UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

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

Oleh:

Alifiyah Rohmatul Hidayati,

Arief Senja Fitrani

Teknik Informatika

Universitas Muhammadiyah Sidoarjo

Maret, 2023

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.

Pada penelitian ini, peneliti melakukan Analisa sentiment terhadap judul berita online mengenai pemilihan umum yang dilaksanakan pada tahun 2019. 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].

Metode Penelitian

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.

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].

Metode Penelitian

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].

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.

Hasil

Handling Imbalanced Data

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].

Hasil

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%.

Pembahasan

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.

Referensi

[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.


UMSIDA
DARI SINI PENCERAHAN BERSEMI