Jurnal Komputasi Terdistribusi

Vol.6 No.4, September 2023
ISSN : 27224707

KLASIFIKASI BERITA MENGENAI VAKSIN BOOSTER COVID-19 DI SOSIAL MEDIA TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER

Reinhard Hutagaol[1], Debi Yandra Niska[2]
[12] Universitas Negeri Medan
E-mail: reinhardhutagaol1@gmail.com[1], debiyandraniska@unimed.ac.id[2]

Abstrak

Perkembangan pesat teknologi internet telah mengakibatkan keterkaitannya yang erat dengan media sosial. Media massa saat ini sering menggunakan platform media sosial sebagai sarana untuk menyebarkan berita, termasuk berita palsu. Hoax adalah informasi yang sebenarnya tidak benar, namun disajikan seolah-olah benar. Hoax dibedakan dari rumor, ilmu semu, berita palsu, dan candaan. Tujuan dari penelitian ini adalah mengembangkan model machine learning dengan menggunakan metode Naive Bayes Classifier untuk mengklasifikasikan berita di platform Twitter. Model ini dibangun menggunakan dataset berita yang diperoleh dari laman resmi kominfo. Setiap berita telah dikategorikan sebagai hoax atau fakta. Totalnya, terdapat 2472 berita dalam dataset ini, dengan 1415 di antaranya merupakan berita fakta, dan 1057 berita lainnya merupakan berita hoax. Hasil evaluasi model menunjukkan akurasi sebesar 71%, presisi sebesar 51%, recall sebesar 51%, dan f1-score sebesar 51%. Setelah itu, model yang telah dikembangkan digunakan untuk mengklasifikasikan 97 berita Twitter yang belum diberi label mengenai vaksin booster. Hasilnya menunjukkan bahwa dari 97 berita tersebut, terdapat 62 berita yang mengandung fakta dan 35 berita yang merupakan hoaks.

Kata Kunci *— Hoax, Klasifikasi, Naive Bayes Classifier.*

PENDAHULUAN

Pada era masyarakat modern, perkembangan teknologi internet sangatlah pesat, dan media sosial telah menjadi bagian yang tidak terpisahkan. Banyak media massa yang menggunakan platform media sosial untuk menyebarkan berita. Setiap hari, berita-berita baru terus muncul dengan beragam topik. Di Indonesia, perkembangan teknologi informasi juga sangat cepat, dengan semakin bertambahnya pengguna internet. Namun, hal ini juga menyebabkan masalah persebaran berita yang sulit terkontrol, terutama berita-berita palsu

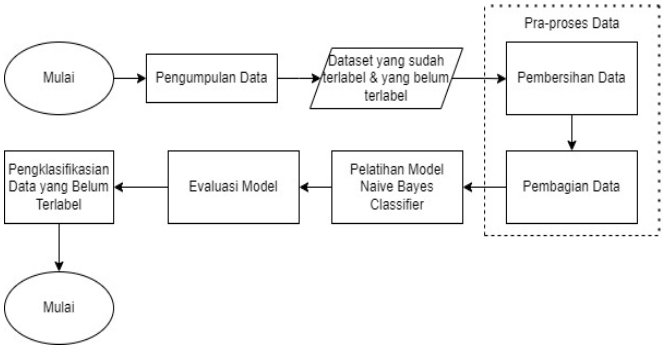
(hoax).

Hoax adalah informasi palsu yang disajikan sedemikian rupa sehingga terlihat benar. Berbeda dengan rumor, ilmu semu, atau berita palsu, hoax bertujuan untuk menciptakan ketidakamanan, ketidaknyamanan, dan kebingungan di kalangan masyarakat. Dalam kondisi bingung, orang-orang dapat mengambil keputusan yang lemah, tidak meyakinkan, bahkan salah. Hoax berpotensi memberikan dampak negatif melalui tulisan dengan mempengaruhi pikiran seseorang, sementara gambar dapat menimbulkan rasa takut dan ancaman. Menurut penelitian yang dilakukan oleh Lembaga Ilmu Pengetahuan Indonesia (LIPI), masyarakat yang bersikap fanatik lebih rentan terkena hoax. Jika dibiarkan, berita hoax dapat menjadi sangat persuasif, maka diperlukan strategi untuk mengidentifikasi dan menangkal berita-berita palsu yang menyebar di media sosial. Media sosial merupakan salah satu inovasi teknologi terkini yang berperan sebagai sumber dan sarana untuk bertukar serta menerima informasi. Beberapa platform media sosial yang sudah akrab dan banyak digunakan saat ini adalah Facebook, Instagram, dan Twitter. Dalam penelitian ini, akan dilakukan analisis berita tentang vaksin Covid-19 yang diposting di Twitter. Twitter sendiri adalah salah satu platform media sosial yang terkenal, didirikan oleh Jack Dorsey, yang digunakan untuk menyebarkan informasi berbasis mikroblogging dan mulai diluncurkan pada 13 Juli 2006.(Eka Sembodo et al., 2016)

Naïve Bayes Classifier, teknik pengklasifikasian dalam data mining, menjadi salah satu pilihan yang populer.. Sebelumnya, penelitian juga telah dilakukan dalam deteksi berita hoax di berbagai bidang. Contohnya, Hery Mustofa, dkk telah mengkaji klasifikasi berita hoax dengan menggunakan metode Naïve Bayes. Hasilnya menunjukkan tingkat akurasi sebesar 85.28%, terdiri dari 307 dokumen terklasifikasi relevan dan 53 dokumen yang tidak relevan, dengan error rate sebesar 14.72%.(Mustofa & Mahfud, 2019.

METODE PENELITIAN

Penelitian ini berfokus pada membangun model pengklasifikasian data teks berita yang sudah diberi label dan kemudian menguji model dengan menggunakan data teks berita yang belum terklasifikasikan di media sosial twitter terkait tentang vaksin booster. Yang pertama kali dilakukan adalah menyiapkan dataset untuk data latih dan data uji, dataset yang diambil ialah berita yang sudah memiliki label hoax dan fakta yang diperoleh dari laman resmi kominfo. Dan kemudian mengambil berita tentang vaksin booster di media sosial twitter yang belum memiliki label untuk diklasifikasikan dengan model yang telah dibuat.



Gambar 1 Skema Penelitian

- Pengumpulan Data, Data yang digunakan untuk membangun model adalah data yang sudah memiliki label dari laman resmi kominfo. Kemudian data yang sudah didapat disimpan dalam format csv. Dan untuk data yang belum memiliki label diambil secara manual dari Twitter terkait dengan vaksin booster untuk menguji model yang sudah dibangun.
- Pembersihab Data, proses awal dalam text mining yang terdiri dari beberapa tahapan yaitu: Cleansing, Casefolding, Tokenization, Stopword Removal, Stemming, dan Detokenization.
- Pembagian Data, digunakan untuk membagi data menjadi data latih dan data uji dengan rasio data yang digunakan.
- Pelatihan Model Naïve Bayes Classifier, Hasil data yang sudah mengalami proses preprocessing akan dilatih dengan menggunakan algoritma Multinomial Naïve Bayes.
- Evaluasi Model, setelah pengujian model dilakukan, selanjutnya akan dilakukan penghitungan akurasi dengan menggunakan confussion matrix.
- Pengklasifikasian Data yang Belum Terlabel, proses ini melakukan prediksi terhadap data yang telah dikumpulkan dari twitter yang belum memiliki label, kemudian hasil nya ditampilan dalam grafik.

HASIL DAN PEMBAHASAN

1. Pengumpulan data

Pengumpulan data untuk model adalah, dataset yang sudah diberi pelabelan diambil secara manual dari laman kominfo dan data tersebut disimpan ke dalam file dengan format csv. Sedangkan untuk pengambilan data untuk menguji model (data twitter) diambil secara manual dengan cara mencari satu per satu tweet/postingan dengan *#(hashtag) #covid #covid-19 #vaksin #vaksinbooster* lalu disimpan ke dalam file dengan format csv juga

No.	***Tweet***
1.	Berakhir di 2022, Vaksin booster akan berbayar
2.	Kementrian Kesehatan merilis kebijakanvaksin Covid-19 booster fase kedua atau vaksin dosis keempat
3.	Dinas Kesehatan (Dinkes) DKI Jakarta menginformasikan sejumlah 201.927 orang sudah disuntik vaksin covid-19 booster kedua. Data itu Terekam sampai Minggu, 29 Januari 2023

2. Cleansing

Proses ini bertujuan untuk menghilangkan komponen yang tidak berpengaruh dalam klasifikasi, seperi karakter, simbol, angka.

No.	Berita	*Cleansing*	Keterangan
1	Anjuran Berbelanja Ketika Menjalankan "S*ocial Distancing*"	Anjuran Berbelanja Ketika Menjalankan *Social Distancing*	Menghilangkan isyarat (") dan Hashtag (#)
2	Manusia Akan Menjadi Mayat Hidup "*Zombie"* Ketika Divaksin	Manusia Akan menjadi Mayat Hidup *Zombie* Ketika Divaksin	
3	#ayovaksin Agar Keluarga Menjadi Aman	Ayovaksin Agar Keluarga Menjadi Aman	

3. Case Folding

Proses ini dilakukan untuk mengubah semua ukuran huruf kapilat (huruf besar) di dalam kalimat menjadi huruf kecil.

a. Tampilan Halamn loading

Tampilan halaman loading muncul setelah tombol start di halaman menu utama ditekan pada aplikasi, memunculkan animasi loading yang berjalan

No.	*Cleansing*	*Case Folding*	Keterangan
1	Anjuran Berbelanja Ketika Menjalankan *Social Distancing*	anjuran berbelanja ketika menjalankan *social distancing*	Mengubah semua huruf capital menjadi huruf kecil
2	Manusia Akan menjadi Mayat Hidup *Zombie* Ketika Divaksin	manusia akan menjadi mayat hidup *zombie* ketika divaksin	
3	Ayo Vaksin Agar Keluarga Menjadi aman	ayo vaksin agar keluarga menjadi aman	

4. Tokenization

Proses ini akan digunakan untuk memecah kalimat teks menjadi beberapa kata.

No.	*Case Folding*	*Tokenization*	Keterangan
1	anjuran berbelanja ketika menjalankan *social distancing*"	anjuran, berbelanja, ketika, menjalankan, *social, distancing*	Mengubah teks menjadi sebuah kata-kata.
2	manusia akan menjadi mayat hidup *zombie* ketika divaksin	manusia, akan, menjadi, mayat, hidup, *zombie*, ketika, divaksin	
3	ayo vaksin agar keluarga Menjadi aman	ayo, vaksin, agar, keluarga, menjadi, aman	

5. Stopword Removal

Proses ini dilakukan untuk menghapus kata yang sering muncul namun tidak memiliki makna sama sekali, contoh kata: yang, di, atau, karena, ke, dengan, tetapi, dan lainnya. Proses ini dilakukan secara otomatis dengan bantuan library python dan menyertakan corpus (kamus) tambahan dalam bentuk file txt yang terdiri dari kata kata yang tidak dibutuhkan.

No.	*Tokenization*	*Stopword Removal*	Keterangan
1	anjuran, berbelanja, ketika, menjalankan, *social*, *distancing*	anjuran, berbelanja, menjalankan, *social*, *distancing*	Menghilangkan kata yang tidak memiliki nilai/makna.
2	manusia, akan, menjadi, mayat, hidup, *zombie*, ketika, divaksin	manusia, akan, menjadi, mayat, hidup, *zombie*, divaksin	
3	ayo, vaksin, agar, keluarga, menjadi, aman	ayo, vaksin, keluarga, menjadi, aman	

6. Stemming

Proses ini dilakukan agar dapat mengubah kata berimbuhan menjadi kata dasar.

No.	*Stopword removal*	*Stemming*	Keterangan
1	usul, berbelanja, menggerakkan, *social*, *distancing*	usul, belanja, gerak, *social*, *distancing*	Mengganti kata berimbuhan membentuk kata awal
2	manusia, bakal, menjadi, mayat, hidup, *zombie*, divaksin	manusia, bakal, jadi, mayat, hidup, *zombie*, vaksin	
3	ayo, vaksin, supaya, keluarga, menjadi, aman	ayo, vaksin, supaya, keluarga, jadi, aman	

7. Detokenizing

Proses ini adalah kebalikan dari tokenization, yaitu menghilangkan semua tanda pemisah dan tanda baca sehingga tercipta sebuah kalimat baru. Data yang di detokenization adalah data yang telah mengalami proses stemming.

No.	*Stemming*	*Detokenizing*	Keterangan
1	anjur, belanja, ketika, jalan, *social, distancing*	anjur belanja jalan *social distancing*	Menghilangkan tanda pemisah kata, sehingga menciptakan kalimat baru
2	manusia, akan, jadi, mayat, hidup, *zombie*, vaksin	manusia akan jadi mayat hidup zombie vaksin	
3	ayo, vaksin, agar, keluarga, jadi, aman	ayo vaksin agar keluarga jadi aman	

8. Pembagian Rasio Data

Sebelum proses klasifikasi, dataset yang ada dilakukan split data terlebih dahulu. Split data adalah langkah pembagian data menjadi 2 bagian, data training dan data testing. Dilakukan pelabelan data sebanyak 4 rasio yaitu: 90: 10, 80: 20, 70: 30, 60: 40. Dilakukan nya pembagian data tersebut bertujuan untuk menguji pembagian data mana yang menunjukkan akurasi paling tinggi.

9. Pelatihan Model Naïve Bayes Classifier

Multinomial Naïve Bayes adalah supervised learning yang memprediksi kelas suatu anggota dengan probalitas yang difokuskan untuk klasifikasi teks. Dimulai dengan menghitung nilai prior probality (nilai kata) pada setiap label atau topic menggunakan Rumus persamaan:

P(c)= $\frac{N_c}{N}$

Keterangan:

P(c) = prior probability kelas c

N_c = jumlah kelas c pada seluruh dokumen

N = jumlah seluruh dokumen

Nilai dari prior probability tersebut digunakan pada tahap pengujian. Tahap pengujian dilakukan perhitungan probabilitas kata di suatu label dengan mempertimbangkan bobot (TF-IDF) dari kata tersebut. Rumus Multinomial Naïve Bayes yang digunakan dengan pembobotan TF-IDF seperti pada rumus berikut:

$$P(t_n|c) = \frac{W_{ct} + 1}{\left(\sum W'^{\in VW'_{ct}}\right) + B'}$$

Keterangan:

W_ct = nilai pembobotan TF-IDF pada term t di kelas c

$\sum W'^{\in VW'_{ct}}$ = jumlah total bobot dari keseluruhan term yang berada di kelas c

B' = jumlah bobot dari term yang unik pada seluruh dokumen.

Selanjutnya dilakukan perhitungan probabilitas suatu tweet masuk ke dalam suatu label dengan menggunakan Persamaan:

$$P(c|term\ dokumen\ d) = P(c) * P(t_1|c) * P(t_2|c) * \dots * P(t_n|c)$$

Keterangan:

P(c | term dokumen d) = probabilitas suatu dokumen d berada di kelas c

P(c) = probabilitas prior dari kelas c

P(t_n | c) = probabilitas kata ke-n pada kelas c

t_n = term ke-n dokumen d

Rasio	Accuracy
90 : 10	66%
80 : 20	71%
70 : 30	68%
60 : 40	66%

Dari tabel diatas dapat disimpulkan bahwa pembagian data rasio yang memiliki akurasi tertinggi yaitu 80 : 20 yang memiliki akurasi 71%. Jadi, pembagian rasio data yang digunakan untuk membangun model ialah 80 : 20 (1977 data latih dan 495 data uji).

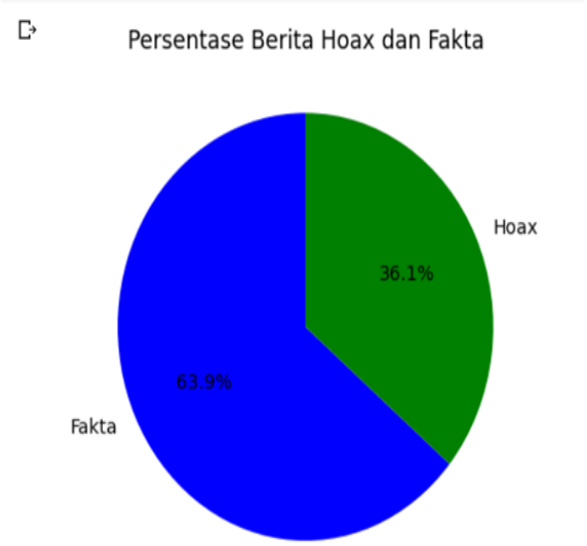
10. Evaluasi Model

Hasil evaluasi didapatkan dengan menggunakan Confussion Matrix, maka di dapatkan hasil Accuracy, Recal, Precision, F1-Score.

		Actual Value	
		fakta	hoax
Predict Value	fakta	1139	59
	hoax	225	270

11. Pengklasifikasian Data yang Belum Terlabel

Setelah mendapatkan nilai dari model klasifikasi yang menggunakan Naïve Bayes yang memiliki nilai Accuracy sebesar 71%, Precision sebesar 51%, Recall sebesar 51% dan f1-score sebesar 51%. Kemudian melakukan pengklasifikasian data yang belum memiliki label, data yang belum memiliki label adalah berita twitter yang belum memiliki label sebanyak 97 berita.



Gambar diatas adalah presentase hasil berita twitter yang sudah terlabeli sebanyak 97 berita yang terdiri dari, 63.9% berita fakta (62 berita) dan 36.1% berita hoax (35 berita).

KESIMPULAN

Model yang dibangun menggunakan dataset (berita) yang sudah memiliki label yang didapatkan di laman resmi kominfo dengan total 1415 berita berlabel fakta, dan 1057 berita berlalbel hoax. Kemudian dataset yang sudah diberi label dilakukan preprocessing. Preprocessingg sendiri terdiri dari Cleansing, Case Folding, Tokenization, Stopword Removal, Stemming, dan Detokenizing. Kemudian, melakukan klasifikasi Naïve Bayes dengan membagi data training dan data testing. Setelah itu dilakukan proses perhitungan evaluasi menggunakan confusion matrix untuk mendapatkan nilai precision, recall, f1-score, dan accuracy.

Hasil evaluasi model adalah sebagai berikut: nilai accuracy sebesar 71%, nilai precision sebesar 51%, nilai recall sebesar 51%, dan nilai f1-score sebesar 51%. Kemudian model yang sudah dibangun melakukan pengklasifikasian berita twitter yang belum memiliki label. Berita yang digunakan sebanyak 97 berita dan hasil klasifikasi terdapat 62 jumlah berita terklasifikasi fakta dan 35 berita terklasifikasi berita hoax.

DAFTAR PUSTAKA

Romzi, M., & Kurniawan, B. (2020). Implementasi Pemrograman Python Menggunakan Visual Studio Code. In JIK: Vol. XI (Issue 2). www.python.org

Rozaqi, A., Triayudi, A., & Aldisa, R. T. (2022). Analisis Sentimen Vaksinasi Booster Berdasarkan Twitter Menggunakan Algoritma Naïve Bayes dan K-NN. Jurnal Sistem Komputer Dan Informatika (JSON), 4(1), 184. https://doi.org/10.30865/json.v4i1.4907

Ananda, Ayu (2021). Klasifikasi Berita Hoax Covid-19 Menggunakan Metode Naïve Bayes Berbasis PSO. Universitas Muhammadiyah: Malang

Abdan, S. (2017). Analisis Sentimen Masyarakat terhadapat E-Commerce pada Media Sosial Menggunakan Metode Naive Bayes Classifier (NBC) dengan Seleksi Fitur Information Gain (IG).

Eka Sembodo, J., Budi Setiawan, E., & Abdurahman Baizal, Z. (2016). Data Crawling Otomatis pada Twitter. October 2018, 11–16. https://doi.org/10.21108/indosc.2016.111

Mustofa, H., & Mahfudh, A. A. (2019). Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes. Walisongo Journal of Information Technology, 1(1), 1. https://doi.org/10.21580/wjit.2019.1.1.3915

Setiadi. (2016). Pemanfaatan Media Sosial Untuk Efektifitas Komunikasi. Jurnal Humaniora Universitas Bina Sarana Informatika. 16(2)

Sofyan, A., & Santosa, S. (2016). Lapor Menggunakan Metode C4.5 Berbasis Forward Selection. In Jurnal Teknologi Informasi (Vol. 12, Issue 1). http://research.pps.dinus.ac.id

Syahnur, M. H., Bijaksana, M. A. & Mubarok, M. S.. (2016). Kategorisasi Topik Tweet di Kota Jakarta, Bandung, dan Makassar dengan Metode Multinomial Naïve Bayes Classifier. eProceeding of Engineering, vol. 3, p. 3631.

Twitter @ id.wikipedia.org. (n.d.).