**VARIANSI: Journal of Statistics and Its Application on Teaching and Research** **ISSN 2684-7590 (Online)**
**Vol. 5 No. 03 (2023), 155-162**

DOI: 10.35580/variansiunm116

# IMPLEMENTASI ANALISIS REGRESI LOGISTIK DENGAN METODE MACHINE LEARNING UNTUK MENGKLASIFIKASI BERITA DI INDONESIA,

**Muhammad Fahmuddin, Muhammad Kasim Aidid*, Muhammad Jabbar Taslim**

*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Makassar, Indonesia saja*

*Keywords:* Text Classification, Machine Learning, Berita

**Abstract:**

Perkembangan internet sangat pesat, internet menjadi sumber informasi yang mudah untuk diakses seperti halnya berita. Perkembangan ini selain membawa dampak yang positif tentu juga dampak yang negatif di dalamnya. Penelitian ini bertujuan untuk mengetahui hasil evaluasi dan tingkat akurasi klasifikasi berita di Indonesia dengan menggunakan analisis regresi logistik beserta metode *supervised learning*. Data yang digunakan diperoleh dari data.mendeley.com diantaranya berita dengan total berita 600. Setelah dilakukan *preprocessing* data, diperoleh jumlah kata dalam *dataset* sebanyak 104.020 kata. Setelah membagi *dataset* menjadi data latih sebanyak 80% atau 480 data dan data uji sebanyak 20% atau 120 data, diperoleh hasil akurasi dalam mengklasifikasi berita menggunakan analisis regresi logistik dengan metode *supervised learning* sebesar 78,3%.

## 1. Pendahuluan

Internet merupakan bagian penting dari hidup saat ini, sulit untuk bisa lepas dari penggunaan internet. Dengan perkembangan internet yang semakin pesat, internet menjadi sumber informasi yang mudah untuk diakses seperti halnya dengan berita.Perkembangan ini selain membawa dampak yang positif tentu terdapat juga dampak yang negatif di dalamnya. Perkembangan teknelogi informasi seperti media sosial dan berita yang terdapat di web turut serta mendorong penyebaran berita palsu dengan sangat mudah dan sangat cepat (Pan & Chiou, 2011).

Dampak buruk dari penyebaran berita palsu apabila tidak dicegah sedini mungkin diantaranya adalah mengakibatkan perpecahan antar kelompok, dapat memberikan reputasi buruk akan seseorang dan ada pihak yang diuntungkan dari kejadian tersebut, berita palsu juga dapat menjadikan masyarakat menjadi panik dari berita yang diberikan oleh penyebar berita palsu tersebut.

Penelitian terkait berita palsu pernah pada tahun 2019 lalu dengan menggunakan metode *Decision Stump, Logistic model Tree* dan J48 dengan mengklasifikasikan antara berita palsu dan berita asli, penelitian tersebut mendapatkan persentase hasil berturut-turut sebesar 56,4% untuk Decision Stump, 60,7% untuk Logistic Model Tree dan 55,8% untuk J48 (Ozbay & Alatas, 2020).

* Corresponding author.
*E-mail address*: kasimaidid@unm.ac.id

## 2. Tinjauan Pustaka

### *2.1. Machine Learning*

*Machine learning* merupakan salah satu diantara cabang ilmu yang merupakan kecerdasan buatan (*artificial intelligence*), dengan pemrograman untuk memungkinkan komputer memiliki kecerdasan dalam berperilaku sebagai mana manusia, dan dapat meningkatkan pemahamannya melalui pengalaman secara otomatis (Kusuma, 2020).

### *2.2. Regresi Logistik*

Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares* (OLS) *Regression*. Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah (Ristekdik, 2018).

#### *2.2.1. Fungsi Sigmoid Regresi Logistik*

Fungsi *sigmoid* regresi logistik merupakan fungsi aktivasi atau bisa disebut *squashing function*, dimana fungsi ini membatasi keluaran prediksi antara 0 dan 1 yang pada akhirnya menjadikan fungsi ini berguna dalam probabilitas prediksi. Berikut merupakan fungsi *sigmoid* dari regresi logistik. Dapat dilihat pada Persamaan 2.1. Fungsi *sigmoid* regresi logistik, sebagai berikut::

$$\text{logit}(p) = \ln\frac{p}{1-p} = B_0 + B_1X_1 + \ldots + B_nX_n \quad (2.1)$$

dimana:

$$\frac{p}{1-p} = e^{-(B_0+B_1X+\ldots+B_nX_n)} \quad (2.2)$$

$$p = \frac{1}{1+e^{-(B_0+B_1X+\ldots+B_nX_n)}} \quad (2.3)$$

#### *2.1.2. Text Preprocessing*

*Text Preprocessing* merupakan satu teknik dari *text mining*, dan dalam *text preprocessing* dilakukan untuk mengubah data tekstual yang tidak memiliki struktur menjadi data yang terstruktur lalu disimpan kedalam basis data (Dewa Krisdaynata, 2022). Tahapan yang ada dalam *preprocessing text*, diantaranya sebagai berikut:

#### *2.2.3. Case Folding*

Pada tahapan ini peneliti mampu mengubah atau mengkonversi *text* menjadi bentuk standar. Sehingga data masukan atau data primer akan diubah menjadi *lowercase* atau huruf kecil (Ratnawati, 2018).

#### *2.1.4. Tokenizing*

Pada tahapan ini untuk memudahkan peneliti dalam *preprocessing text* maka peneliti membagi atau memisahkan *text* menjadi token token (Yusra & Fikry, 2018).

*2.1.5. Filtering*

Pada tahap ini peneliti melakukan proses *filtering* dengan cara proses *stopwords*, proses s*topwords* akan menghapus kata-kata yang tidak memiliki arti yang signifikan atau biasa disebut *meaning less* seperti "ny", "gk", "dg", "sih", "hehe" dan lain-lain (Rian Hidayat, 2017).

*2.1.6. Pembobotan Kata*

Pembobotan kata pada penelitian ini menggunakan nilai TF dan IDF. *Term frequency* (TF) adalah metode sederhana dalam pembobotan setiap kata (Kurniawati, 2016). Sedangkan *Infers Document Frequency* (IDF) adalah perhitungan kata yang di distribusikan dalam sebuah dokumen (Astono dkk., 2019).

*2.1.7. Confusion Matrix*

*Confusion matrix* merupakan hasil dari proses klasifikasi berupa visualisasi data yang benar atau salah diprediksi. Untuk mengetahui hasil akurasi klasifikasi yang sesuai dengan *confusion matrix* digunakan persamaan berikut:

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \; x \; 100\% \qquad (2.4)$$

## 3. Metode Penelitian

*3.1. Jenis Penelitian*

Jenis penelitian yang digunakan peneliti adalah eksploratif kuantitatif. Penelitian eksploratif kuantitatif merupakan penelitian yang melakukan penjelajahan atau mengembangkan pengetahuan pada ilmu baru terkait dengan hal-hal tertentu, demi menyimpulkan suatu permasalahan yang rinti ataupun mengembangkan hipotesis dan bukan menguji hipotesis (Nderu,2021).

*3.2. Sumber Data*

Sumber data yang digunakan peneliti dalam penelitian ini berupa data sekunder yang diperoleh dari website data.mendeley.com dengan judul "Indonesian Hoax News Detection Dataset". Data terdiri dari 600 berita berbahasa Indonesia dari 12 topik berita yang telah diberi label valid dan hoax dari tiga orang penilai.

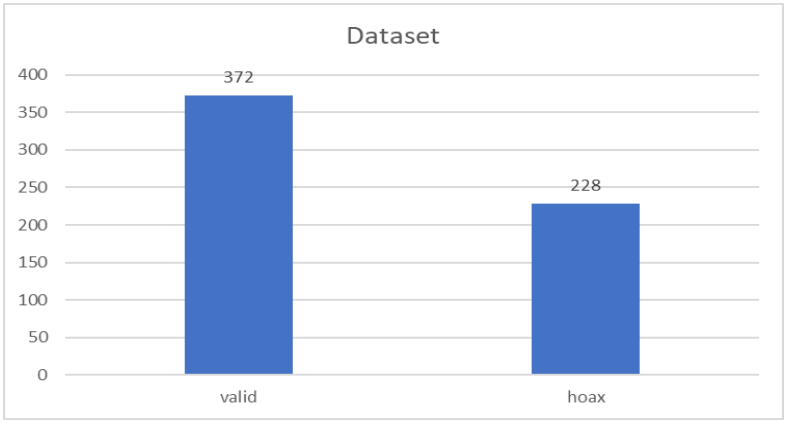
*3.3. Teknik Analisis Data*

Teknik analisis data yang dilakukan peneliti sebagai berikut:

1. Mengambil dataset yang berada di data.mendeley.com dengan judul "Indonesian Hoax News Detection Dataset"
2. Melakukan preprocessing data yang terdiri dari tiga tahapan, yaitu:
   a) Case folding, mengkonfersi teks menjadi bentuk standar.
   b) Tokenizing, memisah teks menjadi token token.
   c) Filtering, menghapus kata yang tidak memiliki arti.
3. Melakukan klasifikasi pada dataset yang telah diperoleh.
4. Term Weighting, Melakukan pembobotan kata menggunakan TD-IDF.
5. Melakukan analisis menggunakan metode regresi logistik dengan metode *supervised learnng*.

## 4. Hasil dan Pembahasan

Proses pengambilan data dilakukan dengan mengunduh dataset pada situs data.mendeley.com sebanyak 600 dataset berita dengan valid dan hoax. Jumlah berita yang berlabel valid terdapat 372 dan jumlah berita yang berlabel hoax terdapat 228.




**Gambar 4.1. Dataset**

Proses Preprocessing data dilakukan dengan beberapa tahapan, diantaranya,case folding, filtering tokenizing. Hasil dari tahap *case folding* dan *filtering*. Dapat dilihat pada Tabel 4.1. Hasil Case Folding dan Filtering, sebagai berikut:

**Tabel 4.1.** Hasil Case Folding dan Filtering

| Sebelum Preprocessing | Setelah Preprocesing |
|---|---|
| Iksan Korea Selatan Praktisi tusuk jarum boleh saja mengklaim berbagai manfaat dari pengobatan alternatif tersebut. Namun pada stroke pene litian membuktikan ternyata akupun-ktur tidak bisa memulihkan kondisi setelah terjadi serangan. Selain gagal memperbaiki kemampuan untuk me- … **Dapat dilihat pada Lamprian 1** | iksan korea selatan praktisi tusuk jarum mengklaim manfaat peng-obatan alternatif stroke penelitian membuktikan akupunktur memu-lihkan kondisi serangan gagal memperbaiki kemampuan aktvitas sehari akupunktur gagal mem-perbaiki fungsi neurologis saraf … **Dapat di-lihat pada Lamprian 3** |

Setelah preprocessing berhasil, dilakukan tahap tokenizing. Dapat dilihat pada Tabel 4.2. Hasil Tokenizing, sebagai berikut:

**Tabel 4.1.** Hasil Case Folding dan Filtering

| Sebelum Tokenizing | Setelah Tokenizing |
|---|---|
| iksan korea selatan praktisi tusuk jarum mengklaim manfaat peng-obatan alternatif stroke penelitian membuktikan akupunktur memu-lihkan kondisi serangan gagal memperbaiki kemampuan aktvitas sehari akupunktur gagal | ['iksan', 'korea', 'selatan', 'praktisi', 'tusuk', 'jarum', 'mengklaim', 'man-faat', 'pengobatan', 'alternatif', 'stroke', 'penelitian', 'membuktikan', 'akupu-nktur', 'memulihkan', 'kondisi', 'sera-ngan', 'gagal', 'memperbaiki', 'ke-mampuan', 'aktivitas', 'sehari', ' |

Setelah melakukan tokenizing selsai, dilakukan proses pembobotan kata. Dapat dilihat pada Tabel 4.3. Pembobotan Kata, sebagai berikut:

**Tabel 4.3.** *Term Frequency*

| **Kata** | | $\boldsymbol{tf_{ij}\,(t,d) = \frac{f_d(i)}{Nf_d(j)}}$ |
|---|---|---|
| $tf_{ij}(\text{“}jakarta\text{”}, 163)$ | = | $\frac{2}{163}$ |
| | = | 0.012269938650306749 |
| $tf_{ij}(\text{“}jejaring\text{”}, 163)$ | = | $\frac{1}{163}$ |
| | = | 0.006134969325153374 |
| $tf_{ij}(\text{“}sosial\text{”}, 163)$ | = | $\frac{1}{163}$ |
| | | 0.006134969325153374 |

Setelah memperoleh nilai TF seluruh kata pada 600 berita, dilakukan proses IDF. Dapat dilihat pada Tabel 4.4. Infers Document Frequency, sebagai berikut:

**Tabel 4.4.** *Infers Document Frequency*

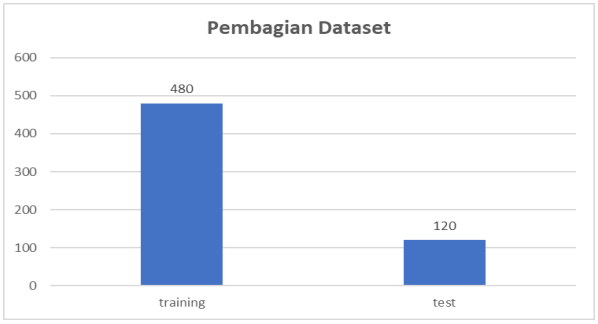
| **Kata** | | $\boldsymbol{idf(t,d) = \log\left(\frac{N}{df(t)}\right)}$ |
|---|---|---|
| $idf_{ij}(\text{“}jakarta\text{”}, 177)$ | = | $\log\left(\frac{600}{177}\right)$ |
| | = | 0.5277312480747496 |
| $idf_{ij}(\text{“}jejaring\text{”}, 33)$ | = | $\log\left(\frac{600}{33}\right)$ |
| | = | 1.2466723333413885 |
| $idf_{ij}(\text{“}sosial\text{”}, 222)$ | = | $\log\left(\frac{600}{222}\right)$ |
| | | 0.42984638733548297 |

Setelah memperoleh nilai TF-IDF seluruh kata pada 600 berita, dilakukan proses pembagian data dilakukan proses TF-IDF.

**Tabel 4.5.** *Term Frequency*

| **Kata** | | $tfidf(t,\,d,\,D) = tf(t,\,d) \times idf(t,\,D)$ |
|---|---|---|
| $tfidf(\text{“}jakarta\text{”}, 177{,}600)$ | = | 0.012269938650306749 X 0.5277312480747496 |
| | = | 0.006475230037726989 |
| $tfidf(\text{“}jejaring\text{”}, 33{,}600)$ | = | 0.006134969325153374 X 1.2466723333413885 |
| | = | 0.007648296523566801 |
| $tfidf(\text{“}sosial\text{”}, 222{,}600)$ | = | 0.006134969325153374 X 0.42984638733548297 |
| | = | 0.0026370944008311838 |

Sebelum melakukan analisis regresi logistik, peneliti membagi data latih dan data uji. Pada penelitian ini peneliti akan menggunakan data latih 80% dan data uji 20%. Data latih akan digunakan untuk memvalidasi model. Penelitian ini menggunakan *software* python untuk melakukan analisis regresi logistik dalam mengklasifikasi berita palsu di indonesia.




Dari proses klasifikasi akan diperoleh hasil yang akan disajikan dalam bentuk confusion matrix. Confusion matrix merupakan hasil dari proses klasifikasi berupa visualisasi dari berita yang diprediksi (Amrin & Saiyar, 2018). Pada penelitian ini berisi hasil klasifikasi berita hoax dan valid. Nilai akurasi dihitung dari perhitungan jumlah prediksi benar yang sesuai (TP) ditambah jumlah prediksi benar tidak sesuai (TN) dibandingkan dengan jumlah prediksi benar yang sesuai (TP), jumlah prediksi tidak sesuai (TF), jumlah prediksi salah yang sesuai (FP) dan jumlah prediksi salah tidak sesuai (FN) (Catur Supriyanto, 2013).

Pembagian dataset, menambahkan kolom ‘label_id’ dengan isinya merupakan hasil dari transformasi kolom ‘label’ menjadi angka 0 adalah valid dan 1 adalah hoax. Dapat dilihat pada **Tabel 4.6.** Penambahan kolom ‘label_id’, sebagai berikut:

**Tabel 4.6.** Penambahan kolom ‘label_id’

| No | berita | label | label _id |
|---|---|---|---|
| 1 | jejaring sosial beredar informasi menyebut lele ikan jorok sesuap daging ikan lele terkandung sel kanker julukan ikan jorok merujuk sifat lele doyan mengonsumsi jenis limbah perairan artikel viral internet kotoran manusia dijadikan pakan budidaya lele kota haikou china habitat aslinya lele catfish dikenal spesies ikan tangguh ikan dilengkapi alat … **Dapat dilihat pada Lamprian 1** | *valid* | *0* |
| | jejaring sosial beredar informasi lele ikan jorok sesuap daging ikan lele terkandung sel kanker kabar pembubidaya ikan lele jawa timur jatim tersinggung terpukul julukan ikan jorok merujuk sifat lele doyan mengonsumsi jenis limbah perairan artikel viral internet kotoran manusia dijadikan pakan budidaya lele … **Dapat dilihat pada Lamprian 1** | *hoax* | 1 |

Setelah penambahan kolom 'label_id' peneliti melakukan analisis regresi logistik dengan menganalisis data latih terlebih dahulu untuk mendapatkan hasil yang baik pada langah analisis data uji. Dapat dilihat pada **Tabel 4.7.** Hasil confusion matrix, sebagai berikut:

**Tabel 4.7.** *Confusion matrix*
*training test* 80:20

| **Kalasifikasi Prediksi** | **Klasifikasi Aktual** | |
|---|---|---|
| | **Valid** | **Hoax** |
| **Valid** | 27 | 13 |
| **Hoax** | 13 | 67 |

Berdasarkan **Tabel 4.7.** confusion matrix, dapat dilihat bahwa jumlah berita di indonesia dengan jumlah topik sebanyak 12 secara keseluruhan berhasil diklasifikasi secara benar terdapat 6, adapun berita benar terdapat 71. Untuk mendapatkan nilai akurasi menggunakan persamaan 2.6, sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN}\ x\ 100\%$$
$$= \frac{27 + 67}{27 + 13 + 13 + 67}\ x\ 100\%$$
$$= \frac{94}{120}\ x\ 100\%$$
$$= 0.783x\ 100\%$$
$$= 78{,}3\%$$

Dengan menggunakan persamaan untuk menghitung akurasi, didapatkan nilai akurasi sebesar 78,3%

## 5. Kesimpulan

Berdasarkan hasil evaluasi klasifikasi regresi logistik, proses pengunduhan dataset melalui situs data.mendeley.com. Dataset terdiri dari 600, dimana terdapat 372 berita dengan label valid dan 228 berita dengan label hoax. Sebelum melakukan analisis regresi logistik, dilakukan preprocessing terlebih dahulu. Preprocessing ini bertujuan untuk kelayakan dataset untuk menjadi data uji dan data latih. Setelah dilakukan preprocesinng, dataset dibagi menjadi 80% data latih dan 20% data uji. Setelah melakukan analaisis regresi logistik dengan metode *supervised learning* berdasarkan data latih dan data uji diperoleh dengan bentuk *confusion matrix*. Berdasarkan hasil tingkat akurasi diperoleh hasil akurasi dalam mengklasifikasi berita menggunakan analisis regresi logistik dengan metode *supervised learning* sebesar 78,3%.

## Daftar Pustaka

A. Yudi Permana. (2017). Implementasi Stemming Porter KBBI untuk Klasifikasi Topik Soal Ujian Nasional Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Journal of Chemical Information and Modeling*, *7*(1), 17–24.

Amrin, A., & Saiyar, H. (2018). Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes. *JURIKOM (Jurnal Riset Komputer)*, *5*(5), 498–502.

Catur Supriyanto, I. N. D. (2013). Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naïve Bayes. *Simantik 2013, 2013*(November), 156–160.

Dewa Krisdaynata, R. (2022). *Implementation of Self Training Classifier Using Logistic Regression in Classification of News Article Categories*. Universitas Multimedia Nusantara.

Eaton, E. (2008). Introduction to Machine Learning What is Machine Learning. *October*, 1–17.

Fikriya, Z. A., Irawan, M. I., & Soetrisno, S. (2017). Implementasi extreme learning machine untuk pengenalan objek citra digital. *Jurnal Sains Dan Seni ITS*, *6*(1), A1–A6.

Kurniawati. (2016). Term weighting berbasis indeks kelas menggunakan metode tf.idf.ics. *Term Weighting Berbasis Indeks Kelas Menggunakan Metode TF.IDF.ICSF Untuk Perengkingan Dokumen Al-Quran*. http://etheses.uin-malang.ac.id/3759/1/12650009.pdf

Kusuma, P. D. (2020). *Machine Learning Teori, Program, dan Studi Kasus*. Deepublish.

Mabruri, A. (2018). Produksi Program TV Drama Manajemen Produksi dan Penulisan Naskah. *Jakarta: PT. Gramedia Widiasarana Indonesia*.

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and Its Applications*, *540*, 123174.

Pan, L.-Y., & Chiou, J.-S. (2011). How much can you trust online information? Cues for perceived trustworthiness of consumer-generated online information. *Journal of Interactive Marketing*, *25*(2), 67–74.

Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *INOVTEK Polbeng-Seri Informatika*, *3*(1), 50–59.

Rian Hidayat. (2017). *Uin Syarif Hidayatullah Jakarta Uin Syarif Hidayatullah Jakarta*. *95*, 1–28. http://repository.uinjkt.ac.id/dspace/bitstream/123456789/ 33026/1/NITA FITRIANI-FKIK.pdf

Ristekdik, K. (2018). *Praktikum Data Mining*. 1–6.

Suryawati, I. (2016). *Jurnalistik Suatu Pengantar*.

Trisna Astono Putri, T., Warra, H. S., Yanti Sitepu, I., & Sihombing, M. (2019). Analysis and Detection of Hoax Contents in Indonesian News Based on Machine Learning. *Journal Of Informatics Pelita Nusantara*, *4*(1), 19–26. http://e-jurnal.pelitanusantara.ac.id/index.php/JIPN/article/view/489/291

Tsangaratos, P., & Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena*, *145*, 164–179.

Yusra, Y., & Fikry, M. (2018). Klasifikasi Tweet E-Commerce dengan Menggunakan Metode Support Vector Machine. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *4*(2), 50. https://doi.org/10.24014/coreit.v4i2.5205