

ABSTRAK

Di era digital saat ini, kemudahan akses informasi melalui internet dan media sosial telah mengubah cara masyarakat mendapatkan berita, namun juga meningkatkan penyebaran berita bohong atau *hoax* yang dapat menimbulkan ketidakpahaman dan keresahan. Penelitian ini bertujuan untuk mengatasi permasalahan penyebaran berita *hoax* dengan memanfaatkan teknologi *Machine Learning*, khususnya melalui metode klasifikasi teks. *Logistic Regression* digunakan sebagai model dasar yang dikenal efektif untuk tugas klasifikasi biner, dikombinasikan dengan teknik *Ensemble Learning* menggunakan *Bootstrap Aggregating* (Bagging). Metode ini bertujuan untuk meningkatkan akurasi model dengan mengurangi bias dan varian melalui pelatihan beberapa model pada subset data yang berbeda. Selain itu, penelitian ini mengeksplorasi pengaruh parameter *Logistic Regression*, seperti *Learning Rate*, Konvergensi, dan jumlah iterasi, terhadap akurasi model. TF-IDF digunakan sebagai metode ekstraksi fitur untuk mengubah teks berita menjadi vektor numerik yang dapat diproses oleh model. Dataset yang digunakan adalah berita *hoax* berbahasa Inggris dari Kaggle. Hasil penelitian ini diharapkan dapat menghasilkan sistem deteksi berita *hoax* yang akurat dan efisien, berguna bagi media, platform sosial, dan individu dalam mengidentifikasi serta mengurangi penyebaran informasi palsu. Selain itu, penelitian ini juga memberikan pengalaman bagi penulis dalam menerapkan metode *Ensemble Logistic Regression* dengan TF-IDF dalam pengembangan sistem deteksi yang lebih canggih.

Kata kunci: klasifikasi teks, berita *hoax*, *Logistic Regression*, Bagging, *Ensemble Learning*.

ABSTRACT

In today's digital era, easy access to information via the internet and social media has changed the way people get news, but has also increased the spread of fake news or hoaxes that can cause misunderstanding and unrest. This study aims to overcome the problem of the spread of hoax news by utilizing Machine Learning technology, especially through text classification methods. Logistic Regression is used as a basic model that is known to be effective for binary classification tasks, combined with Ensemble Learning techniques using Bootstrap Aggregating (Bagging). This method aims to improve model accuracy by reducing bias and variance through training several models on different data subsets. In addition, this study explores the effect of Logistic Regression parameters, such as Learning Rate, Convergence, and number of iterations, on model accuracy. TF-IDF is used as a feature extraction method to convert news text into numeric vectors that can be processed by the model. The dataset used is English-language hoax news from Kaggle. The results of this study are expected to produce an accurate and efficient hoax news detection system, useful for the media, social platforms, and individuals in identifying and reducing the spread of false information. In addition, this study also provides experience for the author in applying the Ensemble Logistic Regression method with TF-IDF in developing a more sophisticated detection system.

Keywords: text classification, hoax news, Logistic Regression, Bagging, Ensemble Learning.

DAFTAR ISI

HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN ORISINALITAS	iv
KATA PENGANTAR	v
ABSTRAK.....	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiii
DAFTAR PERSAMAAN	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.2.1 Permasalahan	3
1.2.2 Solusi.....	3
1.2.3 Pertanyaan Penelitian.....	4
1.3 Tujuan Penelitian	4
1.4 Manfaat Penelitian	4
1.5 Batasan Masalah	4
1.6 Sistematika Proposal.....	5
BAB II KAJIAN PUSTAKA	7
2.1 Berita <i>Hoax</i>	7
2.2 Klasifikasi Teks.....	7
2.3 <i>Pre-Processing</i>	8
2.3.1 <i>Case Folding</i>	8

2.3.2	<i>Punctuation Removal</i>	8
2.3.3	<i>Tokenization</i>	9
2.3.4	<i>Stop Word Removal</i>	9
2.3.5	<i>Stemming</i>	10
2.4	Ekstraksi Fitur	10
2.4.1	TF-IDF	10
2.5	<i>Logistic Regression</i>	11
2.5.1	<i>Log-Likelihood</i>	12
2.5.2	<i>Gradient Descent</i>	12
2.6	<i>Ensemble Learning</i>	13
2.6.1	<i>Bootstrap Aggregating</i>	13
2.7	Penelitian Terkait	14
BAB III METODOLOGI PENELITIAN		18
3.1	Arsitektur Sistem.....	18
3.2	Input Dataset	19
3.3	Teks <i>Preprocessing</i>	19
3.3.1	<i>Case Folding</i>	20
3.3.2	<i>Punctuation Removal</i>	20
3.3.3	<i>Tokenization</i>	21
3.3.4	<i>Stop word Removal</i>	21
3.3.5	<i>Stemming</i>	22
3.4	Ekstraksi Fitur.....	22
3.4.1	TF-IDF	22
3.5	Split Data	22
3.6	Bagging <i>Ensemble Learning</i>	23
3.6.1	Membuat Sampel Data.....	23

3.7 Model <i>Logistic Regression</i>	23
3.8 Uji Model.....	23
3.9 <i>Voting</i>	24
3.10 <i>Confusion Matrix</i>	24
3.11 Skenario Pengujian	24
BAB IV HASIL DAN PEMBAHASAN	28
4.1 Spesifikasi Perangkat	28
4.2 Pembuatan Program	30
4.2.1 Menginput Dataset	30
4.2.2 Mengubah kolom label menjadi angka pada dataset	31
4.2.3 Analisis Jumlah Sampel Berdasarkan Label.....	31
4.2.4 Preprocessing data	32
4.2.5 Ekstraksi fitur menggunakan TF-IDF	34
4.2.6 Pengujian model dengan Logistic Regression Sebelum Menggunakan Ensemble Learning.....	35
4.3 Hasil Skenario Pengujian.....	38
4.3.1 Jumlah Model Ensemble Learning	38
4.3.2 Nilai Konvergensi	39
4.3.3 Jumlah Iterasi	40
4.3.4 Learning Rate.....	41
BAB V KESIMPULAN DAN SARAN.....	43
5.1 Kesimpulan	43
5.2 Saran	43
REFERENSI	45
BIODATA PENULIS	48
LAMPIRAN.....	49

DAFTAR GAMBAR

Gambar 2.1 Bagging <i>Ensemble Learning</i>	14
Gambar 3.1 Desain Arsitektur Sistem.....	18
Gambar 3.2 Confusion Matrix	24
Gambar 4.1 Kode program input dataset	30
Gambar 4.2 Kode program mengubah label pada dataset	31
Gambar 4.3 Kode program analisis jumlah sampel berdasarkan label...	32
Gambar 4.4 Kode program preprocessing data.....	33
Gambar 4.5 Kode program Ekstraksi fitur dengan TF-IDF	34
Gambar 4.6 Kode program pengujian model dengan logistic regression tanpa ensemble learning.....	36
Gambar 4.7 Grafik Akurasi Berdasarkan Jumlah Model Ensemble	39
Gambar 4.8 Grafik Akurasi Berdasarkan Nilai Konvergensi	40
Gambar 4.9 Grafik Akurasi Berdasarkan Jumlah Iterasi	41
Gambar 4.10 Grafik Akurasi Berdasarkan Learning Rate	42

DAFTAR TABEL

Tabel 2.1 Penelitian Terkait	15
Tabel 3.1 Contoh data yang digunakan dalam penelitian.	19
Tabel 3.2 Contoh data yang telah di Preprocessing	22
Tabel 3.3 Tabel Skenario Pengujian.....	25
Tabel 4.1 Perangkat Keras	28
Tabel 4.2 Perangkat Lunak	28
Tabel 4.3 Library.....	29

DAFTAR PERSAMAAN

(2.1) Rumus TF	10
(2.2) Rumus IDF	11
(2.3) Rumus TF-IDF	11
(2.4) Rumus <i>Sigmoid</i>	11
(2.5) Rumus <i>Log-Likelihood</i>	12
(2.6) Rumus <i>Loss Function</i>	12
(2.7) Rumus <i>Gradient Descent</i>	12
(2.8) Rumus Turunan <i>Loss Function</i>	13
(2.9) Rumus Pembaruan Parameter <i>Gradient Descent</i>	13