2.2	Implementasi <i>Interface</i> (Antarmuka).....	IV-34
4.5	Fase Transisi.....	IV-36
4.5.1	Pemodelan Bisnis.....	IV-37
4.5.2	Rencana Pengujian.....	IV-37
4.5.3	Implementasi.....	IV-38
4.6	Kesimpulan.....	IV-39
BAB V HASIL DAN ANALISIS PENELITIAN.....		V-1
5.1	Pendahuluan.....	V-1
5.2	Data Hasil Penelitian.....	V-1
5.2.1	Konfigurasi Percobaan.....	V-1
5.2.2	Data Hasil Percobaan <i>Keyphrase Extraction</i> Menggunakan <i>Position Aware Graph</i> .....	V-18
5.2.3	Data Hasil Percobaan <i>Keyphrase Extraction</i> Menggunakan <i>Traditional Graph Centrality</i> .....	V-31
5.3	Analisis Hasil Penelitian.....	V-44
5.4	Kesimpulan.....	V-47

BAB VI KESIMPULAN DAN SARAN .....	VI-1
6.1    Pendahuluan.....	VI-1
6.2    Kesimpulan .....	VI-1
6.3    Saran .....	VI-2
DAFTAR PUSTAKA .....	vii
DAFTAR LAMPIRAN .....	x

## DAFTAR TABEL

<b>Tabel III-1.</b> Contoh Data Judul Dokumen yang digunakan .....	III-2
<b>Tabel III-2.</b> Contoh Data Abstrak Dokumen yang digunakan .....	III-2
<b>Tabel IV-1.</b> Kebutuhan Fungsional Sistem.....	IV-2
<b>Tabel IV-2.</b> Kebutuhan Non-Fungsional Sistem .....	IV-3
<b>Tabel IV-3.</b> Contoh Data Judul Dokumen .....	IV-4
<b>Tabel IV-4.</b> Contoh Data Abstrak Dokumen .....	IV-5
<b>Tabel IV-5.</b> Data Input Hasil Penggabungan Judul dan Abstrak Dokumen .....	IV-5
<b>Tabel IV-6.</b> Hasil Penghapusan Karakter Spesial dan Tokenisasi Data Input .....	IV-7
<b>Tabel IV-7.</b> Hasil Penghapusan Stopword pada Data Input .....	IV-8
<b>Tabel IV-8.</b> Hasil <i>POS Tagging</i> Data Input .....	IV-9
<b>Tabel IV-9.</b> Pola <i>Grammar</i> Pembentukan <i>Noun Phrase</i> .....	IV-10
<b>Tabel IV-10.</b> <i>Golden Keyphrase</i> yang telah Mengalami <i>Lowercasing</i> .....	IV-11
<b>Tabel IV-11.</b> Vektor Representasi dari Proses <i>Embedding</i> .....	IV-12
<b>Tabel IV-12.</b> Skor setiap Kandidat <i>Keyphrase</i> dihitung dengan PAG.....	IV-14
<b>Tabel IV-13.</b> Kandidat <i>Keyphrase</i> yang Memperoleh Skor Tertinggi.....	IV-17
<b>Tabel IV-14.</b> Hasil Perhitungan Evaluasi <i>Confusion Matrix</i> .....	IV-17
<b>Tabel IV-15.</b> Definisi <i>Actor</i> .....	IV-19
<b>Tabel IV-16.</b> Definisi <i>Use Case</i> .....	IV-19
<b>Tabel IV-17.</b> Skenario <i>Use Case</i> Mengekstraksi <i>Keyphrase</i> .....	IV-20
<b>Tabel IV-18.</b> Skenario <i>Use Case</i> Evaluasi Hasil .....	IV-21
<b>Tabel IV-19.</b> Implementasi <i>Class</i> Perangkat Lunak .....	IV-33

<b>Tabel IV-20.</b> Rencana Pengujian <i>Use Case</i> Mengekstraksi Keyphrase .....	IV-37
<b>Tabel IV-21.</b> Rencana Pengujian <i>Use Case</i> Mengevaluasi Hasil .....	IV-37
<b>Tabel IV-22.</b> Pengujian <i>Use Case</i> Mengekstraksi <i>Keyphrase</i> .....	IV-38
<b>Tabel IV-23.</b> Pengujian <i>Use Case</i> Mengevaluasi Hasil.....	IV-38
<b>Tabel V-1.</b> Sampel Data Uji Judul Dokumen .....	V-3
<b>Tabel V-2.</b> Sampel Data Uji Abstrak Dokumen.....	V-4
<b>Tabel V-3.</b> Kandidat Keyphrase dari Sampel Data Uji.....	V-10
<b>Tabel V-4.</b> <i>Keyphrase</i> Terpilih Top 5 dari Percobaan <i>Position Aware Graph</i> .....	V-19
<b>Tabel V-5.</b> <i>Keyphrase</i> Terpilih Top 10 dari Percobaan <i>Position Aware Graph</i> .....	V-20
<b>Tabel V-6.</b> <i>Keyphrase</i> Terpilih Top 20 dari Percobaan <i>Position Aware Graph</i> .....	V-22
<b>Tabel V-7.</b> <i>Keyphrase</i> Terpilih Top 30 dari Percobaan <i>Position Aware Graph</i> .....	V-24
<b>Tabel V-8.</b> Hasil Evaluasi i <i>Extraction</i> PAG Top 5.....	V-28
<b>Tabel V-9.</b> Hasil Evaluasi <i>Keyphrase Extraction</i> PAG Top 10.....	V-29
<b>Tabel V-10.</b> Hasil Evaluasi <i>Keyphrase Extraction</i> PAG Top 20.....	V-29
<b>Tabel V-11.</b> Hasil Evaluasi <i>Keyphrase Extraction</i> PAG Top 30.....	V-30
<b>Tabel V-12.</b> Rata-rata Nilai Evaluasi <i>Confusion Matrix Keyphrase</i> <i>Extraction</i> PAG .....	V-30
<b>Tabel V-13.</b> <i>Keyphrase</i> Terpilih Top 5 dari Percobaan <i>Traditional</i> <i>Graph Centrality</i> .....	V-32
<b>Tabel V-14.</b> <i>Keyphrase</i> Terpilih Top 10 dari Percobaan <i>Traditional</i> <i>Graph Centrality</i> .....	V-33
<b>Tabel V-15.</b> <i>Keyphrase</i> Terpilih Top 20 dari Percobaan <i>Traditional</i> <i>Graph Centrality</i> .....	V-35

<b>Tabel V-16.</b> Keyphrase Terpilih Top 30 dari Percobaan <i>Traditional</i>	
<i>Graph Centrality</i> .....	V-37
<b>Tabel V-17.</b> Hasil Evaluasi Keyphrase Extraction <i>Traditional</i>	
<i>Graph Centrality</i> Top 5 .....	V-41
<b>Tabel V-18.</b> Hasil Evaluasi Keyphrase Extraction <i>Traditional</i>	
<i>Graph Centrality</i> Top 10.....	V-41
<b>Tabel V-19.</b> Hasil Evaluasi Keyphrase Extraction <i>Traditional</i>	
<i>Graph Centrality</i> Top 20.....	V-42
<b>Tabel V-20.</b> Hasil Evaluasi <i>Keyphrase Extraction Traditional</i>	
<i>Graph Centrality</i> Top 30.....	V-42
<b>Tabel V-21.</b> Rata-rata Nilai Evaluasi <i>Confusion Matrix Keyphrase Extraction</i>	
<i>Traditional Graph Centrality</i> .....	V-43



## DAFTAR GAMBAR

<b>Gambar II-1.</b> Arsitektur RoBERTa untuk proses MLM dan NSP.....	II-5
<b>Gambar II-2.</b> Tabel <i>Confusion Matrix</i> .....	II-9
<b>Gambar III-1.</b> Kerangka Kerja Penelitian.....	III-3
<b>Gambar III-2.</b> Kerangka Kerja Penelitian.....	III-6
<b>Gambar IV-1.</b> <i>Use Case Diagram</i> Perangkat Lunak.....	IV-18
<b>Gambar IV-2.</b> Rancangan Antarmuka Halaman Utama .....	IV-24
<b>Gambar IV-3.</b> Desain Antarmuka Halaman Hasil dan Pembahasan .....	IV-25
<b>Gambar IV-4.</b> Rancangan Antarmuka Halaman Hasil dan Evaluasi Hasil .....	IV-26
<b>Gambar IV-5.</b> <i>Activity Diagram</i> Mengekstraksi <i>Keyphrase</i> .....	IV-28
<b>Gambar IV-6.</b> <i>Activity Diagram</i> Mengevaluasi Hasil .....	IV-29
<b>Gambar IV-7.</b> <i>Sequence Diagram</i> Mengekstraksi <i>Keyphrase</i> .....	IV-30
<b>Gambar IV-8.</b> <i>Sequence Diagram</i> Mengevaluasi Hasil .....	IV-31
<b>Gambar IV-9.</b> <i>Class Diagram</i> Perangkat Lunak .....	IV-32
<b>Gambar IV-10.</b> Implementasi Halaman Utama Perangkat Lunak.....	IV-34
<b>Gambar IV-11.</b> Implementasi Halaman Hasil dan Pembahasan.....	IV-35
<b>Gambar IV-12.</b> Implementasi Halaman Hasil dan Evaluasi Hasil .....	IV-36
<b>Gambar V-1.</b> Diagram Batang Evaluasi Percobaan <i>Keyphrase Extraction</i> PAG ..	V-31
<b>Gambar V-2.</b> Diagram Batang Evaluasi Percobaan <i>Keyphrase Extraction</i> <i>Traditional Graph Centrality</i> .....	V-44
<b>Gambar V-3.</b> Perbandingan Hasil Konfigurasi <i>Position Aware Graph</i> dan <i>Traditional Graph Centrality</i> .....	V-45

# **BAB I**

## **PENDAHULUAN**

### **1.1 Pendahuluan**

Pada bab ini akan diuraikan mengenai latar belakang masalah yang diangkat pada penelitian ini, rumusan masalah, tujuan penulisan, manfaat penelitian, batasan penelitian, dan sistematika penulisan. Secara garis besar, bab ini akan membahas penjelasan umum dari keseluruhan penelitian.

### **1.2 Latar Belakang Masalah**

Dewasa ini, manusia dituntut untuk dapat mengambil informasi dengan cepat selaras dengan banyaknya sumber informasi yang dipublikasikan. Salah satunya adalah banyaknya jurnal ilmiah yang telah dipublikasikan akan membuat para peneliti, akademisi, maupun masyarakat pada umumnya kesulitan dan memakan waktu yang lama untuk mencari jurnal yang diinginkan. Untuk mengatasi masalah tersebut, dimanfaatkanlah *keyphrase* yang merepresentasikan konteks keseluruhan isi dari sebuah jurnal. Menurut Mothe et al. (2018), kata kunci atau *keyword* atau bisa disebut juga *keyphrase* dapat menggambarkan isi dari publikasi ilmiah secara umum.

*Keyphrase extraction* atau ekstraksi kata kunci adalah sebuah proses untuk mengidentifikasi dan mengekstrak kata atau frase yang paling penting dan relevan dari sebuah dokumen (Campos et al., 2018). Ekstraksi kata kunci juga dapat diartikan sebagai sebuah proses untuk menemukan kata atau frase yang paling mewakili isi dari

sebuah dokumen (Firdausillah & Udayanti, 2021). Banyak metode yang telah digunakan dalam proses *keyphrase extraction*, salah satunya adalah *Position Aware Graph*.

*Position Aware Graph* adalah algoritma pemeringkatan berbasis grafik yang mempertimbangkan bobot posisi untuk menentukan pentingnya kandidat frasa kunci dalam ekstraksi frasa kunci yang tidak diawasi. Algoritma ini menggunakan grafik yang dibangun berdasarkan dokumen sumber, di mana kandidat frasa kunci dianggap sebagai simpul dan tepi antara mereka diberi bobot berdasarkan kesamaan mereka (Zhang et al., 2023). Dalam penelitian yang dilakukan oleh Zhang et al. (2023) menunjukkan bahwa *Position Aware Graph* terbukti mengungguli beberapa algoritma pemeringkatan lain seperti SIFRank (Sun et al., 2020), UKERank (Liang et al., 2021), dan EmbedRank (Bennani-Smires et al., 2018).

Menurut Zhang et al. (2023), untuk menghasil kandidat kata kunci yang berkualitas dan akurat maka diperlukan *Pretrained Language Model*. *Pretrained Language Model* (PLM) adalah model-model yang dipelajari dengan menggunakan sinyal pelatihan yang tidak diawasi dari korpus teks yang besar (Elazar et al., 2021). PLM dapat memberikan penyematan (*embedding*) teks berkualitas tinggi dengan mengkodekan kata, kalimat, atau dokumen dengan konteks yang dinamis. Representasi kontekstual yang diperoleh dari PLM dapat memberikan informasi semantik yang lebih baik dibandingkan dengan metode representasi statis seperti Word2Vec atau GloVe (Zhang et al., 2023). Salah satu contoh *Pretrained Language Model* yang telah optimal adalah RoBERTa (Liu et al., 2019).

Menurut Liu et al. (2019), RoBERTa atau *A Robustly Optimized BERT Pretraining Approach* adalah pendekatan pretraining yang dioptimalkan untuk model BERT atau *Bidirectional Encoder Representations* (Devlin et al., 2019) dari Transformers. RoBERTa menggabungkan beberapa perubahan desain yang menghasilkan peningkatan kinerja dibandingkan dengan BERT. Dengan kombinasi dari *pretraining* dan *finetuning*, RoBERTa dapat menghasilkan representasi yang kaya dan kontekstual dari teks, yang dapat digunakan untuk berbagai tugas pemahaman bahasa alami (Liu et al., 2019).

Berdasarkan penelitian tersebut, RoBERTa yang merupakan bentuk optimalisasi dari BERT (Liu et al., 2019) dapat digunakan bersama *Position Aware Graph* untuk melakukan proses *keyphrase extraction*. Jadi, peneliti akan melakukan penelitian *Keyphrase Extraction* menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*.

### 1.3 Rumusan Masalah

Berdasarkan latar belakang masalah maka dirumuskanlah masalah sebagai berikut.

1. Bagaimana mengembangkan sistem *keyphrase extraction* dengan menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*?
2. Bagaimana nilai *F1-score* sistem *keyphrase extraction* dengan menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*?

#### 1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Menghasilkan sebuah sistem yang dapat melakukan ekstraksi kata kunci menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*.
2. Mengetahui nilai *F1-score* sistem ekstraksi kata kunci menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*.

#### 1.5 Manfaat Penelitian

Manfaat yang didapat dari penelitian ini adalah sebagai berikut.

1. Sistem yang dibuat dapat digunakan untuk melakukan proses *keyphrase extraction*.
2. Penelitian ini dapat digunakan sebagai referensi dibidang penelitian terkait *keyphrase extraction*.

#### 1.6 Batasan Penelitian

Batasan dalam penelitian ini adalah sebagai berikut.

1. Publikasi ilmiah yang digunakan adalah publikasi berbahasa Indonesia.
2. Data pada publikasi ilmiah yang digunakan adalah judul dan abstrak berjumlah 100 *dataset*.

## **1.7 Sistematika Penulisan**

Sistematika penulisan yang digunakan pada penelitian ini mengikuti Standar Operasional Penulisan (SOP) tugas akhir yang ditetapkan oleh Fakultas Ilmu Komputer Universitas Sriwijaya yaitu sebagai berikut.

### **BAB I. PENDAHULUAN**

Pada bab ini akan dibahas mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan yang digunakan. Secara garis besar, bab ini akan menjelaskan keseluruhan isi dari penelitian ini.

### **BAB II. KAJIAN LITERATUR**

Pada bab ini akan dibahas mengenai rujukan teori yang digunakan sebagai penunjang penelitian. Bab ini akan memuat berbagai kajian literatur dan juga penelitian lain yang relevan dengan topik penelitian yaitu mengenai *keyphrase extraction*, *Pretrained Language Model RoBERTa*, dan *Position Aware Graph*.

### **BAB III. METODOLOGI PENELITIAN**

Pada bab ini akan dibahas mengenai tahapan-tahapan yang dilakukan dalam penelitian seperti metode pengumpulan data, penentuan kerangka penelitian, metode

pengujian dan penjadwalan penelitian. Setiap tahapan penelitian dibahas secara mendetail sesuai kerangka kerja yang telah dirancang.

#### **BAB IV. PENGEMBANGAN PERANGKAT LUNAK**

Pada bab ini akan dibahas mengenai perancangan perangkat lunak mulai dari analisis kebutuhan perangkat lunak hingga pengujian pada perangkat lunak untuk evaluasi.

#### **BAB V. HASIL DAN ANALISIS PENELITIAN**

Pada bab ini akan dibahas hasil penelitian berdasarkan pengujian metode dan perangkat lunak yang telah dilakukan. Analisis penelitian tersebut akan menjadi dasar kesimpulan penelitian.

#### **BAB VI. KESIMPULAN DAN SARAN**

Pada bab ini akan dibahas mengenai kesimpulan yang didapat dari penelitian yang telah dilakukan serta memberikan saran untuk pembangunan sistem dan penelitian serupa selanjutnya.

### **1.8 Kesimpulan**

Berdasarkan penjelasan yang dibahas pada sub bab sebelumnya, penelitian ini akan membahas *Keyphrase Extraction* menggunakan *Pretrained Language Model* RoBERTa dan *Position Aware Graph*.

## DAFTAR PUSTAKA

- Banga, R., & Mehndiratta, P. 2018. A survey on part of speech tagging. *International Journal of Computer Applications*, 180(1), 1–7.
- Bawden, D., & Robinson, L. 2020. Information overload: An overview. (<https://openaccess.city.ac.uk/id/eprint/23544/>)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Cambridge Dictionary. Noun phrase. In Cambridge Dictionary. Diakses pada 7 Oktober 2023 dari <https://dictionary.cambridge.org/dictionary/english/noun-phrase>.
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. 2018. YAKE! collection-independent automatic keyword extractor. *European Conference on Information Retrieval*. Springer, Cham.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186.
- Duwila, S. 2020. The use of noun phrases in academic writing by Indonesian EFL learners. *Journal of English Language Teaching and Linguistics*, 5(2), 173–186.
- Dwitama, Bintang. 2022. Keyphrase Extraction Berbahasa Indonesia Menggunakan Metode PositionRank. Universitas Sriwijaya.



- Elazar, Y., Goldberg, Y., & Berant, J. 2021. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment analysis. *Natural Language Engineering*, 27(1), 1–23.
- Firdausillah, F., & Udayanti, E. D. 2021. Keyphrase Extraction on Covid-19 Tweets Based on Doc2Vec and YAKE Algorithm. *Journal of Physics: Conference Series*, 1807(1), 012006.
- Hermawan, I., & Ismiati, A. 2019. Text preprocessing pada data teks bahasa Indonesia menggunakan metode text mining. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(3), 379-385.
- Hidayat, R. 2018. Text preprocessing pada data teks bahasa Indonesia menggunakan metode text mining. *Jurnal Ilmiah Teknologi Informasi Asia*, 12(1), 1-10.
- Kim, Y., Kim, J. H., Lee, J. M., Jang, M. J., Yum, Y. J., Kim, S., ... & Song, S. 2022. A pre-trained BERT for Korean medical natural language processing. *Scientific Reports*, 12(1), 13847.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Mothe, J., Ramiandrisoa, F., & Rasolomanana, M. 2018. Automatic keyphrase extraction using graph-based methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 728-730).
- Nasar, Zara., Jaffry, S, Q., and Malik, M, K. 2019. "Textual keyword extraction and summarization: State-of-the-art." *Information Processing & Management* 56.6: 102088.

- Plakasa, Gerald. 2022. Ekstraksi Kata Kunci Pada Bahasa Indonesia Menggunakan Metode Yake. Universitas Sriwijaya.
- Sun, C., Qiu, X., Xu, Y., & Huang, X. 2022. How to fine-tune BERT for text classification? *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 1–14.
- Sun, K., Lin, Z., Zhu, Z., Zhou, J., Li, J., & Yu, Y. 2022. Confusion matrix based on position-aware graph neural networks for classification performance evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1), 1–14.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., ... & Pavlick, E. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.
- You, J., Ying, R., Ren, X., Hamilton, W. L., & Leskovec, J. 2019. Confusion matrix based on rough set theory for classification performance evaluation. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 7134–7143).
- You, J., Ying, R., Ren, X., Hamilton, W. L., & Leskovec, J. 2019. Position-aware graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 7134–7143)
- Zhang, Z., Liang, X., Zuo, Y., & Lin, C. 2023. Improving unsupervised keyphrase extraction by modeling hierarchical multi-granularity features. *Information processing & Management*, 60(4), 103356.