Analisa Sentimen Pemilu 2019 Pada Judul Berita Online Menggunakan Metode Logistic Regression

Alifiyah Rohmatul Hidayati[1], Arief Senja Fitrani[2], Mochamad Alfan Rosid[3]

[1,2,3]Fakultas Sains dan Teknologi, Informatika, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia
Email: [1]halifiyahrohma@gmail.com, [2]asfjim@umsida.ac.id, [3]alfanrosid@umsida.ac.id

Abstract

Online news is a report that discusses an event packaged by the media as a means of publication in the form of news that can be accessed online. The 2019 election was one of the topics that was very much discussed at that time. In its implementation, the 2019 election reaped many critical notes related to the implementation and the issue of the integrity of the election itself. In this study, researchers took titles from various online news portals related to the 2019 election for sentiment analysis. The classification process is divided into three classes, namely positive, neutral and negative. The data used in this study amounted to 395 records. The stages carried out in this study are preprocessing which includes casefolding, remove punctuation, handling whitespace, stopword removal, stemming and tekonization. Next is handling imbalanced data to balance the number of classes. After going through the preprocessing stage, the next is the processing stage using the logistic regression method and randomized search cross validations. This is used to find the best parameters where the results of these parameters are then carried out by fitting the model. The results of the combined logistic regression method and randomized search cross validation show an accuracy score of 86%.

Keywords: *Analysis Sentiment, Online News, Classification, Logistic Regression, Pemilu 2019*

Abstrak

Berita online merupakan laporan yang membahas suatu peristiwa yang dikemas oleh media sebagai sarana publikasi berupa berita yang dapat diakses secara online. Pemilu 2019 merupakan salah satu topik yang sangat ramai diperbincangkan pada waktu itu. Pada pelaksanaannya, pemilu 2019 menuai banyak catatan kritis terkait penyelenggaraan hingga persoalan integritas pemilu itu sendiri. Pada penelitian ini, peneliti mengambil judul dari berbagai portal berita online yang berhubungan dengan pemilu 2019 untuk dilakukan Analisa sentiment. Proses klasifikasi dibagi menjadi tiga kelas yaitu positif, netral dan negatif. Data yang digunakan pada peelitian ini berjumlah 395 record. Adapun tahapan yang dilakukan pada penelitian ini yaitu preprocessing yang meliputi casefolding, remove punctuation, handling whitespace, stopword removal, stemming dan tekonization. Selanjutnya yaitu dilakukan handling imbalanced data untuk menyeimbangkan jumlah kelas. Setelah melalui tahapan preprocessing, selanjutnya yaitu tahapan processing dengan menggunakan metode logistic regression dan randomized search cross validation. Hal ini digunakan untuk mencari parameter terbaik dimana hasil dari parameter tersebut selanjutnya dilakukan fitting model. Hasil dari kombinasi metode logistic regression dan randomized search cross validation menunjukkan skor akurasi sebesar 86%.

Keywords: *Analisa Sentimen; Berita Online; Klasifikasi; Logistic Regression; Pemilu 2019*

1. Pendahuluan

Indonesia merupakan salah satu negara yang menganut sistem demokrasi [1]. Hal ini dibuktikan dengan diadakannya suatu pemilihan umum terhadap presiden dan wakil presiden. Pemilihan umum ini biasanya diselenggarakan secara periodik yaitu 5 tahun sekali. Tahun 2019 merupakan tahun dilaksanakannya pemilihan umum yang menuai banyak catatan kritis mulai dari penyelenggaraan, keakuratan data bahkan hingga persoalan integritas pemilu itu sendiri. Saat ini bidang jurnalistik telah terpengaruhi oleh perkembangan media social. Portal berita online menjadi suatu produk yang berasal dari perkembangan teknologi dalam dunia jurnalisme [2]. Kemunculan portal berita online ini sejalan dengan perkembangan audiens yang semakin dinamis dalam mencari informasi. Dengan adanya portal berita online yang semakin tumbuh membuat persaingan industri portal berita online menjadi semakin ketat.

Pada penelitian ini, peneliti melakukan Analisa sentiment terhadap judul berita online mengenai pemilihan umum yang dilaksanakan pada tahun 2019. Dimana pada masa pemilu tersebut banyak sekali warga yang membentuk beberapa kubu untuk mendukung kedua pasangan calon presiden. Adapun data yang diambil untuk penelitian ini berasal dari beberapa portal berita online yang diambil judulnya saja. Selanjutnya data tersebut dilakukan pra pemrosesan dengan beberapa tahapan. Setelah itu dilakukan pemrosesan menggunakan algoritma logstic regression. Logistic regression merupakan model statistik yang digunakan untuk mencari hubungan antara input dengan probabilitas hasil output [3]. Terdapat beberapa tipe logistic regression. Namun pada penelitian ini, peneliti

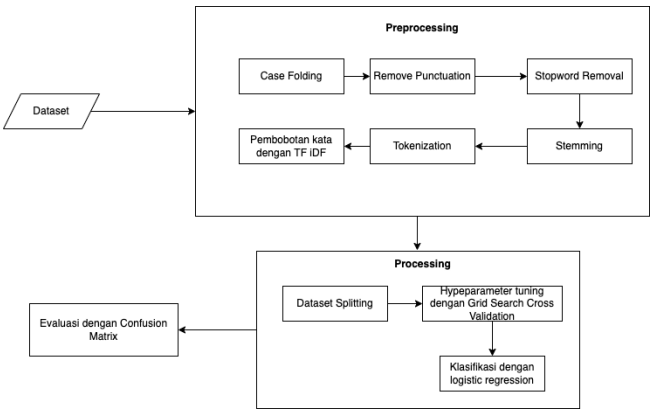
menggunakan multinomial logistic regression dimana penelitian ini mengklasifikasikan ke dalam 3 kelas berbeda yaitu positif, netral dan negatif.
Penelitian mengenai pemilu 2019 pernah dilakukan sebelumnya dengan data dan metode yang berbeda-beda. Salah satunya yaitu penelitian analisis sentimen yang dilakukan oleh Lia Durrotul Mahbubah dan Eri Zuliarso pada tahun 2019 yang berjudul "Analisa Sentimen Twitter Pada Pilpres 2019 Menggunakan Algoritma Naïve Bayes". Pada penelitian ini, peneliti mengambil tweets dari twitter dengan kata kunci #pilpres2019 dan #prabowo untuk diolah dan mengklasifikasikan teks tersebut menjadi du akelas yaitu kelas sentiment positif dan kelas sentimen negative. Data yang diolah berjumlah 300 tweets. Hasil dari penelitian ini menunjukkan bahwa klasifikasi data twitter dengan algoritma naïve bayes menghasilkan akurasi sebesar 73% [1].

Berdasarkan penjelasan yang telah dipaparkan, akan dilakukan penelitian tentang Analisa sentimen untuk mengklasifikasikan judul berita online tentang pemilu 2019. Data yang didapatkan dari portal berita online tersebut dibagi menjadi tiga kelas yaitu negative, netral dan positif. Dari data tersebut selanjutnya diproses dengan algoritma logistic regression. Tujuan dari penelitian ini yaitu untuk melihat sentiment masyarakat terhadap pelaksanaan pemilu 2019 membuahkan hasil yang positif, negative atau netral.

2. Metode Penelitian

2.1 Tahapan Penelitian

Tahapan penelitian merupakan gambaran umum terkait alur penelitian yang akan dilakukan dalam pengerjaan penelitian ini dari awal hingga akhir. Tahapan yang dilakukan dalam penelitian ini dapat dipaparkan melalui diagram alir seperti pada Gambar berikut.



Gambar 1 Tahapan Penelitian

2.2 Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan judul yang diambil dari portal berita online dengan topik pemilu 2019. Setelah data terambil, selanjutnya dilakukan labeling dengan tiga kelas yaitu positif, negatif dan netral. Proses labeling dilakukan berdasarkan sentimen terhadap judul dari portal berita online yang telah didapatkan.

2.3 Text Preprocessing

Tahapan selanjutnya yaitu preprocessing. Tahapan ini bertujuan untuk mengubah data teks yang tidak terstruktur menjadi data yang terstruktur. Preprocessing digunakan untuk menghindari dataset yang kurang sempurna [4]. Adapun tahapan preprocessing adalah sebagai berikut.

2.3.1 Case Folding

Tahapan preprocessing yang pertama yaitu case folding. Pada sebuah data, biasanya terdapat penulisan huruf yang tidak standar seperti adanya huruf kapital dan huruf kecil. Sehingga diperlukan adanya tahapan untuk mengkonversi semua huruf menjadi huruf kecil agar menjadi suatu bentuk standar. Tahapan tersebut disebut dengan case folding dimana case folding ini digunakan untuk mengkonversi alfabet ke huruf kecil atau lower case [5].

2.3.2 Remove Punctuation

Tahapan selanjutnya yaitu remove punctuation. Dimana tahapan ini digunakan untuk menghilangkan tanda baca atau symbol yang ada pada teks [6]. hal ini dikarenakan tanda baca atau simbol ini dihapus karena tidak berpengaruh pada hasil sentiment analisis. Sehingga penghapusan tanda baca atau simbol perlu dilakukan pada tahapan preprocessing.

2.3.3 Stemming

Selanjutnya yaitu tahapan stemming. Stemming merupakan proses yang digunakan untuk menemukan kata dasar dari sebuah kata dengan menghilangkan imbuhan yang terdiri dari awalan, sisipan, akhiran dan kombinasi dari awalan dan akhiran [7]. Tujuan dari tahapan stemming yaitu untuk memperkecil jumlah indeks yang berbeda dari satu data.

2.3.4 Pembobotan Kata

Tahapan ini digunakan untuk menaikkan kemampuan analisis sentiment pada proses text mining. Peneliti menggunakan metode Term Frequency – Inverse Document Frequency (TF-IDF). Proses pembobotan dilakukan dengan menghitung masing-masing nilai Term Frequency dan Inverse Document Frequency. TF menyatakan jumlah berapa banyak keberadaan suatu term dalam satu dokumen. Sedangkan IDF digunakan untuk mengurangi bobot suatu term jika kemunculannya banyak tersebar di seluruh koleksi dokumen [8].

2.4 Handling Imbalanced Data

Pada penelitian ini, kondisi data memiliki kelas yang tidak seimbang. Dimana jumlah kelas negatif terdapat 85 record, Netral 95 record dan Positif 215 record. Berdasarkan distribusi kelas tersebut, perlu dilakukan adanya handling imbalanced data. Peneliti menggunakan Teknik SMOTE untuk melakukan oversampling dimana kelas nantinya kelas netral dan positif akan dinaikan ukuran sampelnya menyesuaikan jumlah kelas negatif [9].

2.5 Processing

Setelah melalui tahapan preprocessing, tahapan selanjutnya yaitu processing dimana di dalam tahapan ini terdapat sub tahapan yaitu hyperparameter tuning dan klasifikasi dengan salah satu algoritma machine learning yaitu Logistic Regression

2.5.1 Hyperpameter tuning dengan Randomized Cross Validation

pada metode machine learning, terdapat beberapa nilai parameter yang diperkirakan dapat meningkatkan kinerja model yang disebut dengan hyperparameter. Metode hyperparameter yang akan diaplikasikan adalah Randomized Search Cross Validation. Metode ini fungsinya sama dengan Grid Search CV. Namun bedanya adalah Randomized Search CV dapat digunakan untuk mencari parameter terbaik dalam waktu yang lebih cepat dibandingkan Randomized Search CV. Teknik ini sangat berguna ketika parameter dan data yang dimiliki memiliki jumlah yang banyak [10].

2.5.2 Klasifikasi dengan Logistic Regression

Dataset hasil preprocessing selanjutnya dibagi menjadi data latih dan data uji dengan prosentase sebesar 80% untuk data latih dan 20% untuk data uji. Setelah melalui pembagian data dan hyperparameter tuning serta cross validation sebanyak 3 kali. selanjutnya dilakukan fitting model terhadap data latih. Data latih digunakan untuk membangun model klasifikasi sedangkan data uji digunakan untuk menguji model. Proses pemodelan dilakukan dengan menggunakan metode Logistic Regression.

2.6 Evaluasi

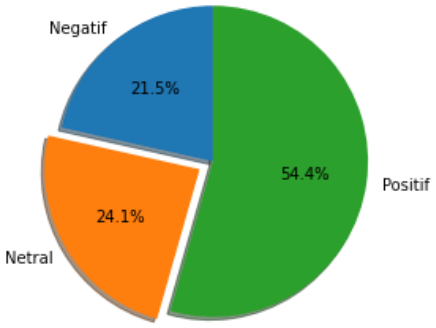
Tahapan ini digunakan untuk mengukur performa dari model *machine learning* yang telah dibuat. pada tahapan evaluasi, peneliti menggunakan *confusion matrix* untuk menganalisa performa model yang telah dibuat. confusion matrix merupakan pengujian performa untuk klasifikasi machine learning dimana hasil dari confusion matrix tersebut berupa dua kelas atau lebih. Confusion matrix merupakan suatu metode yang digunakan untuk melakukan perhitungan akurasi [11]. Prediksi kelas yang benar disebut dengan True Positif (TP), sedangkan prediksi kelas yang salah disebut dengan

False Positif (FP). Untuk kelas negative yang diprediksi benar disebut dengan True Negatif (TN), sedangkan kelas negative yang diprediksi salah disebut dengan False Negatf (FN).

3. Hasil dan Pembahasan

3.1 Analisa Data

Data masukan dari penelitian ini berupa judul dari berbagai portal berita online dengan topik pemilu 2019 Dengan 395 record yang selanjutnya dilakukan proses labeling. Adapun hasil dari proses labeling tersebut didapatkan label positif berjumlah 215 record, data dengan label negatif berjumlah 85 record dan data dengan label netral berjumlah 95 record.



Gambar 2 Perbandingan jumlah sentimen

Dari visualisasi data pada gambar 2 dapat diketahui bahwa persebaran data seimbang dimana sentiment positif terdapat 54.4% dari keseluruhan jumlah data. Sedangkan sentiment negative 21.5% dan netral 24.1%.

3.2 Text Preprocessing

Dalam proses text mining, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu sebelum dapat digunakan untuk proses utama. Proses mempersiapkan dataset mentah disebut juga dengan proses text preprocessing. Text preprocessing berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Preprocessing dilakukan untuk menghindari dataset yang kurang sempurna, terdapat noise pada dataset, data-data yang tidak konsisten dan mempercepat pemrosesan terhadap dokumen [5]. Pada bab sebelumnya telah dijelaskan mengenai alur text preprocessing yang meliputi case folding, remove punctuation, stemming, pembobotan kata. Berikut merupakan tabel hasil preprocessing.

Tabel 1 Hasil Preprocessing

Text	Hasil
Benarkah Polsek Tanjung Bumi Tak Gubris Laporan Tim Farid Alfauzi?	benar polsek tanjung bumi tak gubris lapo r tim farid alfauzi
Bila Pendaftaran Dibuka, Farid Alfauzi Sudah Bisa Mendaftar Pilkada Bangkalan	bila daftar buka farid alfauzi daftar pilkad a bangkal
Ratusan Baleho Bakal Calon Bupati Bangkalan Dirusak	ratus baleho bakal calon bupati bangkal r usak
Serikat Buruh Sepakat Deklarasi Damai Pemilu 2019	serikat buruh sepakat deklarasi damai pe milu 2019
Antisipasi Rawan Pelanggaran, Bawaslu Minta Masyarakat Aktif	antisipasi rawan langgar bawaslu minta m asyarakat aktif

3.2 Handling Imbalanced Data

berdasarkan gambar 2 dapat diketahui bahwa jumlah kelas pada penelitian ini tidak seimbang dimana kelas positif memiliki prosentase yang lebih banyak dibanding kelas negative dan netral. Sehingga perlu dilakukan handling imbalanced data agar model yang dihasilkan tidak condong ke satu kelas. Pada penelitian ini, peneliti melakukan oversampling menggunakan SMOTE. SMOTE merupakan Teknik yang digunakan untuk menyeimbangkan jumlah distribusi data pada kelas minoritas dengan

cara meyeleksi data sampel tersebut hingga jumlah datanya menjadi seimbang dengan jumlah kelas mayoritas [12]. Adapun hasil dari oversampling tersebut adalah berikut.

Tabel 2 Hasil Oversampling

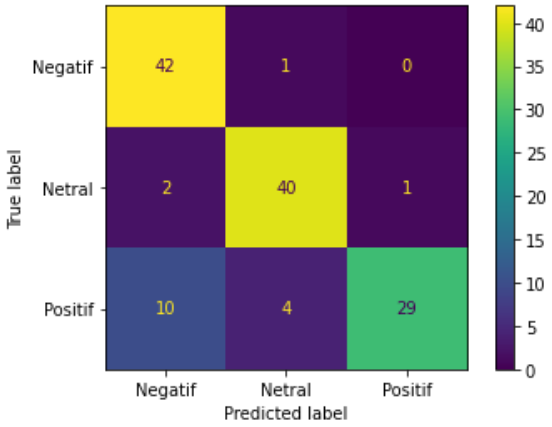
Kelas	Jumlah kelas sebelum oversampling	Jumlah kelas setelah oversampling
Positif	215	215
Negative	85	215
Netral	95	215

3.3 Klasifikasi dengan Logistic Regression

Setelah melalui seluruh tahapan text preprocessing dan oversampling, selanjutnya dilakukan klasifikasi dengan salah satu algoritma machine learning yaitu logistic regression. Data dibagi menjadi dua bagian yaitu data latih dan data uji dengan prosentase 80% untuk data latih dan 20% untuk data uji. setelah data dibagi menjadi dua bagian, selanjutnya peneliti melakukan hyperparameter tuning untuk mendapatkan parameter terbaik. Teknik yang digunakan untuk hyperparameter tuning adalah randomized search cross validation dimana teknik ini dapat melakukan tuning parameter dengan waktu yang lebih singkat dibanding grid search cross validation [13]. Hasil dari tuning tersebut mendapatkan parameter terbaik yaitu C=493.529 dan fit_intercept = False. Sehingga dari kombinasi metode tersebut didapatkan skor latih 98% dan skor uji 86%.

3.4 Evaluasi

Hasil dari tahapan klasifikasi selanjutnya dilakukan evaluasi untuk mengetahui seberapa besar skor akurasi, precision, recall dan f1 score. Berdasarkan gambar 4, skor akurasi yang dihasilkan pada penelitian ini adalah 86%. Untuk menampilkan confusion matrix, peneliti menggunakan library scikit learn dengan memanggil fungsi metrics. Sedangkan untuk menampilkan perhitungan dari confusion matrix, peneliti menggunakan library scikit learn dengan memanggil fungsi classification_report.



Gambar 3 Confusion Matrix

	precision	recall	f1-score	support
0	0.78	0.98	0.87	43
1	0.89	0.93	0.91	43
2	0.97	0.67	0.79	43
accuracy			0.86	129
macro avg	0.88	0.86	0.86	129
weighted avg	0.88	0.86	0.86	129

Gambar 4 Classification Report

4. Kesimpulan

Penelitian ini telah berhasil melakukan seluruh tahapan preprocessing hingga processing dan evaluasi. Hasil yang didapatkan pada penelitian ini mendapatkan nilai akurasi sebesar 86%. Namun

model yang dihasilkan pada penelitian ini masih mengalami overfitting karena skor uji yang dihasilkan memiliki gap yang cukup tinggi dengan skor latih dimana gap tersebut sebesar 12%. Dari hasil tersebut, pada penelitian selanjutnya diharapkan dapat ditemukan metode mulai dari tahapan preprocessing hingga processing yang dapat mengatasi kelemahan yang ada pada penelitian ini. Serta dapat menerapkan metode machine learning lain untuk menjadi tolak ukur antara algoritma logistic regression dengan algoritma machine learning lainnya.

Daftar Pustaka

[1] L. D. Mahbubah and E. Zuliarso, "Analisa Sentimen Twitter Pada Pilpres 2019 Menggunakan Algoritma Naive Bayes," *Sintak*, 2019.

[2] M. K. Anam, B. N. Pikir, and M. B. Firdaus, "Penerapan Na ̈ive Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen danPemeritah," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, 2021, doi: 10.30812/matrik.v21i1.1092.

[3] vincent michael, "Machine Learning: Mengenal Logistic Regression," *https://vincentmichael089.medium.com/machine-learning-2-logistic-regression-96b3d4e7b603*, May 09, 2019.

[4] J. Nasional, S. Informasi, H. Hakim, and S. Agustian, "Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter," vol. 03, pp. 107–114, 2022.

[5] Y. S. Nugroho and N. Emiliyawati, "Sistem klasifikasi variabel tingkat penerimaan konsumen terhadap mobil menggunakan metode random forest," *Jurnal Teknik Elektro*, vol. 9, no. 1, pp. 24–29, 2017.

[6] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *Jurnal Transformatika*, vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.

[7] A. Guterres, Gunawan, and J. Santoso, "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based," *Teknika*, vol. 8, no. 2, 2019, doi: 10.34148/teknika.v8i2.224.

[8] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: 10.24843/jim.2020.v08.i03.p04.

[9] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, p. 208, 2022, doi: 10.30865/mib.v6i1.3373.

[10] M. Rizky Mubarok, Muliadi, and R. Herteno, "Hyper-Parameter Tuning pada XGBoost Untuk Prediksi Keberlangsungan Hidup Pasien Gagal Jantung," *Kumpulan Jurnal Ilmu Komputer (KLIK)*, vol. 9, no. 2, pp. 391–401, 2022.

[11] B. P. Pratiwi, A. S. Handayani, and S. Sarjana, "PENGUKURAN KINERJA SISTEM KUALITAS UDARA DENGAN TEKNOLOGI WSN MENGGUNAKAN CONFUSION MATRIX," *Jurnal Informatika Upgris*, vol. 6, no. 2, 2021, doi: 10.26877/jiu.v6i2.6552.

[12] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, and U. Negeri Malang, "Terakreditasi SINTA Peringkat 2 Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *masa berlaku mulai*, vol. 1, no. 3, pp. 196–201, 2017.

[13] E. Agustin, A. Eviyanti, and N. Lutvi Azizah, "Deteksi Penyakit Epilepsi Melalui Sinyal EEG Menggunakan Metode DWT dan Extreme Gradient Boosting," vol. 7, no. 1, pp. 117–127, 2023, doi: 10.30865/mib.v7i1.5412.