JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3979
Hal 251−256
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression

Nur Ghaniaviyanto Ramadhan[1*], Faisal Dharma Adhinata[2], Alon Jala Tirta Segara[3], Diovianto Putra Rakhmadani[4]

[1,2,3] Fakultas Informatika, Rekayasa Perangkat Lunak, Institut Teknologi Telkom Purwokerto, Indonesia
[4] Fakultas Rekayasa Industri dan Desain, Bisnis Digital, Institut Teknologi Telkom Purwokerto, Indonesia
Email: [1*]ghani@ittelkom-pwt.ac.id, [2]faisal@ittelkom-pwt.ac.id, [3]alon@ittelkom-pwt.ac.id, [4]diovianto@ittelkom-pwt.ac.id
Email Penulis Korespondensi: ghani@ittelkom-pwt.ac.id
Submitted **16-03-2022**; Accepted **13-04-2022**; Published **29-04-2022**

Abstrak

Berita palsu merupakan sebuah informasi yang disajikan secara tidak benar atau bohong. Tentunya jika penyebaran berita palsu terus terjadi maka dapat mengakibatkan salah pengetahuan informasi yang didapat. Salah satu upaya untuk mencegah penyebaran berita palsu yaitu dengan cara melakukan deteksi berita apakah berita itu asli atau palsu agar dapat memberikan penjelasan kepada pembaca berita terkait. Pada penelitian ini bertujuan untuk deteksi berita palsu menggunakan model *supervised learning* random forest. Dataset berita yang digunakan berisi 6256 baris judul yang memiliki kelas *fake* atau *real*. Dataset terlebih dahulu melalui proses cleaning, tokenisasi, dan stemming untuk memecah kalimat menjadi sebuah kata. Hasil yang didapatkan menggunakan model random forest sebesar 84%, hasil tersebut lebih tinggi dibandingkan menggunakan model logistic regresi sebesar 77%.

Kata Kunci: Berita; Deteksi; Random Forest; Supervised Learning

Abstract

Fake news is information that is presented incorrectly or falsely. Of course, if the spread of fake news continues, it can result in wrong knowledge of the information obtained. One of the efforts to prevent the spread of fake news is by detecting whether the news is genuine or fake in order to provide an explanation to the readers of the related news. This study aims to detect fake news using a supervised learning random forest model. The news dataset used contains 6256 rows of titles that have a fake or real class. The dataset first goes through a cleaning, tokenization, and stemming process to break sentences into words. The results obtained using the random forest model of 84%, this result is higher than using the logistic regression model of 77%.

Keywords: News; Detection; Random Forest; Supervised Learning

1. PENDAHULUAN

Berita palsu merupakan sebuah informasi yang disajikan secara tidak benar secara fakta atau bohong. Penyebaran berita saat ini sangat cepat dapat kita ketahui. Tentunya jika penyebaran berita palsu terus terjadi maka dapat mengakibatkan salah pengetahuan informasi yang didapat. Misalkan berita palsu mengenai covid 19 dan kebijakan pemerintah. Hal tersebut sangat membahayakan bagi generasi yang akan datang jika tidak dicegah. Salah satu upaya untuk mencegah penyebaran berita palsu yaitu dengan cara melakukan deteksi berita apakah berita itu asli atau palsu agar dapat memberikan penjelasan kepada pembaca berita terkait. Saat ini sudah ada beberapa penelitian terkait deteksi berita palsu dengan menggunakan model pembelajaran mesin dan kecerdasan buatan. Adanya penelitian tersebut tentunya sangat membantu para pembaca berita untuk menghindari berita palsu yang dibaca. Beberapa penelitian akan dibahas mengenai tujuan dilakukannya penelitian, seperti pada penelitian yang dilakukan oleh Supanya membahas tentang deteksi berita palsu menggunakan model pembelajaran mesin *Naïve Bayes*, SVM, dan *Neural Network* [1].

Penelitian lainnya dilakukan untuk deteksi berita bohong supaya menghindari dampak negatif bagi pengguna media social [2]. Penelitian Abdullah mengusulkan model untuk mengenali pesan berita palsu dari postingan twitter dengan mencari tahu bagaimana mengantisipasi penilaian presisi, mengingat komputerisasi identifikasi berita palsu di dataset Twitter dengan menggunakan model *naïve bayes*, *logistic regression*, dan *reccurent neural network* [3]. Penelitian Yang Liu mengusulkan model baru untuk deteksi dini berita palsu di media sosial melalui klasifikasi jalur propagasi berita dengan model *Convolutional Neural Network* (CNN) dan *Reccurent Neural Network* (RNN) [4]. Penelitian Calo membahas tentang klasifikasi berita palsu berdasarkan subjek bahasanya [5]. Penelitian Souvick mengusulkan metode umum berdasarkan Deep Neural Networks untuk mendeteksi apakah berita yang diberikan palsu atau asli [6].

Penelitian lain membahas tentang klasifikasi artikel berita palsu terkait COVID-19 dengan menggunakan *deep learning* [7]. Penelitian Herley mengusulkan pendekatan pembelajaran *ensemble* pada berita palsu Indonesia untuk memisahkan berita palsu dari yang asli dan untuk mengatasi masalah data yang tidak seimbang yang kita hadapi pada dataset yang diberikan dengan menggunakan model random forest [8]. Penelitian Marcos bertujuan untuk deteksi berita palsu dengan menggunakan model data mining berdasarkan pola text yang dihasilkan [9]. Penelitian Pedro mengusulkan untuk mendeteksi berita palsu dengan melatih model dengan hanya sampel palsu dalam kumpulan data pelatihan, melalui Klasifikasi Satu Kelas (OCC) [10]. Pada makalah lain menyajikan solusi untuk tugas deteksi berita palsu dengan menggunakan arsitektur *deep learning* [11]. Makalah lainnya mengusulkan sistem deteksi berita palsu berdasarkan klasifikasi seperti *Logistic Regression* (LR), *Naïve Bayes* (NB), *Support Vector Machine* (SVM), *Random Forest* (RF) dan *Deep Neural Network* (DNN) [12]. Pada penelitian lain menggunakan dataset berita yang berisikan 600 judul menggunakan model naïve bayes [13]. Study lain membahas tentang deteksi berita palsu dengan menggunakan berbagai model berbasis deep learning [14]. Penelitian yang dilakukan Amiri membahas tentang deteksi berita palsu menggunakan

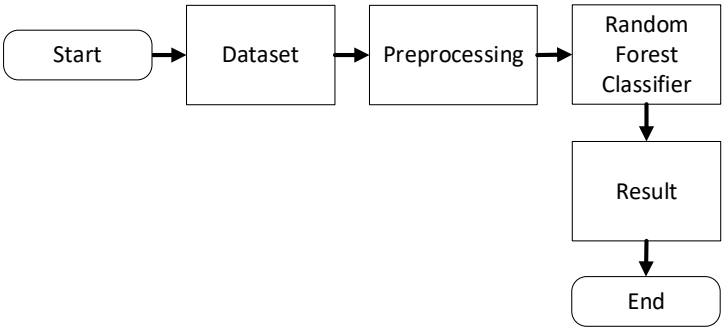
JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3979
Hal 251−256
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

algortima BERT [15]. Study yang dilakukan Setiawan membahas tentang deteksi valudasi berita pada media social twitter berbasis naïve bayes [16].

Berdasarkan penejelasan permasalahan di atas dan pembahasan beberapa penelitian sebelumnya maka pada penelitian ini bertujuan untuk deteksi berita palsu menggunakan model supervised learning random forest.

2. METODOLOGI PENELITIAN

Pada penelitian ini menggunakan alur sistem pada Gambar 1 di bawah ini.



Gambar 1. Alur Sistem Penelitian

2.1 Dataset

Penelitian ini menggunakan dataset yang berisikan atribut judul berita, isi berita, dan kategori. Dataset didapatkan melalui kaggle [17]. Pada tabel 1 merupakan contoh bentuk dataset yang digunakan. Total judul berita yang ada pada dataset sebanyak 6560 baris.

Tabel 1. Contoh Data Berita Fake atau Real

Berita	Kategori
Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Center, is a New York writer focusing on radical Islam.	Fake
Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Pocket Tumblr There are two fundamental truths in this world: Paul Ryan desperately wants to be president. And Paul Ryan will never be president.	Fake
U.S. Secretary of State John F. Kerry said Monday that he will stop in Paris later this week	Real
Women also do just as well in fundraising as men, even if they have to work harder to raise the same amount.	Real
Former Florida governor Jeb Bush last week became the latest Republican to signal a readiness to engage Democrats on what historically has been their turf, putting issues of middle-class wage stagnation, poverty and shared prosperity at the forefront of their political messages	Real

2.2 Preprocessing

Pada tahap ini dilakukan pembersihan dataset dari symbol-simbol dan karakter, tokenisasi, dan stemming. Tokenisasi adalah proses memecah aliran teks menjadi kata, frasa, simbol, atau elemen bermakna lainnya yang disebut token [18][19]. Stemming adalah proses tanpa variasi bentuk kata menjadi bentuk umum yang representative [19]. Pada tabel 2 merupakan hasil penerapan tokenisasi dari kalimat di dalam dataset, sedangkan pada tabel 3 merupakan hasil penerapan stemming.

Tabel 2. Hasil Penerapan Stemming

Berita	Kategori
“Daniel”, “Greenfield”, “a”, “Shillman”, “Journalism”, ”Fellow”, ”at”, “the”, “Freedom”, “Center”, “is”, “a”, “New” “York”, “writer”, “focusing”, “on”, “radical”, “Islam”	Fake
”Women”, ”also”, ”do”, ”just”, ”as”, ”well”, ”in”, ”fundraising”, ”as”, ”men”, ”even”, ”if”, ”they”, ”have”, ”to”, ”work”, ”harder”, ”to”, ”raise”, ”the”, ”same”, ”amount”	Real

Tabel 3. Hasil Penerapan Stemming

Berita	Kategori
“Daniel”, “Greenfield”, “a”, “Shillman”, “Journalis”, ”Fellow”, ”at”, “the”, “Freedom”, “Center”, “is”, “a”, “New” “York”, “writer”, “focus”, “on”, “radical”, “Islam”	Fake
”Women”, ”also”, ”do”, ”just”, ”as”, ”well”, ”in”, ”fundraise”, ”as”, ”men”, ”even”, ”if”, ”they”, ”have”, ”to”, ”work”, ”hard”, ”to”, ”raise”, ”the”, ”same”, ”amount”	Real

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3979
Hal 251−256
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

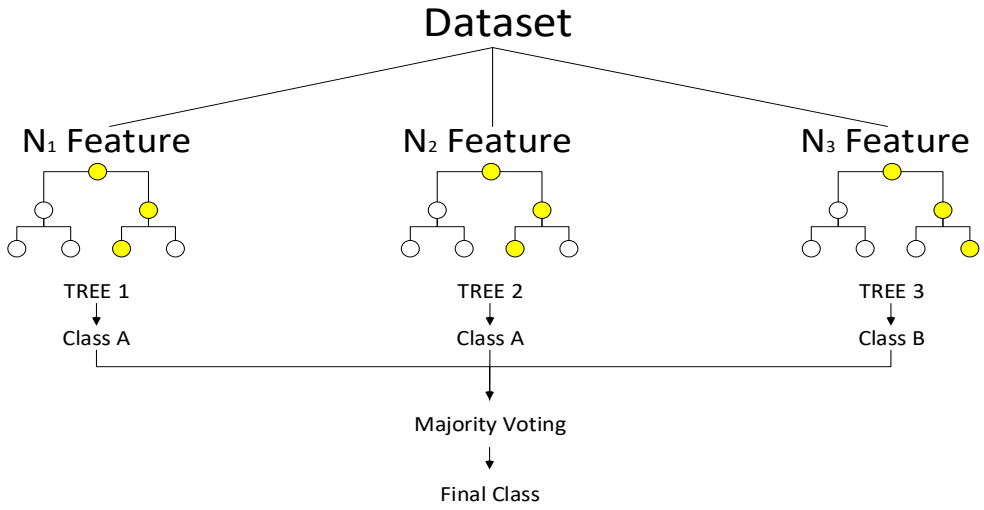
2.3 Random Forest

Random forest adalah metode pembelajaran ensemble untuk klasifikasi atau regresi yg beroperasi menggunakan menciptakan banyak pohon keputusan selama proses *training* dan menaruh hasil berupa mode kelas (klasifikasi) atau prediksi rata-rata (regresi) pohon individu [20]. Selain itu random forest merupakan model klasifikasi supervised learning yang banyak digunakan diberbagai ilmu disiplin klasifikasi [21].

$$Entropy = \sum_{i}^{c} Pi \log 2 \ Pi \quad (1)$$

$$Gini = 1 - \sum_{i}^{c} P^2 i \quad (2)$$

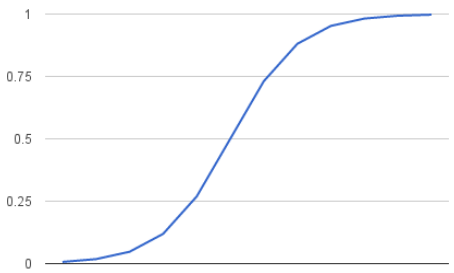
Pada proses deteksi ini menggunakan formula entropy (information gain) dan gini (impurity) seperti pada formula (1) dan (2). Dimana proses deteksi berlangsung dalam bentuk pohon keputusan ke bawah, seperti pada gambar 2.



Gambar 2. Split Random Forest

2.3 Logistic Regression

Logistic Regression merupakan model linier yang terdiri dari satu atau sebagian variabel bebas yang menggambarkan hubungan dengan variabel respon terikat [22]. Salah satu model statistik linier yang paling tidak universal digunakan untuk analisis diskriminan adalah *Logistic Regression*.



Gambar 3. Logistic Regression

Fungsi sigmoid (2) merupakan fungsi aktivasi atau bisa disebut dengan squashing function, dimana fungsi ini membatasi keluaran prediksi antara 0 dan 1 yang pada akhirnya menjadikan fungsi ini berguna dalam probabilitas prediksi.

$$\sigma(Z) = \frac{1}{1+ exp^{-z}} \quad (3)$$

$$p\left(y^{(i)} = 1 \middle| x^{(i)}, w\right) = 1 - \frac{1}{1+ exp^{\left(w^T x^{(i)} + b\right)}} \quad (4)$$

$$p\left(y^{(i)} = 0 \middle| x^{(i)}, w\right) = 1 - \frac{1}{1+ exp^{\left(w^T x^{(i)} + b\right)}} \quad (5)$$

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3979
Hal 251−256
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

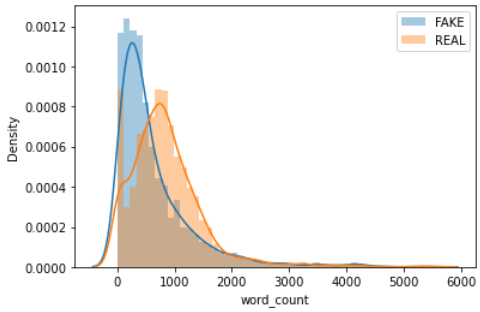
Untuk respon data biner, model regresi logistik dapat diekspresikan dengan menjumlahkan kombinasi linier dari fitur input dan bobot yang sesuai (w) ditambah suku bias (b) untuk setiap contoh seperti yang ditunjukkan pada persamaan (3), (4), dan (5).

3. HASIL DAN PEMBAHASAN

Setelah dilakukan eksperimen terhadap dataset dengan model random forest maka untuk mengukur hasil yang didapat menggunakan formula (6).

$$Akurasi = \frac{TP+TN}{TP+TN+FN+FP} \quad (6)$$

True Positif (TP) merupakan prediksi data benar sesuai dengan faktanya juga benar. True Negatif (TN) merupakan prediksi data salah sesuai dengan faktanya juga salah. False Positif (FP) merupakan prediksi data salah tetapi faktanya benar. False Negatif (FN) merupakan prediksi data benar tetapi faktanya salah.

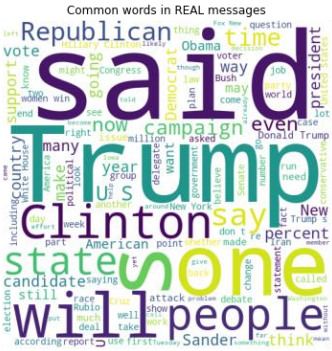


Gambar 4. Density Kelas

Gambar 4 merupakan persebaran nilai *density* terhadap persebaran kelas *fake* atau *real* yang dimana dapat dilihat untuk density kelas *fake* grafiknya cenderung lancip kea rah atas hingga menyentuh angka 0.0012. sedangkan untuk density kelas *real* grafiknya cenderung agak landai di angka 0.0008. Hal tersebut membuktikan bahwa berita dengan kelas *fake* memiliki dominasi daripada kelas *real*.

Gambar 5. Kata yang FAKE

Gambar 5 merupakan *word count* untuk kata dari kelas *fake*. Kata-kata tersebut tentunya setelah dilakukan proses tokenisasi dan stemming sehingga hanya menjadi satu kata saja.



Gambar 6. Kata yang REAL

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 2, April 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i2.3979
Hal 251−256
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Gambar 6 merupakan *word count* untuk kata dari kelas *real*. Kata-kata tersebut sama seperti gambar 4 yang telah dilakukan proses tokenisasi dan stemming. Akan tetapi terlihat bahwa ada kata yang sama untuk kelas *fake* dan *real* misalnya seperti kata *said*, *people*, dan Trump. Hal demikian nantinya jika dilakukan deteksi bergantung kepada nilai density yang dihasilkan sebelumnya.

Tabel 4. Confusion Matrix Random Forest (Entropy)

		Aktual	
		Benar	Salah
Prediksi	Benar	924	165
	Salah	120	611

Tabel 5. Confusion Matrix Random Forest (Gini)

		Aktual	
		Benar	Salah
Prediksi	Benar	894	169
	Salah	128	513

Pada tabel 4 dan 5 merupakan confusion matrix dari eksperimen yang dilakukan menggunakan model random forest. Pada penelitian ini menggunakan ratio 70:30 untuk split dataset. Data training 70%, data testing 30% dari total baris data yang ada pada dataset.

Tabel 6. Hasil Akurasi

Model	Akurasi
Random Forest (Entropy)	84%
Random Forest (Gini)	83%
Logistic Regression	77%

Pada tabel 6 merupakan hasil akurasi yang didapatkan dengan menerapkan model usulan random forest penelitian ini sebesar 84%. Dibandingkan dengan menggunakan model logistic regression lebih tinggi 7%. Hal tersebut membuktikan bahwa model berbasis ensemble mampu melakukan deteksi hingga mendalam. Akan tetapi jika dilihat hasil FP dan FN pada tabel 4 masih berada pada angka di atas 100, hal tersebut menandakan masih ada banyak judul berita yang belum mampu di deteksi oleh model random forest. Hal lainnya jika dilihat dari hasil TP dan TN nya pada tabel 4 model random forest sudah baik melakukan deteksi berita palsu, jika dijumlahkan antara TP dan TN maka sudah ada 1535 text berita yang mampu di deteksi dengan tepat. Untuk mengurangi nilai FP dan FN bisa dilakukan dengan cara menambah jumlah pohon pada parameter random forest untuk melakukan proses deteksi. Akurasi berbeda juga terlihat dengan menggunakan entropy dan gini. Perbedaan hasil akurasi sebesar 1% namun demikian jika dilihat dari confusion matrix (tabel 5) yang dihasilkan tentunya sangat signifikan perubahan yang didapatkan.

4. KESIMPULAN

Berdasarkan eksperimen yang dilakukan untuk mendeteksi berita tergolong ke kelas *real* atau *fake* pada penelitian ini telah mampu melakukan deteksi dengan akurasi yang didapatkan sebesar 84% menggunakan model usulan berbasis ensemble learning random forest. Pada penelitian ini juga membandingkan hasil dengan model lain yang digunakan yaitu logistic regression, hasil menunjukkan model random forest lebih unggul sebesar 7%. Nilai density tertinggi sebesar 0.0012 didapatkan oleh kelas *fake*, sedangkan density untuk kelas *real* tertinggi sebesar 0.0008. Hasil akurasi random forest entropy lebih tinggi 1% daripada random forest gini. Nilai FP dan FN masih tinggi jika dilihat dari confusion matrix maka perlu dilakuan penambahan nilai parameter pada pohon keputusan, selain itu bisa menggunakan model berbasis *deep learning*.

REFERENCES

[1] Aphiwongsophon, Supanya, and Prabhas Chongstitvatana. "Detecting fake news with machine learning method." 2018 15th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON). IEEE, 2018.
[2] Ksieniewicz, Paweł, et al. "Machine learning methods for fake news classification." International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham, 2019.
[3] Mahir, Ehesas Mia, Saima Akhter, and Mohammad Rezwanul Huq. "Detecting fake news using machine learning and deep learning algorithms." 2019 7th International Conference on Smart Computing & Communications (ICSCC). IEEE, 2019.
[4] Liu, Yang, and Yi-Fang Wu. "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[5] Jeronimo, Caio Libanio Melo, et al. "Fake news classification based on subjective language." Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. 2019.

[6] Ghosh, Souvick, and Chirag Shah. "Towards automatic fake news classification." Proceedings of the Association for Information Science and Technology 55.1. 805-807. 2018.
[7] Koirala, Abhishek. "COVID-19 fake news classification with deep learning." Preprint 2020.
[8] Al-Ash, Herley Shaori, et al. "Ensemble learning approach on indonesian fake news classification." 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). IEEE, 2019.
[9] Moraes, Marcos Paulo, Jonice de Oliveira Sampaio, and Anderson Cordeiro Charles. "Data mining applied in fake news classification through textual patterns." Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. 2019.
[10] Faustini, Pedro, and Thiago Covões. "Fake news detection using one-class classification." 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2019.
[11] Thota, Aswini, et al. "Fake news detection: a deep learning approach." SMU Data Science Review 1.3. 10. 2018.
[12] Hiramath, Chaitra K., and G. C. Deshpande. "Fake news detection using deep learning techniques." 2019 1st International Conference on Advances in Information Technology (ICAIT). IEEE, 2019.
[13] Rahutomo, Faisal, et al. "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia Naïve Bayes's Experiment On Hoax News Detection In Indonesian Language." vol 23. 1-15. 2019.
[14] Yunanto, Rio, Apriani Puti Purfini, and Angga Prabuwisesa. "Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning." Jurnal Manajemen Informatika (JAMIKA) 11.2. 118-130. 2021.
[15] Amiri, Aufa Nabil. Deteksi Berita Palsu Otomatis Berbahasa Indonesia Menggunakan BERT. Diss. Institut Teknologi Sepuluh Nopember, 2021.
[16] Setiawan, Esther Irawati, et al. "Deteksi Validitas Berita pada Media Sosial Twitter dengan Algoritma Naive Bayes." Journal of Intelligent System and Computation 3.2. 55-60. 2021.
[17] Fake or Real News, Available online: https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news, Diakses 1 April 2022.
[18] Kannan, Subbu, et al. "Preprocessing techniques for text mining." International Journal of Computer Science & Communication Networks 5.1. 7-16. 2014.
[19] Ramadhan, Nur Ghaniaviyanto. "Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor." Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi) 5.6. 1083-1089. 2021.
[20] Ramadhan, Nur Ghaniaviyanto, and Faisal Dharma Adhinata. "Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara." RADIAL: Jurnal Peradaban Sains, Rekayasa dan Teknologi 9.2. 125-134. 2021.
[21] Ramadhan, Nur Ghaniaviyanto, Ade Romadhony, and Adiwijaya. "Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest.". 2021.
[22] Kirasich, Kaitlin, Trace Smith, and Bivin Sadler. "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets." SMU Data Science Review 1.3. 9. 2018.