REPOSITOR, Vol. 3, No. 4, Agustus 2021, Pp. 419-424
ISSN : 2714-7975
E-ISSN : 2716-1382

Perbandingan Klasifikasi Berita Hoax Kategori Kesehatan Menggunakan Naïve Bayes dan Multinomial Naïve Bayes

Chita Nauly Harahap*[1], Gita Indah Marthasari[2], Nur Hayatin[3]
[1,2,3]Teknik Informatika/Universitas Muhammdiyah Malang
chitaharahap8@gmail.com*[1], gita@umm.ac.id[2], noorhayatin@umm.ac.id[3]

Abstrak

Berita kesehatan merupakan informasi yang paling dicari dan diminati pada masa pandemi, kebutuhan akan perlunya kiat-kiat kesehatan untuk masyarakat membuat berita kesehatan menduduki peringkat atas berita terpopuler. Disaat meningkatnya minat baca masyarakat terhadap berita, banyak pihak tidak bertanggung jawab memanfaatkan keuntungan tersebut dengan menyebarkan berita tidak benar yang menggiring opini masyarakat agar menyudutkan pihak tertentu dan berisi informasi yang melenceng dari pendapat ahli kesehatan. Oleh karena itu salah satu cara untuk mengatasi tersebarnya berita hoax penelitian ini melakukan klasifikasi berita kesehatan berbahasa Indonesia secara otomatis. Pada penelitian ini dataset yang digunakan sebanyak 100 berita kesehatan non-hoax dan 100 berita kesehatan hoax. Proses klasifikasi memalui tahap preproses, pembobotan kata, dan implementasi pada metode naïve bayes dan multimonial naïve bayes. Evaluasi model menggunakan metode 10-fold cross validation, metode multinomial naïve bayes bekerja lebih baik dan efisien daripada metode naïve bayes.

Kata Kunci: *Naïve Bayes, Multinomial Naïve Bayes, Klasifikasi teks, Berita Kesehatan*

Abstract

Health news is the most sought-after information and is in demand during a pandemic, the need for health tips for the public makes health news at the top of most popular news. When the public's interest in reading news increases, many parties are not responsible for taking advantage of this chance by created and spreading hoax news. It may leads public opinion to corner certain parties. Hoax news also contains information that deviates from the opinion of health experts and it can be danger for public. Therefore, one way to overcome the spread of hoax news in this research is classification health news in Indonesian automatically. In this study, the dataset used was 100 non-hoax health news and 100 hoax health news. The classification process goes through preprocessing, word weighting, and implementation of the Naïve Bayes method and the Multinomial Naïve Bayes method. The evaluation of the model uses the 10-fold cross-validation method, the Multinomial Naive Bayes method is more better and efficient than the Naïve Bayes method.

Keywords: *Naïve Bayes, Multinomial Naïve Bayes, Text classification, Health news*

1. Pendahuluan

Pada zaman sekarang teknologi berkembang sangat pesat, selaras dengan kemajuan teknologi. Informasi pun tak kalah cepat beradar dalam kehidupan manusia. Jika pada zaman dahulu informasi kadang kala didapatkan melalui mulut ke mulut karena mungkin keterbatasan teknologi yang hadir pada masa itu. Tetapi untuk era sekarang informasi menjadi salah satu komponen penting hampir untuk semua manusia [1]. Setiap detiknya ada informasi baru yang akan muncul berkaitan tentang segala hal yang terjadi pada dunia, hal ini memiliki dampak berlimpahnya data informasi. Kadang kala tidak semua data informasi yang beredar itu benar, tidak sedikit informasi *hoax* atau bohong yang beredar disetiap harinya [2].

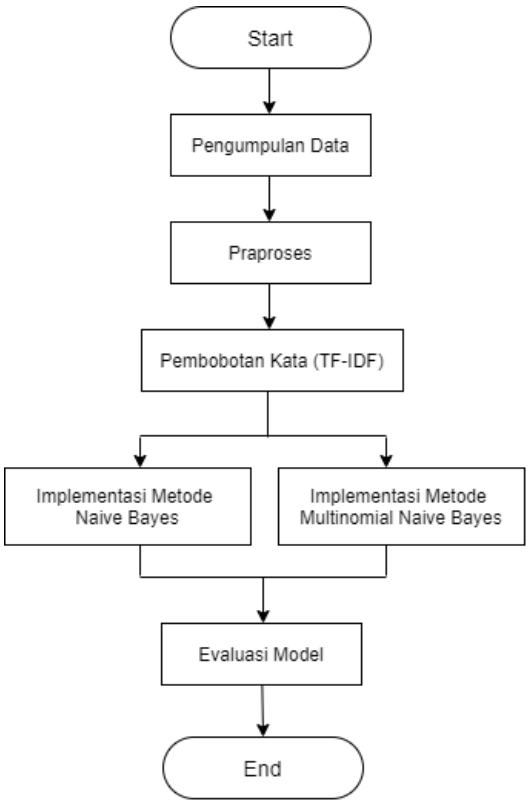
Salah satu cara mengurangi tersebarnya informasi palsu diperlukannya metode klasifikasi berita. Baik itu dilakukan secara manual ataupun dilakukan secara otomatis oleh sistem [3]. Sebelumnya telah banyak penelitian terkait dengan klasifikasi berita *hoax* menggunakan metode-metode tertentu seperti yang dilakukan oleh Hery Mustofa dan Adzhal Arwani Mahfudh penelitian tersebut menggunakan metode Naïve Bayes dengan hasil akurasi 85.28% [3]. Penelitian lain terkait klasifikasi berita *hoax* juga dilakukan oleh Andre Rino Prasetyo dan Indriati menggunakan

metode modifikasi K-Nearest Neighbor dengan hasil akurasi sebesar 75 % [4]. Klasifikasi berita Bahasa Indonesia juga dilakukan menggunakan metode multinomial naïve bayes oleh Amelia, Wiranto, dan Afrizal terbukti mendapatkan hasil akurasi yang lebih baik 86.62 % [5]. Metode Naïve Bayes memiliki hasil akurasi yang tinggi dan terbukti efektif dalam melakukan proses perhitungan [6] dan metode multinomial naïve bayes terkenal dengan keserdahaan dan membuat probabilitas dari gabungan kata-kata dan kategori [7]. Berdasarkan penelitian yang telah dilakukan terdahulu, penelitian ini memutuskan untuk membandingkan metode *naïve bayes* dan *multinomial naïve bayes*.

Kumpulan klasifikasi berita *hoax* tentang kesehatan pada *website turnbackhoax.id* masih dilakukan secara manual. Diperlukannya proses klasifikasi berita *hoax* tentang kesehatan secara sistematis dengan menggunakan metode tertentu. Cara mendapatkan metode terbaik tentunya dengan membandingkan antara hasil akurasi. Sehingga pada penelitian ini mengusulkan menggunakan Metode Naïve Bayes dan Multinomial Naïve Bayes sebagai metode pembanding. Dengan demikian penelitian ini dapat menemukan metode mana yang lebih baik digunakan untuk klasifikasi berita *hoax* kesehatan berbahasa Indonesia.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan teks mining untuk menganalisa informasi dan pengambilan langkah keputusan yang berhubungan dengan data berbentuk teks bertujuan agar mendapat informasi yang lebih berkualitas [8]. Flowchart metode penelitian seperti pada Gambar 1 meliputi studi pustaka, pengumpulan data, praproses, pembobotan kata TF - IDF (*Term Frequency – Invers Document Frequency*), implementasi metode, evaluasi model, kesimpulan dan saran.



Gambar 1. Metode Penelitian

2.1 Pengumpulan Data

Dataset berita diperoleh dengan teknik *web crawler*. *Web crawler* memiliki prinsip kerja dengan cara mengambil dan mengunduh informasi pada sebuah halaman serta memeriksa halaman yang relevan dengan yang pengguna inginkan [9]. Dataset 100 berita hoax dieproleh dari *website turnbackhoax.id*, pengambilan data dalam rentan waktu Agustus 2019 – November

2020. Dataset 100 berita *non hoax* diperoleh dari 4 *website* terkemuka, yaitu *CNN*, *CNBC*, *Health.detik*, dan *health.kompas*.

2.2 Praproses

Praproses merupakan tahapan proses data untuk mendapatkan data latih dan data uji lebih terstruktur yang akan menjadikan nilai akurasi lebih baik dan lebih persisi [10]. Berikut merupakan tahapan praproses.

1. *Case folding* merupakan pemrosesan teks yaitu merubah semua huruf kapital (*uppercase*) pada dokumen berita menjadi huruf kecil (*lowercase*) dan menghilangkan delimiter [11].
2. *Tokenizing* merupakan pemisah kata, membuat kata pada kalimat dan paragraf menjadi potongan kata tunggal atau frasa [3].
3. *Stop forward removal* merupakan penghapusan kata yang tidak memiliki makna dan kata tersebut sering muncul dengan menerapkan library sastrawi [1].
4. *Stemming* merupakan proses mengubah kata kembali menjadi bentuk kata dasarnya dengan menerapkan library sastrawi yang sesuai menurut kaidah tata Bahasa Indonesia [3].

2.3 Pembobotan Kata

Pembobotan kata menggunakan dua seleksi fitur yaitu *Term Frequency* (TF) dan *Invers Document Frequency* (IDF). fitur TF melakukan perhitungan jumlah atau banyaknya kemunculan setiap kata pada dokumen [6]. Bisa disebut juga *frequency* term a muncul pada dokumen teks b lalu dibagi dengan total term pada dokumen b. Persamaan 1 fitur TF seperti pada dibawah.

$$tf_{ab} = \frac{f_d\,(a)}{\max f_d\,(b)} \tag{1}$$

Fitur IDF merupakan pemeberian nilai pada kata dengan melihat kemunculan kata tersebut pada keseluruhan berita didataset [1]. Persamaan IDF dijabarkan pada Persamaan 2 dibawah.

$$idf_a = log_{10}\frac{N}{d\,f_a} \tag{2}$$

2.4 Metode Naïve Bayes

Metode ini merupakan metode probabilitas sederhana dengan didasari oleh teorema bayes yang berupa adanya peluang terjadinya suatu peristiwa yang serupa. Perhitungan tersebut dilakukan dengan cara probabilitas intrinsik (didapatkan dari data sekarang) dikalikan probabilitas bahwa hal yang sama dapat terjadi lagi dimasa depan [3]. Berikut Persamaan 3 formula metode naïve bayes untuk klasifikasi.

$$P(a|b) = \frac{T_{a|b}\ +\ 1}{N_b\ +\ V} \tag{3}$$

2.5 Metode Multinomial Naïve Bayes

Metode ini merupakan pengembangan dari metode naïve bayes yang seringkali digunakan untuk pengklasifikasian data teks [12]. Multinomial naïve bayes merupakan sebuah metode supervised learning [7]. Cara kerja multinomial naïve bayes, kategori dokumen tidak hanya ditentukan oleh kata yang muncul tetapi juga melihat frekuensi kemunculan dari kata tersebut [13]. Berikut Persamaan 4 terkait formula dari metode multinomial naïve bayes.

$$P(w\mid c) = \frac{count(w,c)\ +\ 1}{count(c)\ +\ |V|\ +\ 1} \tag{4}$$

2.6 Evaluasi Model

Tahapan ini dilakukan pengujian menggunakan *stratified k-fold cross validation* yaitu teknik pembagian data yang paling umum digunakan untuk menemukan model terbaik [14]. Cara kerja metode ini dengan membagi data dokumen teks berita hoax menjadi dua bagian, yaitu data uji dan data latih yang jumlah harus sama banyak antar *fold* [15].

Evaluasi model selanjuntya dengan tabel *confusion matrix.* Tabel tersebut dapat mengambarkan suatu kinerja model. *FN (false negative)* pada tabel dianggap sebagai nilai penting, karena semakin rendah nilai FN semakin baik model tersebut [12]. Tabel tersebut juga bertujuan untuk mengetahui seberapa besar dan bagus kinerja dari model yang telah dibangun pada klasifikasi berita *hoax* kesehatan dengan melihat nilai *precision*, *recall,* dan *f-measure* [3].

3. Hasil Penelitian dan Pembahasan

Dataset berita dikumpulkan dengan teknik crwling data menggunakan software data miner pada website turnbackhoax.id sebanyak 100 berita kesehatan hoax dan mengumpulkan 100 berita kesehatan non hoax pada 4 website terpecaya. Setelah data berhasil dikumpulkan, data tersebut masuk tahapan preproses untuk menghilangkan noise didalam berita yang meliputi, karakter bukan huruf, menghilangkan kata tidak memiliki arti, dan menghilangkan imbuhan pada kata. Melakukan ketiga tahap itu menjadikan dataset terstruktur dan bisa diproses pada perhitungan TF-IDF. Pada perhitungan TF IDF didapatkan 10 kata dengan bobot kata paling tinggi seperti pada Tabel 1 dibawah ini.

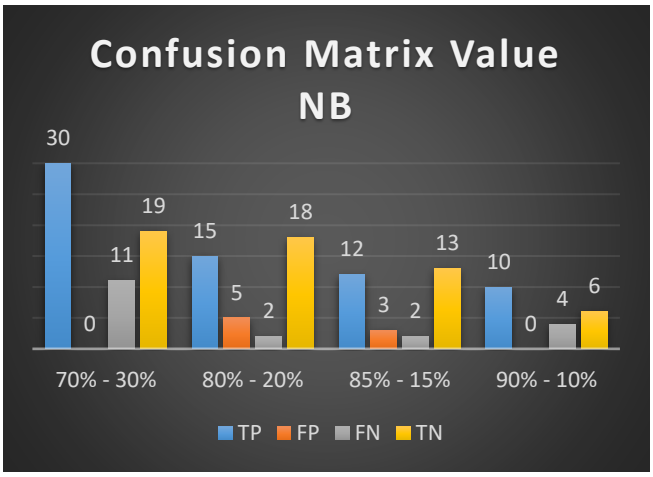
Tabel 1. 10 Kata TF-IDF

Rank	Kata	Nilai TF-IDF	Rank	Kata	Nilai TF-IDF
1	covid	6.32581	6	sehat	4.63474
2	orang	5.35682	7	sakit	4.53238
3	virus	5.35598	8	lelah	4.11698
4	jadi	5.28368	9	banyak	4.04020
5	sebut	4.81115	10	hari	3.75603

Menghidari leaked data antara data latih dan data uji, split data dilakukan pada tahap awal dan membuat empat ratio split data yang berbeda yaitu 70% - 30%, 80% - 20%, 85% - 15%, 90% - 10%. Mengatasi masalah overfiting penelitian ini menambahkan *MinMax scalling* yang dapat mengatur *range* X data menjadi [X_{min} - X_{max}] karena dataset telah berbentuk *numerical*, lalu melakukan *transform* 'yeo-johnson' agar data dipetakan mendekatai distribusi *gaussian*, dan terakhir membuat *tunning parameter*. Kemudian pipeline tersebut diujikan pada metode.

3.1 Evaluasi Naïve Bayes

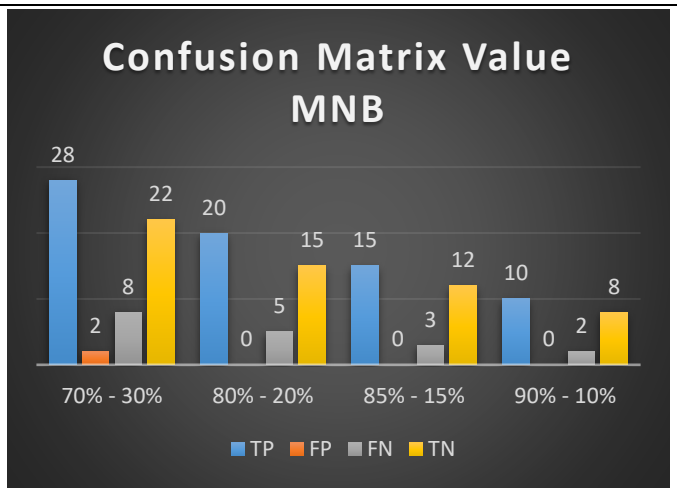
Pengujian menggunakan metode naïve bayes menggunakan *stratified k-fold cross validation* ditemukan hasil terbaik pada **fold 1** dengan ratio split data 80% data uji dan 20% data latih. Visualisai hasil ditampilkan dengan *confusion matrix* seperti Gambar 2 dibawah ini.



Gambar 2. Grafik Confusion Matrix Naïve Bayes

3.2 Evaluasi Multinomial Naïve Bayes

Pengujian menggunakan metode multinomial naïve bayes menggunakan *stratified k-fold cross validation* ditemukan hasil terbaik pada **fold 2** dengan ratio split data 80% data uji dan 20% data latih. Visualisai hasil ditampilkan dengan *confusion matrix* seperti Gambar 3 dibawah ini.



Gambar 3. Grafik Confusion Matrix Multinomial Naïve Bayes

3.3 Analisis Hasil

Tahapan analisis hasil pada penelitian ini menjelaskan setelah melakukan empat model split data yaitu, 70% - 30 %, 80% - 20%, 85% - 25%, dan 90% - 10% didapatkan hasil akurasi seperti Tabel 2 dibawah ini

Tabel 2. Perbandingan Hasil 2 Metode

	Naïve Bayes				Multinomial Naïve Bayes			
	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
I	81.7%	100%	73,10%	84,50%	83.3%	93%	77%	84,80%
II	82.5%	75%	88,20%	81%	87.5%	100%	80%	88,80%
III	83.3%	80%	85,70%	82,70%	90%	100%	83,30%	90,90%
IV	80%	100%	71,40%	83,30%	90%	100%	83,30%	90,90%

Seperti yang terlihat pada tabel 4.6, kinerja metode naïve bayes paling baik pada model split data **III** 85% - 15% yaitu 170 berita (data training) dan 30 berita (data test) dengan hasil pengujian pada data test **25 berita** berhasil terklasifikasi dengan benar dan **5 berita** terklasifikasi salah, diperoleh akurasi 83.3%. Kinerja multinomial naïve bayes sebagai metode pembanding paling baik pada dua model split data **III** dan **IV**. Pada model split data **III** 85% - 15% yaitu 170 berita (data training) dan 30 berita (data test) dengan hasil pengujian pada data test **27 berita** berhasil terklasifikasi dengan benar dan **3 berita** terklasifikasi salah, diperoleh akurasi 90%. Pada model split data **IV** 90% - 10% yaitu 180 berita (data training) dan 20 berita (data test) dengan hasil pengujian pada data test **18 berita** berhasil terklasifikasi dengan benar dan **2 berita** terklasifikasi salah, diperoleh akurasi 90%.

4. Kesimpulan dan Saran

Berdasarkan hasil tersebut, evaluasi kedua metode tersebut. Dapat disimpulkan bahwa metode multinomial naive bayes memiliki kinerja yang lebih baik dan efisien dalam klasifikasi dataset news health dibandingkan metode naive bayes. Membagi data antara data pelatihan dan data pengujian, mempengaruhi performa kedua metode. Saran untuk penelitian selanjutnya adalah memperbanyak dataset untuk proses klasifikasi dan menggunakan metode lain untuk klasifikasi yang lebih variatif dan mencoba menggabungkannya dengan fitur seleksi untuk mendapatkan hasil yang lebih baik.

Daftar Notasi

$f_d\ (a)$ = Banyaknya kata a pada dokumen teks b
$\max f_d\ (b)$ = Total kata pada dokumen b
N = Banyaknya dokumen
$d\ f_a$ = Banyaknya kata a pada keseluruhan dokumen
$T_{a|b}$ = Hasil perhitungan *tf-idf* pada kata tersebut

N_b = Banyaknya kata pada keseluruhan dokumen kategori b
V = Banyaknya kata pada keseluruhan dataset dokumen
$P(w|c)$ = Probabilitas dokumen w yang berada di kelas c
$count(w, c)$ = Frekuensi term c pada dokumen w
$count(c)$ = Frekuensi keseluruhan term pada dokumen c
$V\ (vocab)$ = Banyaknya keseluruhan kata yang ada

Referensi

[1] D. N. Chandra, G. Indrawan, and I. N. Sukajaya, "Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram," vol. 10, no. 1, pp. 11–19, 2016.
[2] T. Trisna *et al.*, "Analysis and Detection Of Hoax Contents In Indonesian News Based On Machine Learning," vol. 4, no. 1, 2019.
[3] H. Mustofa, A. A. Mahfudh, I. Negeri, and W. Semarang, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," vol. 1, no. 1, pp. 1–12, 2019.
[4] A. R. Prasetyo and P. P. Adikara, "Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor," vol. 2, no. 12, pp. 7466–7473, 2018.
[5] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," 2017. [Online]. Available: www.kompas.com.
[6] F. Rahutomo, I. Yanuar, R. Pratiwi, and D. M. Ramadhani, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia Naïve Bayes ' S Experiment On Hoax News Detection In," 2019.
[7] A. Sabrani, I. W. G. P. W. Wedashwara, and F. Bimantoro, "Metode Multinomial Naïve Bayes Untuk Klasifikasi Artikel Online Tentang Gempa Di Indonesia (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia)," 2020. [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/.
[8] A. P. Wijaya and H. A. Santoso, "Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government Naïve Bayes Classification on Document Classification to Identify E-Government Content," *Journal of Applied Intelligent System*, vol. 1, no. 1, pp. 48–55, 2016.
[9] F. R. Aad Miqdad Muadz Muzad, "Korpus Berita Daring Bahasa Indonesia Dengan Depth First Focused Crawling," vol. 01, pp. 15–16, 2016.
[10] Y. Pramudita, U. T. Madura, S. S. Putro, and U. T. Madura, "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," no. August 2018, 2019, doi: 10.25126/jtiik.201853810.
[11] S. Mujilahwati, P. Studi, T. Informatika, F. Teknik, U. I. Lamongan, and D. Mining, "Pre-Processing Text Mining Pada Data Twitter," vol. 2016, no. Sentika, pp. 18–19, 2016.
[12] S. R. Afif, P. Sukarno, and M. A. Nugroho, "Analisis Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Deteksi Serangan Denial of Service (DoS) pada Arsitektur Software Defined Network (SDN)," 2018.
[13] D. H. Kalokasari, I. M. Shofi, and A. H. Setyaningrum, "Implementasi Algoritma Multinomial Naive Bayes Classifier Pada Sistem Klasifikasi Surat Keluar (Studi Kasus : DISKOMINFO Kabupaten Tangerang)," vol. 10, no. 2, 2017, doi: 10.15408/jti.v10i2.6822.
[14] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *Jurnal Ilmu Komputer Agri-Informatika*, vol. 5, pp. 1–10, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jika.
[15] C. Anam and H. Budi Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," vol. 8, no. 1, pp. 13–19, May 2018.