



UNIVERSITAS PERTAHANAN RI

**PERBANDINGAN ALGORITMA *LOGISTIC REGRESSION*
DAN *RANDOM FOREST* DALAM MENGLASIFIKASI
UJARAN KEBENCIAN TERHADAP PEMERINTAH
PADA MEDIA SOSIAL X**

ALFIANO DENI KUSUMA

320200401003

**Skripsi yang Ditulis untuk Memenuhi Sebagian Persyaratan dalam
Mendapatkan Gelar Sarjana**

**FAKULTAS SAINS DAN TEKNOLOGI PERTAHANAN
PROGRAM STUDI INFORMATIKA**

BOGOR 2024

LEMBAR PERSETUJUAN SKRIPSI

Nama : Alfiano Deni Kusuma
NIM : 320200401003
Program Studi : Informatika
Fakultas : Sains dan Teknologi Pertahanan
Judul Skripsi : Perbandingan Algoritma *Logistic Regression*
dan *Random Forest* dalam Mengklasifikasi
Ujaran Kebencian Terhadap Pemerintah pada
Media Sosial X

Pembimbing I,



Nadiza Lediwara, S.T., M.Eng.
Penata Muda Tk.I III/b
NIP. 198802092022032001

Tanggal: 23 - 07 - 2024

Pembimbing II,



Aditya Adiprabowo, S.T., M.Sc.
Penata Muda Tk.I III/b
NIP. 199603212022031005

Tanggal: 23 - 07 - 2024

Mengetahui,

Kepala Program Studi
Informatika,



Adam Mardamsyah, M.Han.
Kolonel Inf NRP. 11940019450871

Tanggal: 23 - 07 - 2024

Dekan
Fakultas Sains dan Teknologi Pertahanan,



Prof. Dr. Ir. Muhamad Asvial, M.Eng.
Pembina Utama Muda IV/c

Tanggal: 25 - 07 - 2024

LEMBAR PENGESAHAN SKRIPSI

Nama : Alfiano Deni Kusuma
 NIM : 320200401003
 Program Studi : Informatika
 Fakultas : Sains dan Teknologi Pertahanan
 Judul Skripsi : Perbandingan Algoritma *Logistic Regression* dan *Random Forest* dalam Mengklasifikasi Ujaran Kebencian Terhadap Pemerintah pada Media Sosial X

No	Nama	Tanda Tangan	Tanggal
1	Dosen Pembimbing I : Nadiza Lediwara, S.T., M.Eng. Penata Muda Tk.I III/b NIP. 198802092022032001		19/7 24
2	Dosen Pembimbing II : Aditya Adiprabowo, S.T., M.Sc. Penata Muda Tk.I III/b NIP. 199603212022031005		19/7 24
3	Dosen Penguji I: Anindito, S.Kom., S.S., MTI., CHFI. PPPK Golongan X NIP. 197805192024211003		19/7 24
4	Dosen Penguji II: Dr. Hondor Saragih, S.T., Msi (Han)., M.M.S.I. Kolonel Laut (KH) NRP. 14633/P		19/7 24 07
5	Dosen Penguji III: Ajeng Hidayati, S.Kom., M.Kom. PPPK Golongan X NIP. 197805192024211003		19/7 24

PERNYATAAN ORISINALITAS

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya atau bagian karya yang pernah diajukan untuk memperoleh gelar kesarjanaan jenjang apa pun di suatu Perguruan Tinggi; dan sepanjang sepengetahuan saya juga tidak terdapat istilah, frasa, kalimat, paragraf, subbab atau bab dari karya yang pernah ditulis atau diterbitkan; kecuali yang secara tertulis diajukan dalam naskah ini dan disebutkan dalam Daftar Referensi.

Apabila di kemudian hari terbukti bahwa terdapat plagiat dalam skripsi ini, saya bersedia menerima sanksi sesuai ketentuan peraturan/undang-undang yang berlaku.

Bogor, 22 Juli 2024



Alfiano Deni Kusuma

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Pertahanan Republik Indonesia, saya yang bertanda tangan di bawah ini:

Nama : Alfiano Deni Kusuma
NIM : 320200401003
Program Studi : Informatika
Fakultas : Fakultas Sains dan Teknologi Pertahanan
Jenis Karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Pertahanan Republik Indonesia **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul:

PERBANDINGAN ALGORITMA *LOGISTIC REGRESSION* DAN *RANDOM FOREST* DALAM MENGLASIFIKASI UJARAN KEBENCIAN TERHADAP PEMERINTAH PADA MEDIA SOSIAL X

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Pertahanan Republik Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemilik Hak Cipta/Karya Intelektual dari skripsi ini.

Demikian pernyataan ini saya buat dengan kesadaran penuh tanpa paksaan dari pihak manapun.

Bogor, 22 Juli 2024

Yang menyatakan



Alfiano Deni Kusuma

KATA PENGANTAR

Puji syukur yang setinggi-tingginya penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas limpahan berkat dan rahmatnya sehingga penulis dapat menyelesaikan skripsi yang judul: “Perbandingan Algoritma *Logistic Regression* dan *Random Forest* dalam Mengklasifikasi Ujaran Kebencian Terhadap Pemerintah pada Media Sosial X” secara tepat waktu. Penyusunan skripsi ini menjadi salah satu syarat bagi penulis untuk memperoleh gelar Sarjana pada Program Studi Informatika Fakultas Sains dan Teknologi Pertahanan Universitas Pertahanan RI. Penulis menyadari bahwa selama menempuh pendidikan di Universitas Pertahanan RI ini dan selama proses penyusunan skripsi ini telah banyak perantara-perantara dari Tuhan yang membantu penulis. Oleh sebab itu penulis mengucapkan terima kasih yang sebesar-sebesarannya kepada pihak-pihak yang telah banyak membantu penulis, pihak-pihak tersebut di antaranya adalah:

1. Ibu Nadiza Lediwara, S.T., M.Eng., selaku dosen pembimbing satu yang telah memberikan banyak ide, saran dan masukan kepada peneliti dalam penyusunan skripsi ini.
2. Bapak Aditya Adiprabowo, S.T., M.Sc., selaku dosen pembimbing dua yang telah membimbing penulis dalam penyusunan skripsi ini dengan baik.
3. Orang Tua Penulis, terkhusus ibu penulis, Ibu Maria Suryaningsih yang telah merawat penulis dengan penuh kasih sayang hingga penulis mampu mencapai titik ini.
4. Segenap dosen program studi Informatika yang telah banyak memberikan ilmu dan pengalaman berharga bagi penulis selama penulis menempuh pendidikan di Universitas Pertahanan RI.
5. Segenap keluarga penulis yang telah memberikan banyak dukungan, motivasi, dan doa selama peneliti melaksanakan skripsi.

6. Rekan-rekan seperjuangan *cohort-1* Informatika yang telah bersama-sama menemani penulis dalam menjalani pendidikan.
7. Adik-adik asuh prodi dan resimen penulis yang senantiasa memberi dukungan dan motivasi bagi penulis dalam menyusun skripsi ini.

Semoga Tuhan Yang Maha Esa membalas kebaikan-kebaikan dari berbagai pihak yang telah membantu penulis. Penulis sadar bahwa skripsi ini masih jauh dari kata sempurna, maka dari itu dengan kerendahan hati mengharapkan saran dan kritik yang membangun guna menyempurnakan skripsi ini. Akhirnya, semoga skripsi ini dapat memberikan kontribusi terhadap pengembangan ilmu pertahanan dan bermanfaat bagi pemangku kebijakan terkait dalam perumusan kebijakan pertahanan.



ABSTRAK

PERBANDINGAN ALGORITMA *LOGISTIC REGRESSION* DAN *RANDOM FOREST* DALAM MENGLASIFIKASI UJARAN KEBENCIAN TERHADAP PEMERINTAH PADA MEDIA SOSIAL X

Media sosial saat ini menjadi tempat di internet yang memungkinkan para penggunanya untuk berinteraksi, bekerja sama, berkomunikasi, dan membentuk ikatan sosial. Salah satu media sosial yang banyak digunakan saat ini adalah media sosial X yang telah mencapai 60.2% pengguna aktif harian. Sering kali dalam interaksi antar pengguna media sosial X muncul *tweet* yang mengandung ujaran kebencian terhadap pemerintah. Hal ini dapat mengakibatkan terjadinya tekanan politik dan memicu emosi masyarakat yang dapat mengakibatkan suatu konflik. Sehingga perlu adanya deteksi dan klasifikasi terhadap *tweet* tersebut. Salah satu metode yang dapat digunakan adalah metode *Natural Language Processing* (NLP) untuk membangun model klasifikasi dengan algoritma seperti *Logistic Regression* dan *Random Forest* serta pengoptimalan model dengan algoritma *boosting* seperti *Adaboost*. Dari beberapa pilihan algoritma tersebut perlu dianalisis algoritma mana yang dapat memberikan tingkat akurasi paling tinggi. Guna melakukan analisis tersebut, maka digunakan metode CRISP-DM untuk mengetahui kombinasi algoritma yang tepat untuk membangun model klasifikasi dengan tingkat akurasi yang tinggi. Melalui metode CRISP-DM didapat bahwa model yang dibangun dengan algoritma *Logistic Regression* mendapat akurasi sebesar 0.8 atau 80%, *Random Forest* 0.7796 atau 77.96%, *Random Forest* dengan *Adaboost* 0.7890 atau 78.90%, dan model *Logistic Regression* yang dioptimalkan dengan *Adaboost* memiliki akurasi tertinggi, yaitu 0.8125 atau 81.25%.

Kata Kunci: Media Sosial X, Ujaran Kebencian, Klasifikasi Teks, *Logistic Regression*, *Random Forest*, *Adaboost*

ABSTRACT

COMPARISON OF LOGISTIC REGRESSION AND RANDOM FOREST ALGORITHMS IN CLASSIFYING HATE SPEECH AGAINST THE GOVERNMENT ON SOCIAL MEDIA X

Social media today has become an internet platform that enables its users to interact, collaborate, communicate, and form social bonds. One of the widely used social media platforms today is social media X, which has reached 60.2% daily active users. Often, in interactions among users of social media X, tweets containing hate speech against the government emerge. This can lead to political pressure and trigger public emotions, potentially resulting in conflict. Therefore, it is necessary to detect and classify such tweets. One method that can be used is Natural Language Processing (NLP) to build a classification model using algorithms such as Logistic Regression and Random Forest, along with model optimization using boosting algorithms like Adaboost. Among these algorithm options, it is essential to analyze which algorithm can provide the highest accuracy rate. To conduct this analysis, the CRISP-DM method is used to identify the right combination of algorithms to build a classification model with a high accuracy rate. Through the CRISP-DM method, it was found that the model built with the Logistic Regression algorithm achieved an accuracy of 0.8 or 80%, Random Forest 0.7796 or 77.96%, Random Forest with Adaboost 0.7890 or 78.90%, and the Logistic Regression model optimized with Adaboost had the highest accuracy, which was 0.8125 or 81.25%.

Keywords: Social Media X, Hate Speech, Text Classification, Logistic Regression, Random Forest, Adaboost

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PERSETUJUAN SKRIPSI	ii
LEMBAR PENGESAHAN SKRIPSI	iii
PERNYATAAN ORISINALITAS	iv
PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS	v
KATA PENGANTAR	vi
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR GAMBAR	xiv
DAFTAR TABEL	xvii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Tujuan Penelitian	4
1.4 Batasan Masalah	5
1.5 Manfaat Penelitian	5
1.5.1 Manfaat Teoritis	5
1.5.2 Manfaat Praktis	5
BAB II TINJAUAN PUSTAKA	7
2.1 Landasan Teori	7
2.1.1 Media Sosial	7
2.1.2 X (Sebelumnya Twitter)	8
2.1.3 Ujaran Kebencian	9

2.1.3.1	Tingkatan Ujaran Kebencian	12
2.1.3.2	Dampak Ujaran Kebencian.....	13
2.1.4	Pemerintah	15
2.1.5	<i>Natural Language Processing</i>	16
2.1.5.1	<i>Crawling</i>	17
2.1.5.2	<i>Preprocessing</i>	18
2.1.5.3	TF-IDF.....	19
2.1.6	Klasifikasi Teks	20
2.1.7	<i>Logistic Regression</i>	20
2.1.8	<i>Random Forest</i>	21
2.1.9	<i>Adaboost</i>	22
2.1.10	Evaluasi Model.....	23
2.1.10.1	<i>Confusion Matrix</i>	24
2.1.10.2	<i>Cross Validation</i>	25
2.1.11	CRISP-DM.....	26
2.2	Hasil Penelitian Terdahulu.....	28
2.3	Kerangka Berpikir	31
BAB III METODOLOGI PENELITIAN.....		34
3.1	Metode Penelitian	34
3.1.1	<i>Business Understanding</i>	34
3.1.2	<i>Data Understanding</i>	35
3.1.3	<i>Data Preparation</i>	36
3.1.4	<i>Modelling</i>	39
3.1.5	<i>Evaluation</i>	41
3.1.6	<i>Deployment</i>	42
3.2	Tempat dan Waktu Penelitian.....	43
3.2.1	Tempat Penelitian.....	43

3.2.2 Waktu Penelitian.....	43
BAB IV HASIL DAN PEMBAHASAN.....	45
4.1 <i>Business Understanding</i>	45
4.1.1 Kebutuhan Perangkat Keras.....	46
4.1.2 Kebutuhan Perangkat Lunak	46
4.2 <i>Data Understanding</i>	46
4.3 <i>Data Preparation</i>	48
4.3.1 <i>Cleaning</i>	49
4.3.2 <i>Case Folding</i>	54
4.3.3 <i>Normalisasi</i>	55
4.3.4 <i>Tokenisasi</i>	56
4.3.5 <i>Stopword Removal</i>	57
4.3.6 <i>Stemming</i>	59
4.3.7 <i>Labelling</i>	60
4.3.8 <i>Pembobotan kata</i>	61
4.3.9 <i>Data Splitting</i>	62
4.4 <i>Modelling</i>	62
4.4.1 <i>Model Logistic Regression</i>	63
4.4.2 <i>Model Random Forest</i>	63
4.4.3 <i>Model Logistic Regression dengan Adaboost</i>	64
4.4.4 <i>Model Random Forest dengan Adaboost</i>	65
4.5 <i>Evaluation</i>	66
4.5.1 <i>Evaluasi Confusion Matrix</i>	67
4.5.1.1 <i>Model Logistic Regression</i>	67
4.5.1.2 <i>Model Logistic Regression Adaboost</i>	69
4.5.1.3 <i>Model Random Forest</i>	70

4.5.1.4 Model <i>Random Forest Adaboost</i>	72
4.5.2 Evaluasi <i>Cross Validation</i>	74
4.5.2.1 Model <i>Logistic Regression</i>	75
4.5.2.2 Model <i>Logistic Regression Adaboost</i>	77
4.5.2.3 Model <i>Random Forest</i>	79
4.5.2.4 Model <i>Random Forest Adaboost</i>	81
4.6 <i>Deployment</i>	82
4.7 Analisis Kinerja Model Klasifikasi	85
4.7.1 Kinerja Model Sebelum Penerapan <i>Adaboost</i>	85
4.7.2 Kinerja Model Setelah Penerapan <i>Adaboost</i>	89
4.8 Analisis Hasil Evaluasi	93
BAB V KESIMPULAN DAN SARAN	96
5.1 Kesimpulan	96
5.2 Saran	97
DAFTAR PUSTAKA	98
LAMPIRAN	111



DAFTAR GAMBAR

Gambar 1.1 <i>Most Used Social Media Platforms in Indonesia</i>	1
Gambar 2.1 Kasus Kejahatan Siber Ujaran Kebencian di Indonesia.....	11
Gambar 2.2 Tahapan CRISP-DM.....	27
Gambar 2.3 Kerangka Berpikir.....	32
Gambar 3.1 Contoh Data Hasil <i>Crawling</i>	36
Gambar 4.1 Jumlah Data Hasil Scrapping	47
Gambar 4.2 Data Hasil Scrapping	47
Gambar 4.3 <i>Tweet</i> Hasil Scrapping.....	48
Gambar 4.4 Kode fungsi <i>remove_retweets</i>	49
Gambar 4.5 kode fungsi <i>remove_mentions</i>	49
Gambar 4.6 Kode fungsi <i>remove_number</i>	50
Gambar 4.7 Kode <i>remove_url</i>	50
Gambar 4.8 Kode <i>remove_hashtags</i>	50
Gambar 4.9 Kode fungsi <i>remove_emoji</i>	51
Gambar 4.10 Kode fungsi <i>remove_punc</i>	51
Gambar 4.11 Kode Hapus Spasi Berlebih.....	52
Gambar 4.12 Kode Hapus <i>tweet</i> Kurang dari dua Kata	52
Gambar 4.13 Kode Hapus kata berulang.....	52
Gambar 4.14 Kode Hapus Duplikasi Data	53
Gambar 4.15 Contoh <i>tweet</i> Setelah Proses <i>Cleaning</i>	54
Gambar 4.16 Kode <i>Case Folding</i>	54
Gambar 4.17 Teks Hasil <i>Case Folding</i>	55
Gambar 4.18 Kode Normalisasi	55
Gambar 4.19 Teks Hasil Normalisasi dan Kata yang Diganti	56
Gambar 4.20 Kode Tokenisasi	56
Gambar 4.21 Teks Hasil Tokenisasi	57
Gambar 4.22 Kode <i>Stopword Removal</i>	58
Gambar 4.23 Teks Hasil <i>Stopword Removal</i>	58
Gambar 4.24 Kode <i>Stemming</i>	59

Gambar 4.25 Teks Hasil <i>Stemming</i>	60
Gambar 4.26 Teks Hasil Pelabelan	61
Gambar 4.27 Kode Pembobotan Kata TF-IDF	61
Gambar 4.28 Kode <i>Data Spliting</i>	62
Gambar 4.29 Pemodelan dengan Algoritma <i>Logistic Regression</i>	63
Gambar 4.30 Pemodelan dengan Algoritma <i>Random Forest</i>	64
Gambar 4.31 Pemodelan dengan <i>Logistic Regression Adaboost</i>	65
Gambar 4.32 Pemodelan dengan <i>Random Forest Adaboost</i>	66
Gambar 4.33 <i>Confusion Matrix</i> Model <i>Logistic Regression</i>	67
Gambar 4.34 Nilai Akurasi, Presisi, <i>Recall</i> , dan <i>F1-Score</i> Model <i>Logistic Regression</i>	68
Gambar 4.35 <i>Confusion Matrix</i> Model <i>Logistic Regression</i> dengan <i>Adaboost</i>	69
Gambar 4.36 Nilai Akurasi, Presisi, <i>Recall</i> , dan <i>F1-Score</i> Model <i>Logistic Regression</i> dengan <i>Adaboost</i>	70
Gambar 4.37 <i>Confusion Matrix</i> Model <i>Random Forest</i>	71
Gambar 4.38 Nilai Akurasi, Presisi, <i>Recall</i> , dan <i>F1-Score</i> Model <i>Random Forest</i>	72
Gambar 4.39 <i>Confusion Matrix</i> Model <i>Random Forest</i> dengan <i>Adaboost</i>	73
Gambar 4.40 Nilai Akurasi, Presisi, <i>Recall</i> , dan <i>F1-Score</i> Model <i>Random Forest</i> dengan <i>Adaboost</i>	74
Gambar 4.41 Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i>	75
Gambar 4.42 Diagram Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i>	76
Gambar 4.43 Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i> dengan <i>Adaboost</i>	77
Gambar 4.44 Diagram Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i> dengan <i>Adaboost</i>	78
Gambar 4.45 Akurasi Setiap <i>Fold</i> Model <i>Random Forest</i>	79
Gambar 4.46 Diagram Akurasi Setiap <i>Fold</i> Model <i>Random Forest</i>	80
Gambar 4.47 Akurasi Setiap <i>Fold</i> Model <i>Random Forest</i> dengan <i>Adaboost</i>	81

Gambar 4.48 Diagram Akurasi Setiap <i>Fold</i> Model <i>Random Forest</i> dengan <i>Adaboost</i>	82
Gambar 4.49 Diagram Hasil Klasifikasi <i>tweet</i> Masing-Masing Model Klasifikasi	83
Gambar 4.50 <i>Website Deployment</i> Model Klasifikasi	84
Gambar 4.51 Diagram Perbandingan Akurasi Model <i>Logistic Regression</i> dengan <i>Random Forest</i> Sebelum Penerapan <i>Adaboost</i>	86
Gambar 4.52 Diagram Perbandingan Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i> dengan <i>Random Forest</i> Sebelum Penerapan <i>Adaboost</i>	87
Gambar 4.53 Diagram Hasil Klasifikasi Data <i>Tweet</i> Model <i>Logistic Regression</i> (sebelah kiri) dan <i>Random Forest</i> (sebelah kanan)	88
Gambar 4.54 Diagram Perbandingan Akurasi Model <i>Logistic Regression</i> dengan <i>Random Forest</i> Setelah Penerapan <i>Adaboost</i>	89
Gambar 4.55 Diagram Perbandingan Akurasi Setiap <i>Fold</i> Model <i>Logistic Regression</i> dengan <i>Random Forest</i> Setelah Penerapan <i>Adaboost</i>	91
Gambar 4.56 Diagram Hasil Klasifikasi Data <i>Tweet</i> Model <i>Logistic Regression</i> (sebelah kiri) dan <i>Random Forest</i> (sebelah kanan)	92



DAFTAR TABEL

Tabel 2.1 <i>Confusion Matrix</i>	25
Tabel 2.2 Penelitian Terdahulu	29
Tabel 3.1 Simulasi <i>k-fold cross validation</i>	42
Tabel 3.2 Jadwal Penelitian	44
Tabel 4.1 Kebutuhan Perangkat Keras	46
Tabel 4.2 Kebutuhan Perangkat Lunak	46
Tabel 4.3 Nilai Akurasi Masing-masing Model Klasifikasi	93
Tabel 4.4 Nilai Akurasi Rata-Rata Setiap Model Klasifikasi	94

