CLINICAL NAMED ENTITY RECOGNITION MODEL BERBASIS TRANSFORMER UNTUK DATA BIOMEDIS

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH :

INDAH GALA PUTRI

09011182126033

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2025

HALAMAN PENGESAHAN

SKRIPSI

CLINICAL NAMED ENTITY RECOGNITION MODEL BERBASIS TRANSFORMER UNTUK DATA BIOMEDIS

Sebagai salah satu syarat untuk penyelesaian studi di
Program Studi S1 Sistem Komputer

Oleh:

INDAH GALA PUTRI

09011182126033

Pembimbing 1	**:**	**Prof. Dr. Ir. Bambang Tutuko, M.T.**
		NIP. 196001121989031002
Pembimbing 2	**:**	**Dr. Firdaus, S.T., M.Kom.**
		NIP. 197801212008121003

Mengetahui

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T.

196612032006041001

AUTHENTICATION PAGE

SKRIPSI

TRANSFORMER-BASED CLINICAL NAMED ENTITY RECOGNITION MODEL FOR BIOMEDICAL DATA

As one of the requirements for completing studies in
the Bachelor's Degree Program in Computer Systems

By:

INDAH GALA PUTRI

09011182126033

Supervisor 1	**:**	**Prof. Dr. Ir. Bambang Tutuko, M.T.**
		NIP. 196001121989031002
Supervisor 2	**:**	**Dr. Firdaus, S.T., M.Kom.**
		NIP. 197801212008121003

Mengetahui

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T.

196612032006041001

HALAMAN PERSETUJUAN

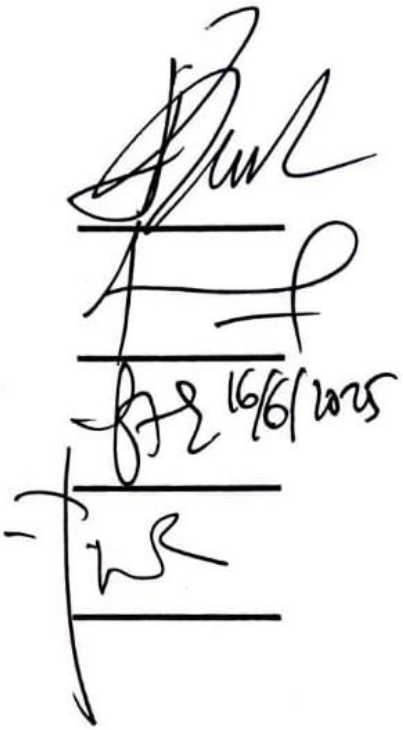
Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 23 Mei 2025

Tim Penguji :

1. Ketua Sidang : Prof. Dr. Erwin, S.Si., M.Si.
2. Penguji Sidang : Dr. M. Fachrurrozi, S.Si., M.T.
3. Pembimbing I : Prof. Dr. Ir. Bambang Tutuko, M.T.
4. Pembimbing II : Dr. Firdaus, S.T., M.Kom.



Mengetahui, 17/6/25

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T.

NIP. 196612032006041001

LEMBAR PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Indah Gala Putri
NIM : 09011182126033
Judul : *Clinical Named Entity Recognition Model* Berbasis Transformer untuk Data Biomedis

Hasil Pengecekan Software Turnitin: 3%

Menyatakan bahwa laporan skripsi saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

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

Indralaya, Juni 2025

INDAH GALA PUTRI
NIM. 09011182126033

HALAMAN PERSEMBAHAN

1. Kepada Papa dan Mama tercinta, Papa Gatot Udiyono dan Mama Harsilah, terima kasih atas setiap do'a yang kalian panjantkan untuk Indah. Do'a kalian adalah cahaya yang tak pernah padam dalam setiap langkah Indah. Selain itu, terima kasih juga atas kasih sayang, pengorbanan, kesabaran, dan segala dukungan yang tak terhitung jumlahnya. Tanpa restu, semangat, dan dukungan dari kalian, Indah tidak akan sampai pada titik ini. Terima kasih sudah menjadi orang tua yang hebat bagi Indah. Kalian selalu menjadi sumber kekuatan dan motivasi terbesar dalam hidup Indah. Indah persembahkan skripsi ini untuk Papa dan Mama sebagai bentuk rasa sayang dan terima kasih atas segala cinta dan perjuangan kalian yang tak ternilai.
2. Kepada adik-adikku yang tersayang, Ine Gala Febrianti dan Irsya Gala Aprilia, yang selalu menjadi penyemangat tersendiri dalam hidupku. Terima kasih telah menjadi *support system* bagiku. Terima kasih atas perhatian, candaan, dan dukungan sederhana yang sering kali datang di saat yang paling aku butuhkan. Kehadiran kalian benar-benar membantuku bertahan dan menyelesaikan proses ini. Skripsi ini aku persembahkan juga untuk kalian, sebagai bentuk rasa sayang dan terima kasihku.

MOTTO

"Maka sesungguhnya bersama kesulitan ada kemudahan. Sesungguhnya bersama kesulitan ada kemudahan."

(Q.S. Al-Insyirah: 5-6)

"Allah tidak membebani seseorang, kecuali menurut kesanggupannya."

(Q.S. Al-Baqarah: 286)

"You cannot fight qadr. If something is written for you, it will be yours. Nothing can stop that. Live your life at peace. Do not fill your heart with worry and anxiety. Your Rabb, the Greatest of the Greatest has written your fate. Have trust in Him."

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Puji dan syukur atas kehadirat Allah SWT yang telah melimpahkan rahmat, kasih sayang dan karunia-Nya, sehingga penulis dapat menyelesaikan Tugas Akhir ini dengan judul "*Clinical Named Entity Recognition Model* Berbasis Transformer untuk Data Biomedis".

Dalam laporan ini, penulis menjelaskan mengenai pemodelan yang digunakan untuk mengimputasi data yang hilang dengan menggunakan data yang penulis temukan saat melakukan penelitian dan pengujian data. Penulis berharap agar tulisan ini dapat bermanfaat bagi orang banyak.

Selama penulisan Tugas Akhir ini, penulis banyak mendapatkan ide, bantuan, serta saran dari semua pihak, baik secara langsung maupun tak langsung. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Allah SWT yang telah melimpahkan berkah serta nikmat kesehatan dan kesempatan kepada penulis sehingga dapat menyelesaikan Tugas Akhir ini.
2. Kedua orang tua, saudara, dan keluarga besar yang telah mendoakan dan memberikan motivasi serta *support*.
3. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Dr. Ir. Sukemi, M.T., selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Prof. Dr. Ir. Bambang Tutuko, M.T. dan Dr. Firdaus, S.T., M.Kom selaku Dosen Pembimbing Tugas Akhir yang telah berkenan meluangkan waktunya guna membimbing, memberikan saran dan motivasi serta bimbingan terbaik untuk penulis dalam menyelesaikan Tugas Akhir ini.
6. Ibu Prof. Dr. Ir. Siti Nurmaini, M.T. selaku *Head of Intelligent System Research Group* (ISysRG) yang telah memberikan kesempatan besar untuk bergabung dan menjadi bagian dari *team research group* ini.
7. Bapak Prof. Dr. Ir. Bambang Tutuko, M.T selaku Dosen Pembimbing Akademik di Jurusan Sistem Komputer Universitas Sriwijaya

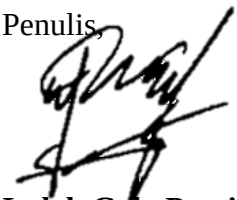
8. Mbak Anggun Islami, M.Kom., selaku mentor divisi *text* ISysRG yang telah memberikan bantuan dan masukan dalam penelitian skripsi ini.
9. Ibu Dr. Ade Iriani Sapitri, M.Kom, Bapak Naufal Rachmatullah, M.T., Ibu Akhiar Wista Arum, S.T., M.Kom., dan Ibu Dr. Annisa Darmawahyuni, M.Kom. selaku mentor di ISysRG yang telah memberikan bantuan dan masukan dalam penelitian skripsi ini.
10. Staff administrasi Fakultas Ilmu Komputer Universitas Sriwijaya Jurusan Sistem Komputer yang telah memberikan kemudahan dalam administrasi sehingga penulis dapat membuat tugas akhir ini dengan lancar.
11. Teman-teman seperjuangan di Jurusan Sistem Komputer Angkatan 2021, terutama sahabat saya Keisyah Sabinatullah Qur'aini, Mutiah Andini dan Zahra Hanifa yang selalu memberikan motivasi dan menjadi *support system* saya.
12. Teman-teman divisi *text processing* di ISysRG yaitu Muhammad Azriel Apriadi, Tiara Oktarina, dan Tria Lailani atas segala dukungan, kerja sama, serta diskusi-diskusi yang membangun selama proses penelitian ini berlangsung.
13. Semua pihak yang telah membantu dalam penyusunan skripsi ini, yang tidak dapat saya sebutkan satu per satu.
14. Almamater

Penulis menyadari bahwa laporan ini masih sangat jauh dari kata sempurna. Untuk itu kritik dan saran yang membangun sangatlah diharapkan penulis. Akhir kata penulis berharap, semoga proposal tugas akhir ini bermanfaat dan berguna bagi khalayak.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, Juni 2025

Penulis,



Indah Gala Putri

NIM. 09011182126033

CLINICAL NAMED ENTITY RECOGNITION MODEL BERBASIS TRANSFORMER UNTUK DATA BIOMEDIS

INDAH GALA PUTRI (09011182126033)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : indahgalaputri@gmail.com

ABSTRAK

Clinical Named Entity Recognition (CNER) merupakan tugas penting dalam pemrosesan bahasa alami untuk mengekstraksi entitas medis dari teks biomedis yang kompleks. Tantangan utama dalam tugas ini terletak pada kompleksitas struktur kalimat dan terminologi medis yang sangat bervariasi. Penelitian ini berfokus pada pengembangan dan evaluasi model CNER berbasis arsitektur Transformer, khususnya BERT, untuk meningkatkan pemahaman dan akurasi dalam mengenali entitas medis dari data biomedis. Penelitian ini mengembangkan dua model berbasis BERT, yaitu EMR-BERT dan PubMed2M-BERT. EMR-BERT merupakan model hasil kustomisasi arsitektur BERT dengan delapan lapisan *encoder* dan dilatih langsung melalui *fine-tuning*. Sementara itu, PubMed2M-BERT adalah hasil *pre-training* lanjutan dari BERT-Base Uncased menggunakan *Masked Language Modeling* (MLM) tanpa *Next Sentence Prediction* (NSP) pada korpus biomedis ViPubMed. Hasil *pre-training* menunjukkan nilai *perplexity* sebesar 2,964 dan kurva *loss* yang stabil. Pada tahap *fine-tuning*, PubMed2M-BERT mencapai F1-*score* tertinggi sebesar 92% pada dataset NCBI-*disease*, mengungguli EMR-BERT yang memperoleh 87%. Temuan ini membuktikan bahwa *pre-training* domain spesifik mampu meningkatkan performa model Transformer dalam tugas CNER pada data biomedis.

Kata Kunci: *Clinical Named Entity Recognition*, Transformer, *Pre-Training*, *Fine- Tuning*, Biomedis.

TRANSFORMER-BASED CLINICAL NAMED ENTITY RECOGNITION MODEL FOR BIOMEDICAL DATA

INDAH GALA PUTRI (09011182126033)

Computer System Department, Computer Science Faculty, Sriwijaya University

Email : indahgalaputri@gmail.com

ABSTRACT

Clinical Named Entity Recognition (CNER) is a critical task in natural language processing (NLP) aimed at extracting medical entities from complex biomedical texts. The main challenges in this task lie in the complexity of sentence structures and the highly variable medical terminology. This study focuses on the development and evaluation of CNER models based on the Transformer architecture, specifically BERT, to improve understanding and accuracy in recognizing medical entities from biomedical data. Two BERT-Base models were developed in this research: EMR-BERT and PubMed2M-BERT. EMR-BERT is a customized model with eight encoder layers trained directly through fine-tuning. In contrast, PubMed2M-BERT is a continuation pre-training of BERT-Base Uncased using the Masked Language Modeling (MLM) objective without Next Sentence Prediction (NSP) on the ViPubMed biomedical corpus. The pre-training results showed a perplexity score of 2.964 and a stable loss curve. During the fine-tuning phase, PubMed2M-BERT achieved the highest F1-score of 92% on the NCBI-disease dataset, outperforming EMR-BERT, which achieved 87%. These findings demonstrate that domain-specific pre-training can significantly enhance the performance of Transformer models in CNER tasks on biomedical data.

Keywords: *Clinical Named Entity Recognition, Transformer, Pre-Training, Fine-Tuning, Biomedical.*

DAFTAR ISI

HALAMAN JUDUL ..... **i**
HALAMAN PENGESAHAN ..... **ii**
AUTHENTICATION PAGE ..... **iii**
HALAMAN PERSETUJUAN ..... **iv**
LEMBAR PERNYATAAN ..... **v**
HALAMAN PERSEMBAHAN ..... **vi**
KATA PENGANTAR ..... **vii**
ABSTRAK ..... **ix**
ABSTRACT ..... **x**
DAFTAR ISI ..... **xi**
DAFTAR GAMBAR ..... **xiv**
DAFTAR TABEL ..... **xvii**
BAB I PENDAHULUAN **1**
1.1. Latar Belakang 1
1.2. Rumusan Masalah 2
1.3. Batasan Masalah 3
1.4. Tujuan 3
1.5. Metodologi Penelitian 4
1.5.1. Metode Studi Pustaka dan Literatur 4
1.5.2. Metode Konsultasi 4
1.5.3. Metode Pembuatan Model 4
1.5.4. Metode Pengujian 4
1.5.5. Metode Analisa dan Kesimpulan 5
1.6. Sistematika Penulisan 5
BAB II TINJAUAN PUSTAKA **7**
2.1. Penelitian Terdahulu 7
2.2. *Natural Language Processing* 8
2.3. *Text Processing* 9
2.3.1. Normalisasi Teks 9
2.3.2. Tokenisasi Teks 10
2.3.3. *Word Indexing* 11
2.3.4. *Sequence Labeling* 11
2.3.5. *Padding* 12

2.4. *Clinical Named Entity Recognition* 12
2.5. *Deep Learning* 13
2.6. Transformer 13
2.7. *Self-Attention Mechanism* 14
2.8. *Multi-Head Attention* 16
2.9. *Bidirectional Encoder Representations from Transformers* 17
2.10. *Pre-Training* pada BERT 19
2.10.1. *Masked Language Model* 19
2.10.2. *Next Sentence Prediction* 20
2.11. *Fine-Tuning* 21
2.12. Dataset Biomedis 22
2.12.1. PubMed *Abstract* 22
2.12.2. *BioCreative* II *Gene Mention* 23
2.12.3. *Joint Workshop on Natural Language Processing in Biomedicine and its Applications* 23
2.12.4. *National Center for Biotechnology Information* 23
2.13. Pengukuran Kinerja 24
2.13.1. *Perplexity* 24
2.13.2. *Confusion Matrix* 24
2.13.3. *Recall* 25
2.13.4. *Precision* 25
2.13.5. F1-*Score* 25
BAB III METODOLOGI PENELITIAN 27
3.1. Kerangka Kerja 27
3.2. Akuisisi Data 29
3.3. *Data Splitting* 30
3.4. *Eploratory Data Analysis* 32
3.4.1. EDA Pada Tahap *Pre-Training* 32
3.4.2. EDA Pada Tahap *Fine-Tuning* 34
3.5. *Data Preprocessing* 39
3.5.1. *Preprocessing* pada Tahap *Pre-Training* 40
3.5.2. *Preprocessing* pada Tahap *Fine-Tuning* 44
3.6. Pelatihan Model 46
3.6.1. EMR-BERT 47
3.6.2. PubMed2M-BERT 51

3.7. Evaluasi Model 60
BAB IV HASIL DAN ANALISIS 62
4.1. Skenario Percobaan 62
4.2. Hasil Pengujian pada Dataset BC2GM 62
4.3. Hasil Pengujian pada Dataset JNLPBA 67
4.4. Hasil Pengujian pada Dataset NCBI-*Disease* 76
4.5. *Learning Curve* pada EMR-BERT 81
4.6. *Macro Average* Hasil Pengujian pada Data Latih, Validasi, dan Uji 83
4.7. *Confusion Matrix* 89
4.7.1. *Confusion Matrix* pada Dataset BC2GM 89
4.7.2. *Confusion Matrix* pada Dataset JNLPBA 91
4.7.3. *Confusion Matrix* pada Dataset NCBI-*Disease* 93
4.8. Rangkuman Waktu Pelatihan 95
4.9. Penambahan Model PubMed2M-BERT 96
4.9.1. Skenario Percobaan 96
4.9.2. Hasil Pengujian pada Tahap *Pre-Training* 97
4.9.3. Hasil Pengujian pada Tahap *Fine-Tuning* 98
4.9.4. *Loss Curve* pada Tahap *Pre-Training* 113
4.9.5. *Macro Average* Hasi Pengujian pada Tahap *Fine-tuning* 114
4.9.6. *Confusion matrix* pada Tahap *Fine-Tuning* 117
4.9.7. Rangkuman Waktu Pelatihan 121
4.10. Rangkuman Hasil Pengujian Terbaik pada EMR-BERT dan PubMed2M-BERT 122
BAB V KESIMPULAN DAN SARAN 125
5.1. Kesimpulan 125
5.2. Saran 126
DAFTAR PUSTAKA 127
LAMPIRAN 132

DAFTAR GAMBAR

Gambar 2.1. Ilustrasi Normalisasi Teks. 10
Gambar 2.2. Ilustrasi Tokenisasi Teks. 10
Gambar 2.3. Arsitektur Transformer 14
Gambar 2.4. *Scaled Dot Product Attention*. 15
Gambar 2.5. *Multi-Head Attention* 17
Gambar 2.6. Arsitektur BERT. 18
Gambar 2.7. Ilustrasi MLM pada BERT 20
Gambar 2.8. Ilustrasi MLM pada BERT 21
Gambar 3.1 Kerangka Kerja. 29
Gambar 3.2. Visualisasi Dataset ViPubMed. 33
Gambar 3.3. Visualisasi Panjang Data ViPubMed. 34
Gambar 3. 4. Sampel Dataset Biomedis. 35
Gambar 3.5. Perbandingan Distribusi Label Entitas pada Dataset Biomedis. 37
Gambar 3.6. *Wordcloud* Label Entitas pada Dataset Biomedis. 38
Gambar 3.7. Perbandingan Panjang Data Tiap Dataset. 39
Gambar 3.8. *Flowchart Data Pre-processing.* 40
Gambar 3.9. Ilustrasi *Wordpiece Tokenizer.* 41
Gambar 3.10. Ilustrasi *Word Indexing.* 42
Gambar 3.11. Ilustrasi *Group Token.* 43
Gambar 3.12. Ilustrasi *Padding.* 43
Gambar 3.13. Ilustrasi Proses Normalisasi Teks. 44
Gambar 3.14. Ilustrasi *Padding* dengan Nilai -100. 46
Gambar 3.15. Arsitektur EMR-BERT. 47
Gambar 3.16. Skema Proses Pelatihan pada EMR-BERT. 49
Gambar 3.17. Arsitektur PubMed2M-BERT. 51
Gambar 3.18. Skema Proses *Pre-Training* PubMed2M-BERT. 52
Gambar 3.19. Ilustrasi Proses *Embedding* pada BERT. 53
Gambar 3.20. Arsitektur *Encoder* BERT. 54
Gambar 3.21. Skema Proses *Fine-Tuning* PubMed2M-BERT. 59
Gambar 4.1. *Bar Chart* Hasil Pengujian pada Dataset BC2GM dengan Metrik *Precision.* 63
Gambar 4.2. *Bar Chart* Hasil Pengujian pada Dataset BC2GM dengan Metrik *Recall.* 65

Gambar 4.3. *Bar Chart* Hasil Pengujian pada Dataset BC2GM dengan Metrik F1-*Score.*.....66
Gambar 4.4. *Bar Chart* Hasil Pengujian pada Dataset JNLPBA dengan Metrik *Precision.*.....69
Gambar 4.5. *Bar Chart* Hasil Pengujian pada Dataset JNLPBA dengan Metrik *Recall.*..72
Gambar 4.6. *Bar Chart* Hasil Pengujian pada Dataset JNLPBA dengan Metrik F1-*Score.*.....75
Gambar 4.7. *Bar Chart* Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik *Precision.*.....77
Gambar 4.8. *Bar Chart* Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik *Recall.*.....79
Gambar 4.9. *Bar Chart* Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik F1-*Score.*.....80
Gambar 4. 10. *Learning Curve* EMR-BERT untuk Hasil Pengujian Terbaik pada Tiap Dataset..... 82
Gambar 4.11. Grafik Perbandingan *Macro Average* Model EMR-BERT Dengan *Batch Size* 16.....85
Gambar 4.12. Grafik Perbandingan *Macro Average* Model EMR-BERT dengan *Batch Size* 32.....88
Gambar 4.13. *Confusion Matrix* EMR-BERT pada Dataset BC2GM 90
Gambar 4.14. *Confusion Matrix* EMR-BERT pada Dataset JNLPBA 92
Gambar 4.15. *Confusion Matrix* EMR-BERT pada Dataset NCBI-*Disease*.....94
Gambar 4.16. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik *Precision.*.....99
Gambar 4.17. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik *Recall.*.....100
Gambar 4.18. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik F1-*Score.*.....102
Gambar 4.19. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik *Precision.*.....104
Gambar 4.20. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik *Recall.*.....106
Gambar 4.21. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik F1-*Score.*.....108

Gambar 4.22. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset NCBI-*Disease* dengan Metrik *Precision.* 110
Gambar 4.23. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset NCBI-*Disease* dengan Metrik *Recall.* 111
Gambar 4.24. *Bar Chart* Hasil Pengujian PubMed2M-BERT pada Dataset NCBI-*Disease* dengan Metrik F1-*Score.*.....113
Gambar 4.25. *Loss Curve* pada Tahap *Pre-Training*.....114
Gambar 4.26. Grafik Perbandingan *Macro Average* Model PubMed2M-BERT. 116
Gambar 4.27. *Confusion Matrix* pada PubMed2M-BERT untuk Dataset BC2GM 118
Gambar 4.28. *Confusion Matrix* pada PubMed2M-BERT untuk Dataset JNLPBA 119
Gambar 4.29. *Confusion Matrix* pada PubMed2M-BERT untuk Dataset NCBI-*Disease*.....120
Gambar 4.30. *Bar Chart* Rangkuman Perbandingan Hasil Terbaik pada EMR-BERT dan PubMed2M-BERT.....123

DAFTAR TABEL

Tabel 2.1. *Confusion Matrix* 25
Tabel 3.1. Persentase dan Persebaran Jumlah Data Latih dan Uji pada Dataset ViPubMed.....31
Tabel 3.2. Persentase dan Persebaran Jumlah Data Latih, Validasi, dan Uji pada Dataset BC2GM, JNLPBA, dan NCBI- *Disease* 31
Tabel 3.3. Keterangan Label Entitas pada Dataset 35
Tabel 3.4. *Spesifikasi* Arsitektur EMR-BERT.....49
Tabel 3.5. *Hyperparameter* Pelatihan Model EMR-BERT 51
Tabel 3.6. Spesifikasi Arsitektur PubMed2M-BERT 56
Tabel 3.7. *Hyperparameter* Pelatihan model PubMed2M-BERT pada Tahap *Pre-Training* 57
Tabel 3.8. *Hyperparameter* Pelatihan Model PubMed2M-BERT pada Tahap *Fine-Tuning*.....60
Tabel 4.1. Skenario Percobaan.....62
Tabel 4.2. Hasil Pengujian pada Dataset BC2GM dengan Metrik *Precision*.....63
Tabel 4.3. Hasil Pengujian Dataset BC2GM dengan Metrik *Recall* 64
Tabel 4.4. Hasil Pengujian pada Dataset BC2GM dengan Metrik F1-*Score* 66
Tabel 4.5. Hasil Pengujian pada Dataset JNLPBA dengan Metrik *Precision*.....68
Tabel 4.6. Hasil Pengujian pada Dataset JNLPBA dengan Metrik *Recall*.....70
Tabel 4.7. Hasil Pengujian pada Dataset JNLPBA dengan Metrik F1-*Score* 73
Tabel 4.8. Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik *Precision* 77
Tabel 4.9. Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik *Recall* 78
Tabel 4.10. Hasil Pengujian pada Dataset NCBI-*Disease* dengan Metrik F1-*Score*.....80
Tabel 4.11. *Macro Average* pada Data Latih, Validasi, dan Uji dengan *Batch Size* 16 84
Tabel 4.12. *Macro Average* pada Data Latih, Validasi, dan Uji dengan *Batch Size* 32 86
Tabel 4.13. Rangkuman Waktu *Training* pada Model EMR-BERT.....95
Tabel 4.14. Skenario Percobaan *Pre-Training* 96
Tabel 4. 15. Skenario Percobaan *Fine-Tuning* 96
Tabel 4.16. Hasil Pengujian pada Tahap *Pre-Training* dengan Data Latih 97
Tabel 4.17. Hasil Pengujian pada Tahap *Pre-Training* dengan Data Uji.....97
Tabel 4.18. Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik *Precision* 99

Tabel 4.19. Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik *Recall* 100
Tabel 4.20. Hasil Pengujian PubMed2M-BERT pada Dataset BC2GM dengan Metrik F1-*Score* 101
Tabel 4.21. Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik *Precision* 103
Tabel 4.22. Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik *Recall* 105
Tabel 4.23. Hasil Pengujian PubMed2M-BERT pada Dataset JNLPBA dengan Metrik F1-*Score* 107
Tabel 4.24. Hasil Pengujian PubMed2M-BERT pada Dataset NCBI-*Disease* dengan Metrik *Precision* 110
Tabel 4.25. Hasil Pengujian PubMed2M-BERT pada Dataset NCBI-*Disease* dengan Metrik *Recall* 111
Tabel 4.26. Hasil Pengujian PubMed2M-BERT pada Dataset Ncbi-*Disease* dengan Metrik F1-*Score* 112
Tabel 4.27. Rangkuman *Macro Average* Data Latih, Validasi, dan Uji pada PubMed2M-BERT 115
Tabel 4.28. Rangkuman Waktu Pelatihan pada Tahap *Pre-Training* dan *Fine-Tuning* ..121
Tabel 4.29. Rangkuman Perbandingan Hasil Terbaik pada EMR-BERT dan PubMed2M-BERT 122

BAB I
PENDAHULUAN

1.1. Latar Belakang

Pertumbuhan informasi tekstual dalam jumlah besar semakin meningkat dalam beberapa tahun terakhir, khususnya di bidang biomedis. Informasi ini dapat berupa literatur biomedis, catatan klinis, atau laporan dalam rekam medis elektronik (RME), sementara data tekstual seperti literatur biomedis merujuk pada data teks yang ditemukan dalam basis data artikel ilmiah dan RME mencakup catatan elektronik mengenai informasi kesehatan pasien [1],[2],[3]. Salah satu fungsi utama teks biomedis adalah sebagai sumber informasi yang kaya akan pengetahuan medis, termasuk gejala penyakit, diagnosis, pengobatan, interaksi obat, serta hasil penelitian ilmiah [1]. Akan tetapi, informasi yang disajikan dalam narasi klinis sering kali tidak terstruktur, panjang, rumit, dan penuh istilah teknis [4]. Oleh karena itu, teknologi dan sistem *Natural Language Preprocessing* (NLP) standar tidak dapat langsung diterapkan pada domain klinis [5]. Hal ini tentu menyulitkan proses analisis dan ekstraksi informasi secara otomatis. Maka dari itu, pada penelitian kali ini memiliki fokus utamanya adalah pada tugas pengenalan entitas bernama atau *Named Entity Recognition* (NER) pada dokumen medis [6].

NER adalah metode untuk mengidentifikasi, mengklasifikasikan, dan memisahkan entitas bernama ke dalam kelompok sesuai dengan kategori yang telah ditentukan seperti nama-nama orang, lokasi, dan organisasi [7], [8]. Dalam domain medis, NER memainkan peran penting dengan mengekstraksi terminologi medis yaitu, segmen teks yang bermakna, seperti penyakit, gejala, obat-obatan, dan lain-lain [9], [10]. Dengan NER informasi penting pada teks tersebut akan diatur dan dikodekan ke dalam format yang bisa dihitung oleh komputer menggunakan istilah yang sudah ditentukan, sehingga bisa disaring secara otomatis oleh komputer[11].

Transformer adalah salah satu arsitektur *neural network* yang paling umum digunakan dalam NLP. Baru-baru ini, model berbasis Transformer telah menunjukkan kinerja yang sangat baik dalam berbagai tugas NLP, termasuk NER

[12]. Arsitektur ini pertama kali diperkenalkan oleh tim peneliti Google untuk masalah penerjemahan Inggris-Jerman dan Inggris-Prancis. Perkembangan penelitian tentang topik NLP telah mencapai tahap yang signifikan setelah kehadiran arsitektur *deep learning* berbasis *attention* yang disebut Transformer pada tahun 2017. Adapun model Transformer ini sendiri memiliki keunggulan dalam memahami dan memodelkan bahasa [13],[14]. Namun, penerapan langsung metodologi NLP mutakhir pada pengenalan entitas biomedis memiliki keterbatasan. Pertama, karena model representasi kata terbaru seperti Word2Vec, ELMo, dan BERT dilatih dan diuji terutama pada dataset yang berisi teks dari domain umum, misalnya Wikipedia. Maka sulit untuk memperkirakan kinerja mereka pada dataset yang berisi teks biomedis. Selain itu, distribusi kata dalam korpus umum dan korpus biomedis sangat berbeda dan jarang ditemui pada korpus umum, yang sering kali menjadi masalah bagi model [3]. Sehingga penggunaan *transfer learning* seperti BERT perlu diperhatikan karena pnggunaan yang tidak tepat dapat menyebabkan *negative transfer*, yaitu penurunan performa jika pengetahuan dari domain sumber tidak sesuai dengan domain target [4].

Berdasarkan dari latar belakang yang diuraikan sebelumnya, penulis akan mengembangkan model berbasis Transformer yang dilatih dengan data biomedis untuk melakukan *Clinical Named Entity Recognition* (CNER). Melalui pengembangan sistem ini, diharapkan model yang dihasilkan mampu membantu dalam memecahkan berbagai permasalahan terkait CNER pada data biomedis.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang diuraikan di atas maka rumusan masalah dari tugas akhir ini diantaranya:

1. Bagaimana mengembangkan dan mengadaptasi model berbasis BERT agar mampu memahami karakteristik data biomedis yang kompleks dalam konteks tugas CNER?
2. Bagaimana pengaruh penggunaan arsitektur dan konfigurasi pelatihan, terhadap performa model Transformer dalam tugas CNER pada berbagai dataset biomedis?

3. Bagaimana perbandingan kinerja model dalam mengenali dan mengekstraksi entitas klinis dari teks biomedis berdasarkan metrik evaluasi seperti *precision*, *recall*, dan F1-*score*?

1.3. Batasan Masalah

Berdasarkan latar belakang yang diuraikan di atas maka batasan masalah dari tugas akhir ini diantaranya:

1. Model yang digunakan dalam penelitian ini adalah model Transformer berbasis *Bidirectional Encoder Representations from Transformer*s (BERT), dengan kustomisasi lapisan *encoder* yang dibatasi sebanyak 8 lapisan *encoder* tanpa dilakukan *pre-training* lanjutan untuk EMR-BERT.
2. Pada model PubMed2M-BERT tahap *pre-training* model dilatih hanya dengan tugas *Mask Language Model* (MLM) tanpa menggunakan tugas *Next Sentence Prediction* (NSP), sehingga fokus penelitian lebih terarah pada tugas CNER.
3. Data yang digunakan untuk melakukan *pretraining* dalam penelitian ini adalah dataset ViPubMed dengan hanya mengambil 2 juta sampel teks berbahasa Inggris. Hal ini bertujuan untuk mempertimbangkan keterbatasan perangkat keras yang digunakan, seperti kapasitas memori GPU dan efisiensi waktu pelatihan.
4. Data yang digunakan untuk melakukan *fine-tuning* terbatas pada dataset biomedis yang bersumber dari tiga dataset, yaitu BC2GM, NCBI, dan JNLPBA dengan format yang telah disesuaikan untuk mendukung tugas pengenalan entitas klinis.
5. Evaluasi model dalam penelitian ini dibatasi pada tiga metrik utama, yaitu *precision*, *recall*, dan F1-*score*.

1.4. Tujuan

Berdasarkan dari latar belakang di atas maka tujuan dari tugas akhir ini diantaranya:

1. Mengembangkan model *deep learning* berbasis transformer untuk melakukan CNER pada data biomedis.

2. Mengevaluasi kinerja model yang dibangun menggunakan metrik evaluasi yang sesuai, seperti *precision, recall*, dan F1-*score,* untuk mengukur tingkat akurasi dan efektivitas model dalam mengenali dan mengekstraksi entitas klinis.

1.5. Metodologi Penelitian

Penelitian ini menggunakan pendekatan metodologi yang sistematis untuk menyelesaikan permasalahan dalam pengenalan entitas klinis. Metodologi yang digunakan dalam penelitian tugas akhir ini adalah:

1.5.1. Metode Studi Pustaka dan Literatur

Metode ini dilakukan dengan mencari dan mengumpulkan referensi dari literatur, khususnya jurnal-jurnal terpercaya, yang berkaitan dengan "*Clinical Named Entity Recognition* Model Berbasis Transformer".

1.5.2. Metode Konsultasi

Metode ini melibatkan konsultasi dengan pihak-pihak yang memiliki pengetahuan serta wawasan yang baik dalam mengatasi permasalahan yang ditemui pada penulisan tugas akhir "*Clinical Named Entity Recognition* Model Berbasis Transformer".

1.5.3. Metode Pembuatan Model

Pembuatan model dalam penelitian ini dilakukan dengan mengadaptasi arsitektur BERT yang disesuaikan untuk tugas NER pada domain biomedis. Proses ini mencakup tahap *pre-training* dan dilatih dengan dataset yang tidak berlabel dan *fine-tuning model* menggunakan dataset anotasi khusus. Adapaun model dibangun dengan menggunakan simulasi pemrograman bahasa *python*.

1.5.4. Metode Pengujian

Metode ini dilakukan dengan menjalankan dan melakukan pengujian terhadap simulasi yang sudah dibuat apakah telah menghasilkan nilai F1-*score* yang baik atau tidak.

1.5.5. Metode Analisa dan Kesimpulan

Analisis dilakukan dengan membandingkan performa model berdasarkan hasil evaluasi terhadap data uji menggunakan metrik seperti *precision*, *recall*, dan F1-*score*. Hasil dari setiap eksperimen dibandingkan untuk menilai pengaruh konfigurasi model terhadap akurasi identifikasi entitas. Dengan cara ini, dapat diketahui konfigurasi mana yang memberikan hasil paling optimal untuk tugas NER.

1.6. Sistematika Penulisan

Agar penyusunan tugas akhir ini dapat disampaikan secara sistematis dan mudah dipahami, maka penulisan dibagi ke dalam beberapa bab sebagai berikut:

BAB I PENDAHULUAN

Bab I memberikan uraian tentang awal dari suatu penulisan, meliputi latar belakang, perumusan dan batasan masalah, tujuan dan manfaat, metodologi penelitian, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab II memaparkan mengenai teori – teori dasar yang menjadi landasan dari penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Bab III berisi penjelasan detail mengenai teknik, metode, serta alur proses yang digunakan dalam penelitian.

BAB IV HASIL DAN ANALISIS

Bab IV menjelaskan hasil pengujian yang diperoleh dan menjelaskan analisa terhadap hasil penelitian yang telah dilakukan.

BAB V KESIMPULAN

Bab V berisi kesimpulan dari hasil dan analisa dari keseluruhan penelitian yang dilakukan.

DAFTAR PUSTAKA

Daftar pustaka berisi daftar refrensi dari sumber – sumber informasi yang digunakan dalam metode literatur.

LAMPIRAN

Lampiran mencakup formulir perbaikan dan juga pemeriksaan tingkat kemiripan karya dengan sumber lain.

DAFTAR PUSTAKA

[1] A. Chaves, C. Kesiku, and B. Garcia-Zapirain, "Automatic Text Summarization of Biomedical Text Data: A Systematic Review," *Inf.*, vol. 13, no. 8, 2022, doi: 10.3390/info13080393.

[2] C. Y. Y. Kesiku, A. Chaves-Villota, and B. Garcia-Zapirain, "Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review," *Inf.*, vol. 13, no. 10, 2022, doi: 10.3390/info13100499.

[3] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[4] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on Named Entity Recognition — datasets, tools, and methodologies," *Nat. Lang. Process. J.*, vol. 3, no. 10, p. 100017, 2023, doi: 10.1016/j.nlp.2023.100017.

[5] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A comparative analysis of active learning for biomedical text mining," *Appl. Syst. Innov.*, vol. 4, no. 1, 2021, doi: 10.3390/asi4010023.

[6] M. Polignano, M. de Gemmis, and G. Semeraro, "Comparing transformer-based NER approaches for analysing textual medical diagnoses," *CEUR Workshop Proc.*, vol. 2936, pp. 818–833, 2021.

[7] M. B. Shishehgarkhaneh, R. C. Moehler, Y. Fang, A. A. Hijazi, and H. Aboutorab, "Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia," *IEEE Access*, vol. 12, no. March, pp. 41829–41851, 2024, doi: 10.1109/ACCESS.2024.3377232.

[8] Y. Zhang and G. Xiao, "Named Entity Recognition Datasets: A Classification Framework," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00456-1.

[9] O. Solarte-Pabón *et al.*, "Transformers for extracting breast cancer information from Spanish clinical narratives," *Artif. Intell. Med.*, vol. 143, no. May, 2023, doi: 10.1016/j.artmed.2023.102625.

[10] A. Villaplana, R. Martínez, and S. Montalvo, "Improving Medical Entity Recognition in Spanish by Means of Biomedical Language Models," *Electron.*, vol. 12, no. 23, pp. 1–12, 2023, doi: 10.3390/electronics12234872.

[11] S. Tian *et al., Transformer-based named entity recognition for parsing clinical trial eligibility criteria*, vol. 1, no. 1. Association for Computing Machinery, 2021.

doi: 10.1145/3459930.3469560.

[12] E. C. Jibril and A. C. Tantug, “ANEC: An Amharic Named Entity Corpus and Transformer Based Recognizer,” *IEEE Access*, vol. 11, no. February, pp. 15799–15815, 2023, doi: 10.1109/ACCESS.2023.3243468.

[13] S. Silalahi, T. Ahmad, and H. Studiawan, “Transformer-Based Named Entity Recognition on Drone Flight Logs to *Support* Forensic Investigation,” *IEEE Access*, vol. 11, no. December 2022, pp. 3257–3274, 2023, doi: 10.1109/ACCESS.2023.3234605.

[14] M. Al-Yahya, H. Al-Khalifa, H. Al-Baity, D. Alsaeed, and A. Essam, “Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches,” *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5516945.

[15] T. ValizadehAslani *et al.*, “PharmBERT: a domain-specific BERT model for drug labels,” *Brief. Bioinform.*, vol. 24, no. 4, pp. 1–10, 2023, doi: 10.1093/bib/bbad226.

[16] M. A. Rahman, S. M. Preum, R. D. Williams, H. Alemzadeh, and J. Stankovic, “EMS-BERT: A Pre-Trained Language Representation Model for the Emergency Medical Services (EMS) Domain,” *Proc. - 2023 IEEE/ACM Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2023*, pp. 34–43, 2023, doi: 10.1145/3580252.3586978.

[17] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, “BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework,” *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–19, 2022, doi: 10.1186/s12859-022-05051-9.

[18] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, “Natural Language Processing Advancements By Deep Learning: A Survey,” pp. 1–23, 2020, [Online]. Available: http://arxiv.org/abs/2003.01200

[19] V. Sorin, Y. Barash, E. Konen, and E. Klang, “Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review,” *J. Am. Coll. Radiol.*, vol. 17, no. 5, pp. 639–648, 2020, doi: 10.1016/j.jacr.2019.12.026.

[20] E. T. R. Schneider *et al.*, “BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition,” pp. 65–72, 2020, doi: 10.18653/v1/2020.clinicalnlp-1.7.

[21] I. S. Central and T. O. All, “Natural Language Processing CFG-Based Unification Grammars,” vol. 253.

[22] J. Zhang *et al.*, "A hybrid text normalization system using multi-head self-attention for mandarin," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 6694–6698, 2020, doi: 10.1109/ICASSP40776.2020.9054695.

[23] A. Javaloy and G. García-Mateos, "Text normalization using encoder-decoder networks based on the causal feature extractor," *Appl. Sci.*, vol. 10, no. 13, 2020, doi: 10.3390/app10134551.

[24] A. Tabassum and R. R. Patil, "A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing," *Int. Res. J. Eng. Technol.*, no. June, pp. 4864–4867, 2020, [Online]. Available: www.irjet.net

[25] G. Kim, J. Son, J. Kim, H. Lee, and H. Lim, "Enhancing Korean Named Entity Recognition with Linguistic Tokenization Strategies," *IEEE Access*, vol. 9, no. Mlm, pp. 151814–151823, 2021, doi: 10.1109/ACCESS.2021.3126882.

[26] I. Hashem, M. Islam, S. M. Haque, Z. I. Jabed, and N. Sakib, "A Proposed Technique for Simultaneously Detecting DDoS and SQL Injection Attacks," *Int. J. Comput. Appl.*, vol. 183, no. 11, pp. 50–57, 2021, doi: 10.5120/ijca2021921428.

[27] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," *IEEE Access*, vol. 11, no. March, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[28] P. Meesad, "Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning," *SN Comput. Sci.*, vol. 2, no. 6, pp. 1–17, 2021, doi: 10.1007/s42979-021-00775-6.

[29] T. Kato, K. Abe, H. Ouchi, S. Miyawaki, J. Suzuki, and K. Inui, "Embeddings of label components for sequence labeling: A case study of fine-grained named entity recognition," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 222–229, 2020, doi: 10.18653/v1/2020.acl-srw.30.

[30] S. M. Jain, *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. 2022. doi: 10.1007/978-1-4842-8844-3.

[31] A. Asroni, K. R. Ku-Mahamud, C. Damarjati, and H. B. Slamat, "Arabic speech classification method based on padding and deep learning neural network," *Baghdad Sci. J.*, vol. 18, no. 2, pp. 925–936, 2021, doi: 10.21123/bsj.2021.18.2(Suppl.).0925.

[32] V. Tunali, "Improved Prioritization of Software Development Demands in Turkish With Deep Learning-Based NLP," *IEEE Access*, vol. 10, pp. 40249–40263, 2022, doi: 10.1109/ACCESS.2022.3167269.

[33] M. Al-Qurishi and R. Souissi, “Arabic Named Entity Recognition Using Transformer-based-CRF Model,” *ICNLSP 2021 - Proc. 4th Int. Conf. Nat. Lang. Speech Process.*, pp. 262–271, 2021.

[34] M. Cho, J. Ha, C. Park, and S. Park, “Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition,” *J. Biomed. Inform.*, vol. 103, no. February 2019, p. 103381, 2020, doi: 10.1016/j.jbi.2020.103381.

[35] C. Shorten, T. M. Khoshgoftaar, and B. Furht, “Deep Learning applications for COVID-19,” *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00392-9.

[36] M. Raparthi *et al.*, “1 Semi Annual Edition,” vol. 1, no. 1, pp. 1–9.

[37] S. Islam *et al.*, “A comprehensive survey on applications of transformers for deep learning tasks,” *Expert Syst. Appl.*, vol. 241, 2024, doi: 10.1016/j.eswa.2023.122666.

[38] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, “AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing,” pp. 1–42, 2021, [Online]. Available: http://arxiv.org/abs/2108.05542

[39] D. Nozza, F. Bianchi, and D. Hovy, “What the [MASK]? Making Sense of Language-Specific BERT Models,” 2020, [Online]. Available: http://arxiv.org/abs/2003.02912

[40] J. Wang *et al.*, “Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges,” *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–33, 2024, doi: 10.1145/3648471.

[41] M. V. Koroteev, “BERT: A Review of Applications in Natural Language Processing and Understanding,” 2021, [Online]. Available: http://arxiv.org/abs/2103.11943

[42] Y. M. Kim and T. H. Lee, “Korean clinical entity recognition from diagnosis text using BERT,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. Suppl 7, pp. 1–10, 2020, doi: 10.1186/s12911-020-01241-8.

[43] K. W. Church, Z. Chen, and Y. Ma, “Emerging trends: A gentle introduction to fine-tuning,” *Nat. Lang. Eng.*, vol. 27, no. 6, pp. 763–778, 2021, doi: 10.1017/S1351324921000322.

[44] Z. Qiu, X. Wu, J. Gao, and W. Fan, “U-Bert,” *Aaai.Org*, 2021, [Online]. Available: www.aaai.org

[45] L. Kryeziu and V. Shehu, “Pre-Training MLM Using Bert for the Albanian Language,” *SEEU Rev.*, vol. 18, no. 1, pp. 52–62, 2023, doi: 10.2478/seeur-2023-0035.

[46] A. H. Mohammed and A. H. Ali, “Survey of BERT (Bidirectional Encoder Representation Transformer) types,” *J. Phys. Conf. Ser.*, vol. 1963, no. 1, 2021, doi: 10.1088/1742-6596/1963/1/012173.

[47] A. Wettig, T. Gao, Z. Zhong, and D. Chen, “Should You Mask 15% in Masked Language Modeling?,” *EACL 2023 - 17th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf.*, pp. 2977–2992, 2023, doi: 10.18653/v1/2023.eacl-main.217.

[48] G. Penha and C. Hauff, “What does BERT know about books, movies and music? Probing BERT for Conversational Recommendation,” *RecSys 2020 - 14th ACM Conf. Recomm. Syst.*, pp. 388–397, 2020, doi: 10.1145/3383313.3412249.

[49] M. Bozuyla and A. ÖzçİIft, “Developing a fake news identification model with advanced deep language transformers for Turkish covid-19 misinformation data,” *Turkish J. Electr. Eng. Comput. Sci.*, pp. 908–926, 2021, doi: 10.3906/elk-2106-55.

[50] Q. Liu, M. J. Kusner, and P. Blunsom, “A Survey on Contextual Embeddings,” 2020, [Online]. Available: http://arxiv.org/abs/2003.07278

[51] J. White, “PubMed 2.0,” *Med. Ref. Serv. Q.*, vol. 39, no. 4, pp. 382–387, 2020, doi: 10.1080/02763869.2020.1826228.

[52] P. Lewis, M. Ott, J. Du, and V. Stoyanov, “Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art,” pp. 146–157, 2020, doi: 10.18653/v1/2020.clinicalnlp-1.17.

[53] S. Serrano, “Language Models : A Guide for the Perplexed,” pp. 1–35.

[54] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, “Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking,” *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[55] A. A. Ramadana Lubis, S. I. Purnama, and M. A. Afandi, “Sistem Pendeteksi Kantuk Berbasis Metode Haar Cascade Untuk Aplikasi Computer Vision,” *Techno.Com*, vol. 22, no. 3, pp. 589–598, 2023, doi: 10.33633/tc.v22i3.8464.

[56] I. Fursov *et al.*, “Sequence Embeddings Help Detect Insurance Fraud,” *IEEE Access*, vol. 10, pp. 32060–32074, 2022, doi: 10.1109/ACCESS.2022.3149480.

[57] Q. Li *et al.*, “A Survey on Text Classification: From Traditional to Deep Learning,” *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, 2022, doi: 10.1145/3495162.