

**IMPLEMENTASI MODEL *INDOBERT* PADA ANALISIS
SENTIMEN TERHADAP ISU FEMINISME DI TWITTER**

SKRIPSI

**Diajukan untuk memenuhi salah satu syarat kelulusan
di Program Studi Sains Data**



Disusun Oleh:

BRESCIA AYUNDINA YUNIAROSSY

20083010021

**UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SAINS DATA
SURABAYA
2024**

LEMBAR PENGESAHAN

**IMPLEMENTASI MODEL INDOBERT PADA ANALISIS SENTIMEN
TERHADAP ISU FEMINISME DI TWITTER**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Sains Data
pada : Senin, 15 Juli 2024

**Program Studi S-1 Sains Data
Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya**

Oleh :

BRESCIA AYUNDINA YUNIAROSSY

NPM. 20083010021

Disetujui oleh Tim Penguji Skripsi :

Penguji 1

Penguji 2

Dr. Ir. Mohammad Ichom, SP., S.Kom., MT.
NIP. 198303102021211006

Trunono, S.Si., M.Si
NIP. 199509032022031003

Pembimbing 1

Pembimbing 2

Kartika Maulida Hindrayani, S.Kom., M.Kom
NIP. 199209092022032009

Aviolla Terza Damaliana, S.Si., M.Stat
NIP. 199408022022032015

**Fakultas Ilmu Komputer
Dekan,**

Mengetahui,

**Program Studi Sains Data
Fakultas Ilmu Komputer
Koordinator,**

Prof. Dr. Ir. Novirina Hendrasarie, MT
NIP. 196811261994032001

Dr. Eng. Ir. Dwi Arman Prasetya, ST, MT, IPU
NIP. 198012052005011002

Surabaya, Juli, 2024

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Brescia Ayundina Yuniarossy
NPM : 20083010021
Program Studi : Sains Data

Menyatakan bahwa judul Skripsi / Tugas Akhir sebagai berikut:

Implementasi Model *IndoBERT* Pada Analisis Sentimen Terhadap Isu Feminisme di Twitter

Bukan merupakan plagiat dari Skripsi/ Tugas Akhir/ Penelitian orang lain dan juga bukan merupakan produk/ *software*/ hasil karya yang saya beli dari orang lain

Saya juga menyatakan bahwa Skripsi/ Tugas Akhir ini adalah pekerjaan saya sendiri, kecuali yang dinyatakan dalam Daftar Pustaka, dan tidak pernah diajukan untuk syarat memperoleh gelar di Universitas Pembangunan Nasional "Veteran" Jawa Timur maupun di institusi pendidikan lain.

Jika ternyata dikemudian hari pernyataan ini terbukti tidak benar, maka Saya bertanggung jawab penuh dan siap menerima segala konsekuensi, termasuk pembatalan ijazah dikemudian hari

Surabaya, 05 Juli 2024

Hormat Saya



Brescia Ayundina Yuniarossy
NPM. 20083010021

ABSTRAK

IMPLEMENTASI MODEL *INDOBERT* PADA ANALISIS SENTIMEN TERHADAP ISU FEMINISME DI TWITTER

Nama Mahasiswa / NPM : **Brescia Ayundina Yuniarossy / 20083010021**
Program Studi : **Sains Data, FASILKOM,UPN Veteran Jatim**
Dosen Pembimbing 1 : **Kartika Maulida Hindrayani, S.Kom., M.Kom**
Dosen Pembimbing 2 : **Aviolla Terza Damaliana, S.Si., M.Stat**

ABSTRAK

Penelitian ini berfokus pada analisis sentimen masyarakat terhadap isu Kekerasan Dalam Rumah Tangga (KDRT) dan Pelecehan Seksual di Indonesia, dua isu sosial yang serius dan berdampak luas. Dalam era digital, media sosial seperti Twitter menjadi *platform* untuk mengekspresikan opini publik. Menggunakan model *IndoBERT*, varian dari *Bidirectional Encoder Representations from Transformers* (BERT) khusus bahasa Indonesia untuk analisis sentimen masyarakat terhadap dua isu sosial. Penelitian ini menunjukkan bahwa *IndoBERT* mampu mengklasifikasikan sentimen publik dengan akurasi 89% pada dataset Pelecehan Seksual tanpa ekstrasi fitur *Bag of Words* (BoW) dan SMOTE. Data yang digunakan dalam penelitian ini didapatkan melalui proses *crawling* dari Twitter dengan total 3007 data dari kedua topik tersebut. Setelah itu, dilakukan tahap *preprocessing* untuk membersihkan data dari *noise*. Sentimen dari *tweet* yang terkumpul diberikan label positif, negatif, dan netral. Penelitian ini menunjukkan bahwa *IndoBERT* efektif dalam mengklasifikasikan sentimen publik dan dapat membantu dalam memahami persepsi masyarakat terhadap isu-isu feminisme, khususnya KDRT dan pelecehan seksual.

Kata Kunci: Analisis Sentimen, Feminisme, *IndoBERT*, *BoW*, *SMOTE*

ABSTRACT

IMPLEMENTATION OF *INDOBERT* MODEL ON SENTIMENT ANALYSIS OF FEMINISM ISSUES ON TWITTER

Student Name / NPM : **Brescia Ayundina Yuniarossy / 20083010021**
Study Program : **Sains Data, FASILKOM,UPN Veteran Jatim**
Advisor 1 : **Kartika Maulida Hindrayani, S.Kom., M.Kom**
Advisor 2 : **Aviolla Terza Damaliana, S.Si., M.Stat**

ABSTRACT

This research focuses on analyzing public sentiment towards the issues of Domestic Violence and Sexual Harassment in Indonesia, two serious and far-reaching social issues. In the digital age, social media such as Twitter has become a platform for expressing public opinion. Using the IndoBERT model, a variant of the Indonesian language-specific Bidirectional Encoder Representations from Transformers (BERT) for public sentiment analysis on two social issues. This research shows that IndoBERT is able to classify public sentiment with 89% accuracy on the Sexual Harassment dataset without Bag of Words (BoW) and SMOTE feature extraction. The data used in this study was obtained through a crawling process from Twitter with a total of 3007 data from both topics. After that, a preprocessing stage is carried out to clean the data from noise. The sentiments of the collected tweets are labeled as positive, negative, and neutral. This research shows that IndoBERT is effective in classifying public sentiment and can help in understanding the public perception of feminism issues, especially domestic violence and sexual harassment.

Keywords: Sentiment Analysis, Feminism, IndoBERT, BoW, SMOTE

KATA PENGANTAR

Puji dan syukur kehadirat ALLAH SWT, atas limpahan Rahmat serta Kasih Sayang-Nya sehingga penulis dapat menyelesaikan Proposal Skripsi yang merupakan persyaratan dalam menyelesaikan mata kuliah Seminar Proposal pada Program Studi S1 Sains Data di Universitas Pembangunan Nasional “Veteran” Jawa Timur.

Dalam penyusunan Skripsi ini tidak terlepas dari bantuan berbagai pihak, dan dalam kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Kedua orang tua dan keluarga yang selalu memberikan doa dan dukungan kepada penulis, sehingga penulis bersemangat dalam menyelesaikan skripsi ini.
2. Ibu Kartika Maulida Hindrayani, S.Kom., M.Kom., selaku dosen pembimbing 1 penulis selama penelitian ini dilakukan.
3. Ibu Aviolla Terza Damaliana, S.Si., M.Stat., selaku dosen pembimbing 2 penulis selama penelitian ini dilakukan.
4. Bapak Dr. Ir. Mohammad Idhom, SP., S.Kom., MT, selaku dosen penguji 1 penulis yang sangat baik dan mengarahkan penulis selama proses revisi berlangsung.
5. Bapak Trimono, S.Si., M.Si, selaku dosen penguji 2 penulis yang mampu mengarahkan penulis dalam kepenulisan skripsi ini.
6. Bapak Tresna Maulana Fahrudin, S.ST., M.T., selaku dosen wali penulis.
7. Kepada diri sendiri (penulis) terima kasih karena kamu sudah bertahan dan mampu menyelesaikan skripsi ini dengan sehat dan semangat juang.
8. Brescia Ayundini Yuniarossy selaku kembaran perempuan yang telah memberikan dukungan dan tahap proses skripsi ini.
9. Rangga Saputra selaku teman penulis yang memiliki peranan penting selama penelitian ini.
10. Alm. Muhammad Jibril Izzah Nur Rahm selaku teman penulis yang sangat berharga.

11. Kepada All Unit NCT, Red Velvet, Aespa, iKON, BlackPink, BABYMONSTER yang sudah mendukung penulis (secara tidak langsung) dengan karyanya yang menjadi penyemangat dalam mengerjakan skripsi ini.
 12. Lee Taeyong, Lee Jeno, Nakamoto Yuta, Winwin, Xiaojun, Koo Junhoe, Kang Seulgi, NingNing, Kwon Yuri, Kyuhyun, Kang Daniel, Ong Seongwu, Park Woojin, Kim Jaehwan yang telah memberikan gambaran perjuangan sampai titik sukses kepada penulis agar cepat menyelesaikan skripsi ini.
 13. Kepada Nurhalimah selaku teman *online* yang saya sayangi dan selalu berada disamping saya sedari saya SMP.
 14. Kepada Dea Salfia Murdiasari selaku teman kost saya yang selalu berada disamping bahkan menemani penulis selama proses pembuatan skripsi ini.
 15. Teman seperjuangan penulis selama menempuh Pendidikan di Sains Data.
- Penulis menyadari bahwa masih terdapat banyak kekurangan dalam Proposal Skripsi ini, namun penulis berharap semoga Proposal Skripsi ini dapat memberikan kontribusi terhadap perkembangan ilmu pengetahuan, khususnya dalam bidang ilmu Sains Data.

Surabaya, 15 Juli 2024

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN	ii
SURAT PERNYATAAN.....	iii
ABSTRAK	iv
ABSTRACT	v
KATA PENGANTAR	vi
DAFTAR ISI.....	viii
DAFTAR GAMBAR	xi
DAFTAR TABEL.....	xiii
LAMPIRAN.....	xv
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Batasan Masalah.....	5
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	6
BAB II TINJAUAN PUSTAKA.....	7
2.1 Dasar Teori.....	7
2.1.1 Analisis Sentimen	7
2.1.2 Twitter	8
2.1.3 <i>Crawling Data</i>	8
2.1.4 <i>Bidirectional Encoder Representation from Transformers (BERT)</i> ...	8
2.1.5 <i>IndoBERT</i>	12
2.1.6 <i>Transformers</i>	17
2.1.7 Ekstrasi Fitur <i>Bag of Words</i>	17
2.1.8 <i>Imbalance Data Menggunakan Synthetic Minority Oversampling Technique (SMOTE)</i>	18
2.1.9 <i>Confusion Matrix</i>	18
2.2 Penelitian Terdahulu	21
BAB III METODOLOGI PENELITIAN.....	26
3.1 Variabel Penelitian dan Sumber Data	26

3.1.1	Variabel Penelitian	26
3.1.2	Sumber Data	26
3.2	Langkah Analisis.....	28
3.2.1	Pengumpulan Data.....	29
3.2.2	<i>Preprocessing Data</i>	29
3.2.3	Pelabelan Data	32
3.2.4	Ekstrasi Fitur.....	33
3.2.5	Evaluasi Model	34
3.3	Diagram Alir Penelitian.....	35
3.4	Jadwal Penelitian.....	36
BAB IV HASIL DAN PEMBAHASAN		37
4.1	Pengumpulan Data	37
4.2	<i>Preprocessing Data</i>	38
4.3	Pelabelan Data.....	46
4.4	<i>Word Cloud</i>	56
4.5	Ekstrasi Fitur	58
4.6	Penanganan Data Tidak Seimbang Menggunakan <i>SMOTE</i>	65
4.7	Skenario 1: Model <i>IndoBERT</i> Tanpa Ekstrasi Fitur <i>BoW</i>	66
4.8	Skenario 2: Model <i>IndoBERT</i> Dengan Ekstrasi Fitur <i>BoW</i>	67
4.9	Skenario 3: Model <i>IndoBERT</i> Dengan <i>Epoch</i> Berbeda (3 dan 5).....	69
4.10	Evaluasi Model.....	70
4.10.1	Skenario 1: Model <i>IndoBERT</i> Tanpa Ekstrasi Fitur <i>BoW</i>	70
4.10.2	Skenario 2: Model <i>IndoBERT</i> Dengan Ekstrasi Fitur <i>BoW</i>	79
4.10.3	Skenario 3: Model <i>IndoBERT</i> Dengan <i>Epoch</i> Berbeda (3 dan 5)	88
4.10.4	Analisis Hasil Skenario Model <i>IndoBERT</i>	90
4.11	<i>Deployment</i> Aplikasi.....	93
BAB V KESIMPULAN		94
5.1	Kesimpulan	94
5.2	Saran.....	95
DAFTAR PUSTAKA		96
LAMPIRAN		99

BIODATA PENULIS	104
-----------------------	-----

DAFTAR GAMBAR

Gambar 2.1 Arsitektur <i>BERT</i>	9
Gambar 2.2 <i>Encoder</i>	11
Gambar 2.3 <i>Decoder</i>	12
Gambar 2.4 Pelatihan <i>BERT</i>	14
Gambar 3.1 Diagram Alir <i>Preprocessing</i> Data.....	30
Gambar 3.2 Proses <i>Labeling Transformers</i>	32
Gambar 3.3 Diagram Alir Penelitian	35
Gambar 4.1 Distribusi Data Pelabelan Kamus Pelecehan Seksual	50
Gambar 4.2 Distribusi Data Pelabelan Kamus KDRT.....	51
Gambar 4.3 Distribusi Data Pelabelan <i>Transformers</i> KDRT	53
Gambar 4.4 Distribusi Data Pelabelan <i>Transformers</i> Pelecehan Seksual.....	54
Gambar 4.5 <i>Word Cloud</i> Data KDRT	56
Gambar 4.6 <i>Word Cloud</i> Data Pelecehan Seksual	57
Gambar 4.7 20 Kata Teratas Label Kamus Data Pelecehan Seksual	60
Gambar 4.8 20 Kata Teratas Label <i>Transformers</i> Data Pelecehan Seksual	62
Gambar 4.9 20 Kata Teratas Label Kamus Data KDRT	64
Gambar 4.10 20 Kata Teratas Label <i>Transformers</i> Data KDRT	65
Gambar 4.11 <i>Cross Validation</i> Metode <i>SMOTE</i>	66
Gambar 4.12 <i>Confusion Matrix</i> Pelecehan Seksual Tanpa <i>BoW</i> dan <i>SMOTE</i> (Kamus).....	71
Gambar 4.13 <i>Classification Report</i> Pelecehan Seksual tanpa <i>BoW</i> (Kamus).....	72
Gambar 4.14 <i>Confusion Matrix</i> KDRT tanpa <i>BoW</i> dan <i>SMOTE</i> (Kamus).....	73
Gambar 4.15 <i>Classification Report</i> KDRT tanpa <i>BoW</i> dan <i>SMOTE</i> (Kamus).....	74
Gambar 4.16 <i>Confusion Matrix</i> Pelecehan Seksual tanpa <i>BoW</i> (<i>Transformers</i>)..	75
Gambar 4.17 <i>Classification Report</i> Pelecehan Seksual Tanpa <i>BoW</i> dan <i>SMOTE</i> (<i>Transformers</i>).....	76
Gambar 4.18 <i>Confusion Matrix</i> Data KDRT tanpa <i>BoW</i> (<i>Transformers</i>)	77
Gambar 4.19 <i>Classification Report</i> Data KDRT tanpa <i>BoW</i> (<i>Transformers</i>)	78
Gambar 4.20 <i>Confusion Matrix</i> Data Pelecehan Seksual dengan <i>BoW</i> (Kamus) .	80

Gambar 4.21 <i>Classification Report</i> Pelecehan Seksual dengan <i>BoW</i> (Kamus)....	81
Gambar 4.22 <i>Confusion Matrix</i> KDRT dengan <i>BoW</i> (Kamus)	82
Gambar 4.23 <i>Classification Report</i> KDRT dengan <i>BoW</i> (Kamus).....	83
Gambar 4.24 <i>Confusion Matrix</i> Pelecehan Seksual dengan <i>BoW</i> dan <i>SMOTE</i> (<i>Transformers</i>)	85
Gambar 4.25 <i>Classification Report</i> Pelecehan Seksual dengan <i>BoW</i> dan <i>SMOTE</i> (<i>Transformers</i>).....	85
Gambar 4.26 <i>Confusion Matrix</i> KDRT dengan <i>BoW</i> (<i>Transformers</i>).....	87
Gambar 4.27 <i>Classification Report</i> KDRT dengan <i>BoW</i> (<i>Transformers</i>)	87
Gambar 4.28 Tampilan <i>Deployment</i>	93

DAFTAR TABEL

Tabel 2.1 Tabel <i>Confusion Matrix</i>	19
Tabel 2.2 Studi Literatur Penelitian Terdahulu	21
Tabel 3.1 Tabel Sumber Data (1)	27
Tabel 3.2 Tabel Lanjutan Sumber Data (2)	27
Tabel 3.3 Tabel Lanjutan Sumber Data (3)	28
Tabel 3.4 Penerapan <i>BoW</i> Berbentuk List	33
Tabel 3.5 Frekuensi Kata Dalam Dokumen	33
Tabel 3.6 Tabel <i>Timeline</i> Penelitian	36
Tabel 4.1 Contoh Tabel Normalisasi Kata	31
Tabel 4.2 Hasil <i>Crawling</i> Data Berdasarkan <i>Keyword</i>	38
Tabel 4.3 Algoritma <i>Preprocessing Case Folding</i>	39
Tabel 4.4 Hasil Proses <i>Case Folding</i>	40
Tabel 4.5 Algoritma <i>Preprocessing</i> Normalisasi Kata	40
Tabel 4.6 Hasil Normalisasi Kata	41
Tabel 4.7 Algoritma <i>Preprocessing Stopword Removal</i>	42
Tabel 4.8 Hasil <i>Stopword Removal</i>	43
Tabel 4.9 Algoritma <i>Preprocessing Stemming</i>	43
Tabel 4.10 Hasil <i>Stemming</i>	44
Tabel 4.11 Algoritma <i>Preprocessing</i> Data Tahapan <i>Cleaning Data</i>	45
Tabel 4.12 Hasil <i>Cleaning Data</i>	46
Tabel 4.13 Algoritma Pelabelan Kamus	47
Tabel 4. 14 Algoritma Pelabelan <i>Transformers</i>	48
Tabel 4.15 Hasil Label Kamus Dataset Pelecehan Seksual	49
Tabel 4.16 Hasil Label Kamus Dataset KDRT	51
Tabel 4.17 Hasil Label <i>Transformers</i> Dataset KDRT	52
Tabel 4.18 Hasil Label <i>Transformers</i> Dataset Pelecehan Seksual	53
Tabel 4.19 Hasil Validasi Pelabelan Kamus dan Pelabelan <i>Transformers</i>	54
Tabel 4.20 Hasil Pemisahan Kata Menjadi <i>List</i>	58
Tabel 4.21 Hasil Ekstraksi Fitur <i>BoW</i>	58
Tabel 4.22 Perbedaan <i>Split</i> Data <i>BoW</i> Dengan <i>Split IndoBERT</i>	68

Tabel 4.23 Model Klasifikasi <i>IndoBERT</i>	69
Tabel 4.24 Tabel Pengujian tanpa Ekstrasi Fitur	71
Tabel 4.25 Tabel Pengujian tanpa Ekstrasi Fitur	79
Tabel 4.26 Tabel Pengujian tanpa Ekstrasi Fitur	88
Tabel 4.27 Hasil Akurasi <i>IndoBERT</i> Pada Dataset Pelecehan Seksual	90
Tabel 4.28 Hasil Akurasi <i>IndoBERT</i> Pada Dataset KDRT	91

LAMPIRAN

Lampiran 1. Hasil uji plagiasi	99
Lampiran 2. <i>Source Code</i> yang digunakan	101