Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332

# Sistem Deteksi Berita Hoax Berbahasa Indonesia Bidang Kesehatan

[1]*Trisna Ari Roshinta, [2] Elya Kumala, [3] Ivan Vausta Dinata,
[1,2,3]Sekolah Vokasi, Universitas Sebelas Maret
Surakarta, Indonesia

[1]trisna.roshinta@staff.uns.ac.id, [2]elya.kumala@student.uns.ac.id,
[3]ivan.fausta@student.uns.ac.id

*Penulis Korespondensi
Diajukan : 15/04/2023
Diterima : 27/04/2023
Dipublikasi : 27/04/2023

## ABSTRAK
Berita palsu (hoax) menyebarkan informasi yang salah atau sangat berbeda dengan fakta yang ada, dimana maksud utamanya adalah untuk memanipulasi dan menipu pembaca. Pertumbuhan internet yang eksponensial, sebanding dengan penyebaran berita hoax yang juga sangat cepat. Saat ini hampir susah untuk membedakan berita yang fakta dan hoax. Hoax menyebabkan banyak kerugian sosial dan nasional dengan dampak destruktif. Misalnya, penelitian medis di Taiwan menyatakan bahwa hoax mengenai vaksin COVID-19 mengurangi kepercayaan vaksin di masyarakat sehingga serapa dosis vaksin tidak sesuai target. Banyak masyarakat menjadi ragu bahkan menjadi anti vaksin. Oleh karena itu, pengendalian dan pendeteksian penyebaran berita hoax sangat mendesak dilakukan, terutama bidang kesehatan. Dalam mendeteksi berita hoax, metode yang dapat digunakan adalah machine learning, khususnya untuk klasifikasi teks. Makalah ini bertujuan untuk merancang dan membangun sistem deteksi hoax untuk berita bersahasa Indonesia khususnya bidang kesehatan dengan algoritma klasifikasi Naive Bayes. Metode ini menggunakan data dari hasil crawling sebagai data training Akurasi dari sistem ini mencapai 90,9%. Sistem ini dibangun dengan menggunakan bahasa CI 4 dengan metode pengembangan Waterfall. Sistem ini telah digunakan oleh dinas Kabupaten Madiun dalam memberikan informasi hoax bagi masyarakat.

**Kata Kunci**: Filtering Hoax, Naive Bayes, Machine Learning

## I. PENDAHULUAN
Ada kutipan dari Mark Twin bahwa "kebohongan dapat menyebar ke belahan dunia lain dengan begitu cepat, sementara kebenaran mengenakan sepatunya". Kutipan ini menjelaskan bahwa kebohongan dapat dengan cepat menyebar ke seluruh penjuru dunia, sedangkan kebenaran terkadang menghilang begitu saja. Era teknologi seperti saat ini, telah menjadi jembatan penyebaran berita palsu (hoax) yang begitu cepat dan terkadang tidak terkendali.

Berita hoax melaporkan informasi palsu yang bertujuan untuk menyesatkan pembaca. Contoh nyata bagaimana hoax dapat membuat kerugian di masyarakat yaitu, penelitian medis di Taiwan menunjukkan bahwa hoax tentang vaksin COVID-19 mengurangi jumlah dosis yang diserap masyarakat karena mereka yang terpapar berita hoax menjadi ragu-ragu dan bahkan anti vaksin (Chen et al., 2022). Contoh lainnya adalah banyaknya berita hoax tentang lowongan kerja yang berujung pada penipuan untuk mendapatkan uang dari calon pelamar. Oleh karena itu, deteksi dini otomatis terhadap berita hoax di internet menjadi sangat penting.

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332

Deteksi otomatis berita hoax, terutama di situs publik atau pemerintah, sangat penting. Namun, dalam penerapannya, penting juga untuk menemukan metode terbaik dalam menentukan apakah suatu berita hoax atau tidak. Penelitian ini berfokus untuk mencari algoritma dan mengimplementasikan pada berita Indonesia. Karakter berita bahasa Indonesia memiliki perbedaan karena istilah kosa kata dan penggunaan fake news berbeda (Prasetijo et al., 2017).

Studi deteksi hoax untuk mencari metode terbaik dalam menentukan berita hoax atau nyata telah dilakukan dan diterbitkan dalam literatur. Hasil terbaru yaitu membandingkan metode SVM dan SGD, dengan hasil SGD yang lebih baik (Prasetijo et al., 2017). Penelitian Robin dengan metode SVM menyimpulkan bahwa penggunaan TF-IDF dalam pembentukan vektor kata mampu meningkatkan recall dan precision, serta proporsi data training antara fake news dan real news sehingga hasilnya valid (Rubin et al., 2016). Pada tahun 2020, Faisal at all, (Rahutomo et al., 2019) melakukan percobaan dengan Naïve Bayes yang menyimpulkan akurasinya adalah 68,33%.

Sistem deteksi berita hoax berbasis information retrieval telah dilakukan oleh Santoso (Santoso et al., 2018). Penelitian tersebut menyatakan bahwa ada beberapa langkah, antara lain input pengguna, pencocokan input dengan database, dan jika tidak ditemukan maka akan dicek di News API untuk menentukan apakah suatu berita masuk ke kategori hoax. Framework yang diusulkan (Rahutomo et al., 2019) tidak dijelaskan secara rinsi, terutama dalam memanfaatkan data klasifikasi yang telah divalidasi untuk digunakan sebagai data pelatihan tambahan.

Dalam penelitian ini akan melakukan studi mengenai akurasi dari penggunaan Naive bayes serta akan dikembangkan suatu sistem bebasis website untuk deteksi hoak dengan metode Naive bayes menggunakan bahasa pemrograman Code Igniter 4.

## II. STUDI LITERATUR

Dalam proses melakukan deteksi hoax, hal yang dilakukan adalah melakukan *text-processing* data berita, menghitung TF-IDF, melakukan pembuatan model dengan algoritma Naive bayes, serta melakukan perhitungan akurasi.

### Text Processing

Didalam proses *text-processing* terdapat beberapa tahapan metode untuk mendapatkan data yang sesuai dan akurat, beberapa Langkah yang dilakukan adalah sebagai berikut :

- *Case Folding* merupakan sebuah metode mengubah kata berhuruf kapital menjadi huruf kecil atau *lowercase* (Utami & Sari, 2018). Metode ini tak hanya mengubah menjadi *lowercase* tetapi juga menghilangkan tanda baca (Rahutomo et al., 2019). Dengan contoh “Perusahan itu Membeli Alat berat Bermerk C.A.T.” setelah melalui proses case folding maka menjadi “perusahaan itu membeli alat berat bermerk cat.”.
- *Data Cleaning* merupakan sebuah proses untuk menghilangkan noise dan data-data yang tidak relevan (Iqbal, 2021). Pada tahapan ini menghilangkan noice dan data-data yang tidak relevan seperti beberapa tanda baca (,:’?/;”).
- *Tokenizing* merupakan pemotongan kalimat berdasarkan tiap kata dan sekaligus mehilangkan beberapa karakter tertentu yang terdapat didalam kata tersebut (Utami & Sari, 2018). Dalam proses ini karakter yang dihilangkan seperti tanda baca, angka, dan karakter selain huruf alphabet, karena karakter tersebut dianggap sebagai pemisah kata (delimiter), tetapi terdapat dibeberapa kasus angka tidak dihilangkan karena masih dianggap penting (Rahutomo et al., 2019) .
- *Vectorize* merupakan tahapan untuk menghitung berapa frekuensi kata yang muncul dalam setiap datanya. Sehingga akan didapatkan ke akuratan data dari berapa banyak frekuensi kata tersebut muncul dan tergolong kata apa yang sering muncul tersebut.
- *Stopword* merupakan langkah dimana menghilangkan kata-kata yang sering kali digunakan tetapi tidak memiliki artian atau informasi yang penting (Rahutomo et al., 2019). Pada tahap ini menggunakan stopword list yang didalamnya berisi kata-kata yang tidak bermakna

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332

penting, seperti yang, di, ke, adalah, akhir, apabila, dari. Sehingga akan didapat data yang lebih presisi dan lebih detail.

**TF-IDF**

TF-IDF *(Term Frequency – Inverse Document Frequency)* adalah metode pembobotan yang paling populer dan memberikan hasil yang baik (Prasetijo et al., 2017). TF-IDF menghubungkan term frequency (TF) dan inverse document frequency (IDF) (Iqbal, 2021). TF-IDF merupakan salah satu metode ekstraksi fitur yang digunakan untuk menentukan vektor fitur yang berpengaruh pada keragaman kelas (Prasetijo et al., 2017). TF berarti menyoroti kata-kata yang lebih sering muncul dalam berita. Semakin banyak kata yang muncul, semakin besar nilainya. Sedangkan IDF adalah kebalikan dari frekuensi kemunculan kata diantara semua dokumen yang mengandung kata tersebut. Di IDF, semakin banyak kata yang muncul, semakin rendah nilainya. Rumus TF-IDF adalah sebagai berikut (Wahyuni et al., 2017).

TFi = N / (jumlah kata dalam D) (1 )
IDFi = log (jumlah $D/df$i) ( 2 )
$W$i,j = TFi,j x IDFi ( 3 )

**Daftar Simbol**
D: dokumen
N: Istilah frekuensi i dalam D
$df$i : jumlah dokumen yang mengandung term i
$W$i,j : bobot suku ke-i pada dokumen j

**Naïve Bayes**

Naïve Bayes menggunakan teorema bayes dalam membagi data ke dalam kelas-kelas berdasarkan probabilitasnya (Santoso et al., 2018). Naive Bayes bekerja untuk memprediksi probabilitas data berdasarkan data sebelumnya sehingga dapat digunakan untuk pengambilan keputusan (Prasetijo et al., 2017). Metode ini direkomendasikan karena beberapa alasan, antara lain karena dalam proses klasifikasi, metode ini hanya membutuhkan sedikit data pelatihan (Rifai et al., 2019) dan mudah dilacak (Dewi, 2016). Dalam menentukan probabilitas suatu kejadian ‘A’ (hoax atau nyata) ketika kejadian ‘B’ benar menggunakan Naïve Bayes dimana perhitungannya adalah sebagai berikut (Santoso et al., 2018)

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \qquad (1)$$

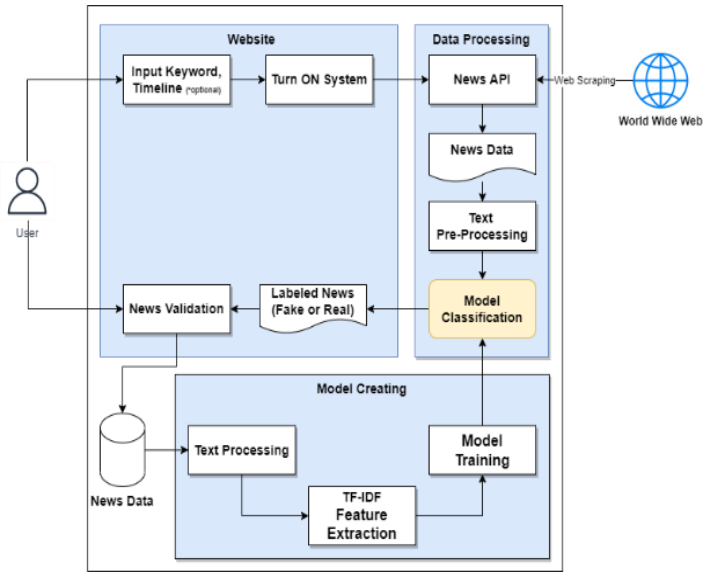
**Daftar Simbol:**
P(B | A) = Probabilitas 'B' benar dimana 'A' benar
P(A) = Probabilitas 'A' benar
P(B) = Probabilitas 'B' benar

## III. METODE

***Framework Pembuatan Sistem***

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332

Pada bagian ini akan dilakukan kerangka pembuatan sistem klasifikasi deteksi hoax yang digunakan sekaligus melihat akurasi dalam pembuatan model (Lihat Gambar 1).




Gambar 1. Framework sistem dekteksi hoax

Ada sistem ini, ada tiga bagian utama, yaitu website sebagai interface pada user, kedua adalah model crating, dan terakhir adalah data classification.

- Website User, bagian ini menangani input user berupa kata kunci. Misalkan ingin fokus dalam menyaring berita kesehatan mapun berita teknologi, maka user dapat input keyword. Kemudian, user dapat mengisi periode waktu, misalkan 1 hari atau 1 miggu berita terakhir. Kemudian, ada tombol on untuk mntrigger proses scraping dan filtering berita.
- Model Generating, proses ini dilakukan untuk melakukan pembuatan model classifikasi. Awal pembuatan model ini menggunakan data train yang sudah diset. Namun, data training ini akan terus diupdate sesuai dengan hasil filtering berita yang sudah divalidasi oleh user.
- News Classification, proses ini meurupakan proses untuk melakukan clasifikasi dari data baru hasil dari scraping. Hasil klasifikasi tersbeut divalidasi oleh user sebelum ditampilkan ke website. Berita yang sudah tervalidasi, selai tampil pada website juga masuk ke dalam database, dan sekaligus ditambahkan menjadi data training dan mengupdate model.

***Metodologi Pengembangan Waterfall***

Pada tahap pengembangan system menggunakan metode SDLC *(Systems Development Life Cycle)*. SDLC menjelaskan mengenai tahapan proses pengembangan sistem dengan menyajikan metodologi atau proses yang terstruktur untuk membangun suatu sistem (Sommerville, 2011). Dengan menggunakan SDLC pengembangan menjadi lebih efektif karena pada SDLC analisisnya disetiap detail tahapan dalam proses dengan rinci. Alasan menggunakan metode pengembangan SDLC waterfall dikarenakan system pendeteksi berita hoax ini membutuhkan proses yang urut yang dimulai dari proses perencanaan, analisa, desain, dan implementasi pada sistem.

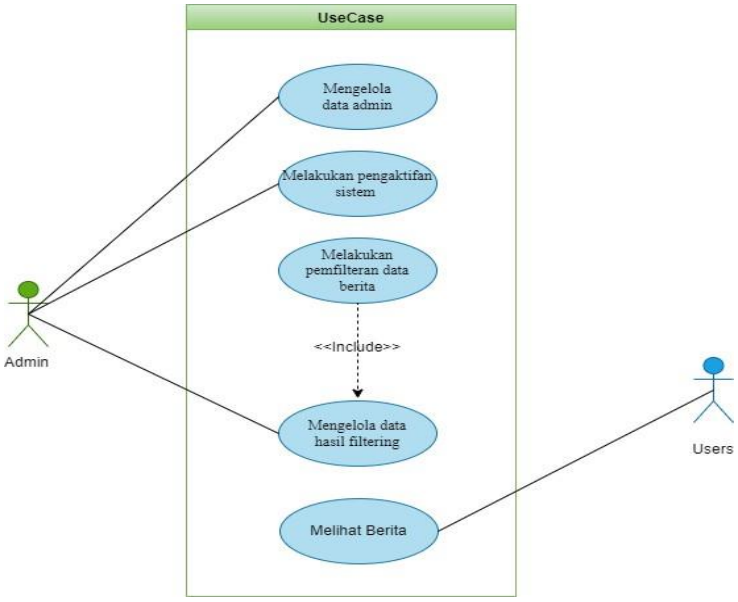
## IV. HASIL DAN PEMBAHASAN

### Hasil Pembuatan Sistem

Pada pengembangan sistem, tahap perancangan fitur dibuat dalam bentuk UML *use case* (Lihat Gambar 2). Pada *use case* pendeteksi fake news memiliki 2 aktor yaitu admin dan user

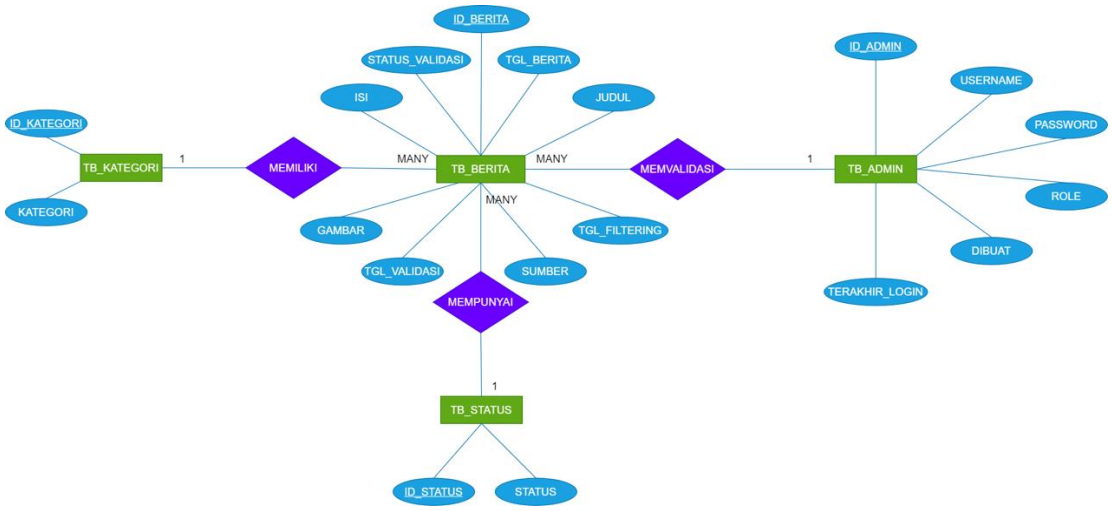
Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332

(pengunjung website). Dari *use case* ini dijelaskan bahwa sistem deteksi hoax ini memiliki beberapa fitur utama, yaitu admin mengelola data pengguna, admin melakukan aktifasi sitem (mulai crawling dan mulai untuk mengupdate model), admin melakukan validasi terhadap hasil dari pemodelan dan klasifikasi filtering hoax.




Gambar 2. Use case Sistem Deteksi Hoax.

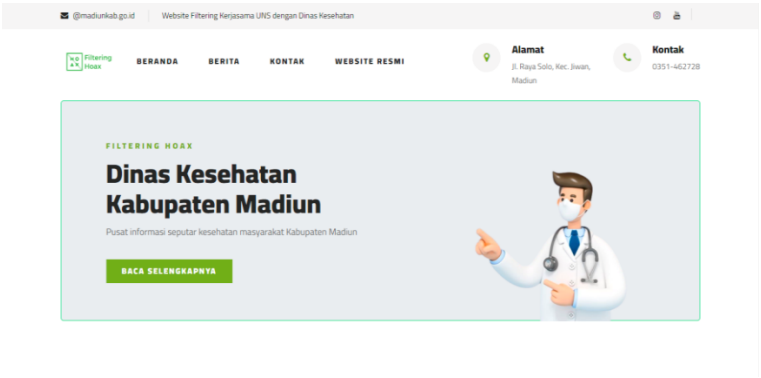
Pembuatan database berdasarkan ERD yang dapat diluhat pada Gambar 3. Tabel yang terlibat adalah admin, kategori, berita, dan hasil klasifikasi.




Gambar 3. ERD Sistem Deteksi Hoax

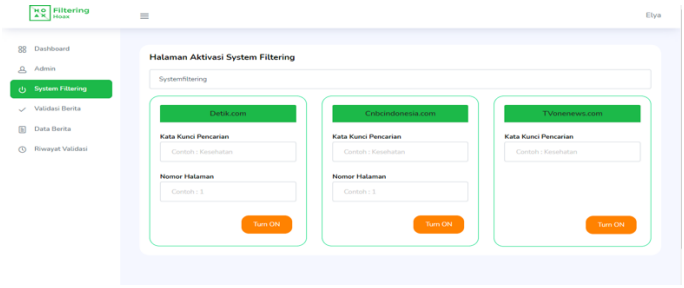
Hasil implementasi sistem deteksi hoax bidang kesehatan untuk Dinas Kabupaten Madiun dapat dilihat pada gambar 4.

Remik: Riset dan E-Jurnal Manajemen Informatika Komputer
Volume 7, Nomor 2, April 2023
http://doi.org/10.33395/remik.v7i2.12369
e-ISSN : 2541-1330
p-ISSN : 2541-1332



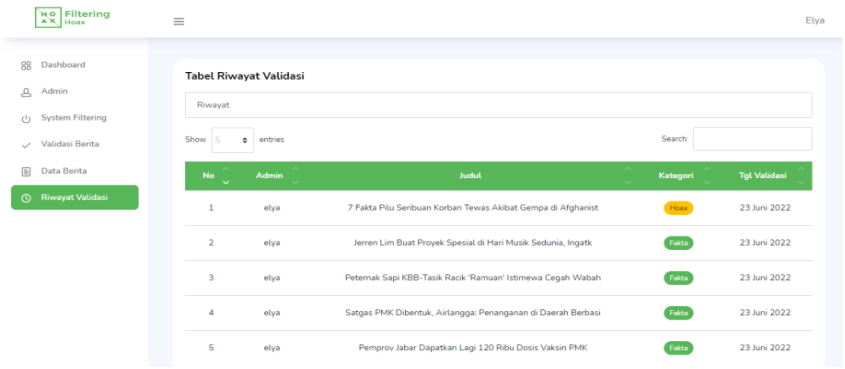

Gambar 4. Halaman Awal Hasil Implementasi

Pada menu pemilihan data untuk crawling, admin memilih sumber berita yang diinginkan. Disini terdapat 3 pilihan yaitu sumber berita dari TVonenews, Detik dan CNBCIndonesia (Lihat Gambar 5). Kemudian admin memilih kategori berita yang ingin difilter. Contoh admin memasukkan kata "Kesehatan" maka sistem akan mengambil semua jenis berita tentang Kesehatan. Admin memasukkan nomor halaman. Misalnya admin memasukan nomor halaman 2. Jika data sudah terisi lengkap pilih fitur "Turn On" untuk mengaktifkan Sisytem Filtering tersebut dan Ketika berhasil akan muncul alert "Berhasil mengaktifkan Filtering Berita




Gambar 5. Halaman Memilih Sumber *Crawling*

Pada bagian Navigation Berita ini ketika user klik menu tersebut maka akan langsung diarahkan ke berita yang telah terdapat hasil klasifikasi berita berupa Fakta dan Hoax (Lihat Gambar 6).




Gambar 6. Hasil Deteksi Berita Hoax Sistem

**Hasil Pengujian Akurasi**

Hasil pengujian memperlihatkan bahwa akurasi dari algoritma Naive Bayes dalam deteksi hoax menunjukkan 90.9%. Nilai ini diperoleh dengan mencocokkan, jumlah berita yang

benar diklasifikasikan sebagai data hoax dibagi dengan total jumlah data. Dimana data yang digunakan dalam pengujian ini berjumlah 200 data berita.

## V. KESIMPULAN

Dalam mendeteksi berita hoax, metode yang dapat digunakan adalah machine learning, khususnya untuk klasifikasi teks. Metode NaiveBayes berhasil diimplementasi dengan akurasi 90,9%. Sistem ini dibangun dengan menggunakan bahasa CI 4 dengan metode pengembangan Waterfall. Sistem ini telah digunakan oleh dinas Kabupaten Madiun dalam memberikan informasi hoax bagi masyarakat.

## VI. REFERENSI

Chen, Y. P., Chen, Y. Y., Yang, K. C., Lai, F., Huang, C. H., Chen, Y. N., & Tu, Y. C. (2022). The Prevalence and Impact of Fake News on COVID-19 Vaccination in Taiwan: Retrospective Study of Digital Media. *Journal of Medical Internet Research*, *24*(4). https://doi.org/10.2196/36830

Dewi, S. (2016). Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan. *Techno Nusa Mandiri*, *XIII*(1), 60–66.

Iqbal, M. A. (2021). Application of Regression Techniques with their Advantages and Disadvantages. *Elektron Magazine*, *September*.

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Alvin, Y., Soetrisno, A., Arfan, M., & Sofwan, A. (2017). *Hoax Detection System with SVM and SGD-2017.pdf*. 45–49.

Rahutomo, F., Pratiwi, I. Y. R., & Ramadhani, D. M. (2019). Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia. *Jurnal Penelitian Komunikasi Dan Opini Publik*, *23*(1). https://doi.org/10.33299/jpkop.23.1.1805

Rifai, M. F., Jatnika, H., & Valentino, B. (2019). Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS). *Petir*, *12*(2), 131–144. https://doi.org/10.33322/petir.v12i2.471

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). *Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News*. 7–17. https://doi.org/10.18653/v1/w16-0802

Santoso, I., Yohansen, I., Nealson, N., Warnars, H. L. H. S., & Hashimoto, K. (2018). Early investigation of proposed hoax detection for decreasing hoax in social media. *2017 IEEE International Conference on Cybernetics and Computational Intelligence, CyberneticsCOM 2017 - Proceedings*, *2017-Novem*, 175–179. https://doi.org/10.1109/CYBERNETICSCOM.2017.8311705

Sommerville, I. (2011). *Software engineering, 9th Edition*. Pearson.

Utami, P. D., & Sari, R. (2018). Filtering Hoax Menggunakan Naive Bayes Classifier. *Multinetics*, *4*(1), 57. https://doi.org/10.32722/vol4.no1.2018.pp57-61

Wahyuni, R. T., Prastiyanto, D., & Supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro Universitas Negeri Semarang*, *9*(1), 18–23. https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659