Pengembangan Aplikasi Pendeteksian Hoaks Pada Berita Covid-19 Berbahasa Indonesia Menggunakan Naïve Bayes Bernoulli

Yovi Friangga [1] Yulia Ery Kurniawati [2]

Informatika, Fakultas Ilmu Komputer dan Desain, Institut Teknologi dan Bisnis Kalbis
Jalan Pulomas Selatan Kav.22, Jakarta 13210
[1] *Email: yovifriangga1@gmail.com*
[2] *Email: yulia.kurniawati@kalbis.ac.id*

Abstract: *The main objective of this research is to develop an application that can detect hoax and not hoax at the covid-19 articles with Indonesian language. The dataset was obtained through web scraping method at turnbackhoax.id website on April until November 2020. Then we will perform the labelling of the dataset, which are zero for hoax and one for not hoax. The labelling result of the data obtained is 499 data of COVID-19 articles with Indonesian language. This research is using incremental method as an application development. The incremental method that used is consist into two steps, and those are incremental part one that contains the development of classification model, which are data preprocessing, Bernoulli Naïve Bayes algorithm with TF-IDF as term weighting method, and the testing of classification model with five new testing data. While, incremental part two contains the development of the application user interface. The interface of the application exists two buttons, which are process and exit button. The results of the classification model are 84% of accuracy, 84,15% of precision, 84,0% of recall, and 84,04% of f-measure.*
Keywords: *Hoax detection, COVID-19, Naïve Bayes Bernoulli, TF-IDF.*

Abstrak: *Tujuan dari penelitian ini adalah menghasilkan aplikasi yang dapat mendeteksi hoaks dan bukan hoaks pada berita COVID-19 berbahasa Indonesia. Pengambilan dataset menggunakan metode web scraping melalui situs turnbackhoax.id pada periode April hingga November 2020. Selanjutnya dilakukan pelabelan dataset, yaitu nol untuk data hoaks dan satu untuk data bukan hoaks. Hasil pelabelan diperoleh 499 data berita COVID-19 berbahasa Indonesia. Penelitian ini menggunakan metode inkremental untuk pengembangan aplikasi. Inkremental yang digunakan terdiri dari dua tahap, yaitu inkremental satu berisi tahapan perancangan model klasfikasi, mulai dari preprocessing data, algoritma Naïve Bayes Bernoulli dengan pembobotan kata menggunakan TF-IDF, dan pengujian model klasifikasi dengan lima data uji baru. Inkremental dua berisi tahapan perancangan user interface. Tampilan aplikasi terdapat dua tombol proses dan exit. Model klasifikasi mendapatkan nilai akurasi sebesar 84% dan nilai evaluasi model sebesar, precision 84,15%, recall 84,0%, dan f-measure 84,04%.*
Kata Kunci: *Pendeteksian hoaks, COVID-19, Naïve Bayes Bernoulli, TF-IDF.*

I. PENDAHULUAN

A. Latar Belakang

Saat ini, seluruh dunia sedang berperang melawan pandemi bernama *Corona Virus Disease* 2019 (COVID-19). COVID-19 adalah penyakit yang disebabkan oleh *virus severe acute respiratory syndrome corona virus* dua (SARS-CoV-2) yang dapat menyebabkan gangguan sistem pernapasan, mulai dari gejala yang ringan seperti flu, hingga infeksi paru-paru, seperti pneumonia [1]. Penyebaran COVID-19 juga diiringi dengan banyaknya hoaks mengenai virus tersebut. Menurut data Kementrian Keminfo hingga 8 April 2020

menemukan adanya 474 hoaks terkait COVID-19 yang tersebar di *platform* digital, seperti Facebook, Instagram, Twitter, Yotube, dll [2]. Efek buruk yang ditimbulkan akibat adanya berita hoaks COVID-19 adalah, menggunakan pengobatan yang berpotensi membahayakan nyawa, kurangnya rasa kepercayaan terhadap pemerintah, dan tidak patuh dalam menjalankan protokol kesehatan.

Pemerintah dan komunitas-komunitas pemberantas hoaks terus berupaya memerangi hoaks dengan membangun *website* yang memberikan informasi tentang berita-berita yang mengandung hoaks. Namun, pengembangan *website* dinilai masih kurang berperan besar dalam menghentikan peredaran hoaks dikarenakan masyarakat Indonesia masih malas untuk melakukan literasi dalam menerima kebenaran informasi.

Saat ini penelitian tentang sistem cerdas mulai banyak dikembangkan, Salah satunya penelitian mengenai sistem klasifikasi berita hoaks dilakukan Mandeep Sing, dkk [3]. Penelitian ini mendeteksi berita hoaks menggunakan algoritma *Naïve Bayes Bernoulli* dan membandingkan hasilnya dengan algoritma *Gaussian Naïve Bayes.* Pelabelan data dibagi menjadi dua kelas yaitu, nol dan satu. Nol adalah berita *Fake* dan satu adalah berita dari *Genuine new article.* Hasil penelitian penggunaan *Naïve Bayes Bernoulli* mendapatkan hasil lebih tinggi dari *Gaussian Naïve Bayes* dengan hasil akurasi ditingkatkan besar 10%, presisi sebesar 15%, dan *F1-measure* sebesar 6%.

Berdasarkan latar belakang masalah di atas, penelitian ini akan mengembangkan sistem cerdas yang menghasilkan aplikasi pendeteksi berita teks hoaks dan bukan hoaks pada berita COVID-19 berbahasa Indonesia pada bulan April sampai November 2020. Berdasarkan penelitian terdahulu, penelitian ini akan menggunakan algoritma *Naïve Bayes Bernoulli* dengan penambahan fitur *term frequency–inverse document frequency (TF-IDF)* dalam pembuatan model klasifikasi untuk aplikasi pendeteksian hoaks dan bukan hoaks. Proses pembuatan model klasifikasi memerlukan *dataset* untuk pembelajaran model. Pengambilan *dataset* melalui proses *webscraping* pada *website turnbackhoax.id* karena memiliki banyak data mengenai hoaks dan sudah diakui oleh Kementerian Komunikasi dan Informatika.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka dapat diambil rumusan masalah yaitu:

1. Bagaimana cara mengembangkan aplikasi pendeteksi hoaks dan bukan hoaks pada berita COVID-19 berbahasa Indonesia menggunakan algoritma *Naive Bayes Bernoulli* dengan pembobotan *TF-IDF*?
2. Bagaimana tingkat akurasi dan nilai evaluasi yang didapatkan dari model klasifikasi teks menggunakan algoritma *Naive Bayes Bernoulli* dengan pembobotan *TF-IDF*?

C. Tujuan Penelitian

Adapun tujuan penelitian yang digunakan dalam peneliti ini adalah:

1. Mengembangkan aplikasi pendeteksi hoaks dan bukan hoaks pada berita COVID-19 berbahasa Indonesia menggunakan algoritma *Naive Bayes Bernoulli* dengan pembobotan *TF-IDF*.
2. Mengetahui tingkat akurasi dan nilai evaluasi yang di dapat model klasifikasi teks menggunakan algoritma *Naive Bayes Bernoulli* dengan pembobotan *TF-IDF*.

II. METODE PENELITIAN

A. Teori Pendukung

1. *Naïve Bayes Bernoulli*

Naïve Bayes Bernoulli menerapkan algoritma pelatihan dan klasifikasi, untuk klasifikasi yang menggunakan *binary* nol dan satu [4]. Berikut perhitungan *Naïve Bayes Bernoulli* yang terdapat pada Persamaan 1.1 dan 1.2 [4].

$$P(c|d) = P(c) \times (P(t_1|c) + P(t_2|c) + \dots + P(t\square|c))$$
(1.1)

Keterangan:

$(c|d)$ = Probabilitas *posterior* suatu dokumen d masuk ke dalam kelas c.

$P(c)$. = Probabilitas *prior* dari kelas c.

$P(tk|c)$ = Probabilitas kata ke-k diketahui sebagai kelas c.

$$P(t\square|c) = \frac{Wct + 1}{((\sum W\prime\in VW\prime ct)+B\prime)}$$
(1.2)

Keterangan:

Wct. = Nilai bobot TF-IDF atau W dari *term* t di kelas c.

$\sum W'\in VW'ct$ = Jumlah total W dari keseluruhan *term* di kelas c.

B' = Jumlah W kata unik (nilai idf) pada seluruh dokumen.

2. Pembobotan *TF-IDF*

Pembobotan *TF-IDF* merupakan metode untuk menghitung bobot setiap kata yang sangat umum untuk digunakan pada *information retrieval*. *TF-IDF* juga terkenal mudah, efisien, dan memiliki hasil yang akurat. Berikut rumus perhitungan pada pembobotan TF-IDF pada Persamaan 2.1 [5]:

$$tfi(\mathrm{b}, d, D) = tf(\mathrm{b}, d) \times \log(N/n)$$
(2.1)

Keterangan:

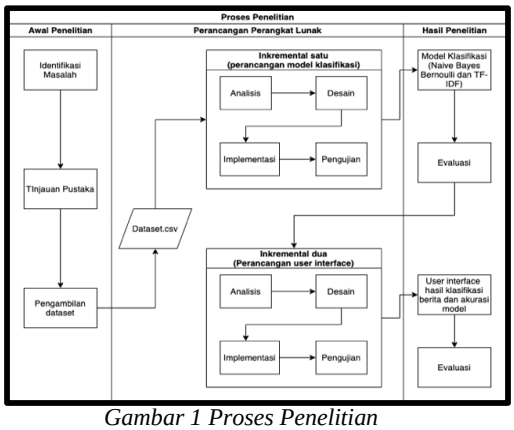
tfidf (b, d, D) = Bobot kata (b) pada sebuah dokumen (d) terhadap sekumpulan dokumen (D).

tf (b, d) = Jumlah kemunculan suatu kata (b) dalam sebuah dokumen (d).

N = Jumlah seluruh dokumen dalam dalam basis data.

n = Jumlah dokumen yang mengandung suatu kata.

B. Proses Penelitian



Gambar 1 Proses Penelitian

Pada Gambar 1 menampilkan alur dalam pengembangan aplikasi. Tahap awal penelitian yaitu melakukan identifikasi masalah. Permasalahan utama dalam penelitian ini yaitu hoaks COVID-19 yang menimbulkan kurangnya kepercayaan terhadap pemerintah, tidak menjalani protokol COVID-19, dan penggunaan pengobatan yang membahayakan nyawa. Sehingga, tujuan penelitian ini adalah mengembangkan aplikasi pendeteksian hoaks dan bukan hoaks pada berita COVID-19 berbahasa Indonesia. Setelah melakukan identifikasi masalah, proses selanjutnya adalah melakukan tinjauan pustaka untuk menemukan metode yang akan digunakan dalam penelitian. Dalam penelitian ini menggunakan algoritma

Naïve Bayes Bernoulli dengan *TF-IDF* untuk membuat model klasifikasi teks.

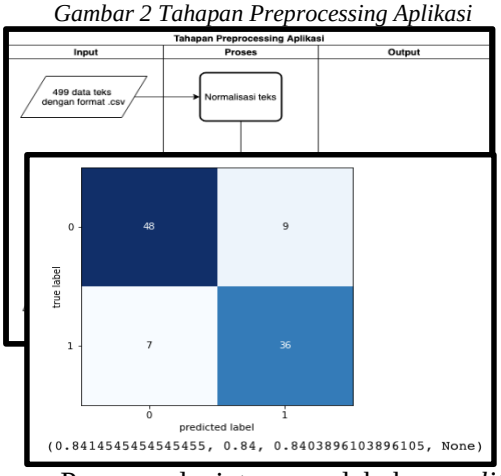
Selanjutnya dilakukan pengumpulan *dataset* untuk dijadikan pembelajaran model. Proses pengambilan data dilakukan dengan cara *scraping* melalui *website turnbackhoax.id. Scraping* data dibatasi hanya berita COVID-19 pada bulan April sampai November 2020. Kemudian data diberi label nol (0) untuk kategori berita hoaks dan label satu (1) untuk kategori berita bukan hoaks. Data yang sudah diberikan label kemudian disimpan dengan format *.csv.*

Proses akhir adalah pengembangan aplikasi menggunakan model inkremental. Inkremental yang digunakan terdiri dari dua tahap, yaitu tahap inkremental satu dan tahap inkremental dua. Tahap inkremental satu adalah perancangan dalam pembuatan model klasifikasi. Inkremental satu terdiri dari beberapa proses, proses pertama adalah *preprocessing.* Terdapat tiga proses *preprocessing* yang digunakan yaitu, *normalisasi*, *stemming*, dan *stopword.* Proses kedua adalah pembuatan model klasifikasi menggunakan algoritma *Naïve Bayes Bernoulli* dengan pembobotan *TF-IDF*. Proses ketiga adalah pengujian model dengan data uji baru untuk mengetahui kemampuan model dalam melakukan pendeteksian hoaks dan bukan hoaks pada berita teks COVID-19 baru. Selanjutnya tahap inkremental dua, yaitu perancangan *user interface*. Dalam penelitian ini, hanya menggunakan tombol proses dan keluar. Tombol proses untuk menjalankan aplikasi dan menampilkan hasil klasifikasi disertai dengan akurasi model. Tombol keluar untuk menghentikan program dan tampilan aplikasi.

III. HASIL DAN PEMBAHASAN

Langkah awal yang akan dilakukan adalah proses *preprocessing* pada *dataset. Dataset* diperoleh dari hasil *web scraping* yang sudah melalui tahap penyeleksian data dan pelabelan data. *Dataset* berjumlah 499 data dengan 240 data dilabeli nol dan 259 data dilabeli satu. Penelitian ini menggunakan tiga proses *preprocessing*, yaitu *normalisasi* teks, *stemming* kata, dan *stopword* kata. Tahapan *preprocessing* pada aplikasi dapat dilihat pada Gambar 2.

Gambar 2 Tahapan Preprocessing Aplikasi



Proses selanjutnya melakukan *split* data hasil *preprocessing*. Pada proses ini, membagi *dataset* hasil *preprocessing* menjadi 80% data *training* dan 20% data *testing*. Kemudian data *training* akan melalui tahap ekstraksi fitur menggunakan *TF-IDF*. Selanjutnya, hasil ekstraksi fitur akan digunakan untuk pelatihan model *Naïve Bayes Bernoulli*. Model *Naïve Bayes Bernoulli* akan diuji dengan data *testing* untuk mendapatkan akurasi model dan nilai evaluasi model. Nilai evaluasi dan akurasi didapatkan dari hasil *confusion matrix* yang menghasilkan perhitungan untuk mendapatkan nilai seperti akurasi, *precision, recall, dan f-measure*. Model klasifikasi disimpan dan akan digunakan pada proses pengujian dengan data uji baru. Gambar hasil *confusion matrix* ditampilkan pada gambar 3.

Gambar 3 Hasil Confusion Matrix

Perhitungan tingkat performansi *classifier* pada tabel *confusion matrix* :

Perhitungan Akurasi :

$$\textbf{Akurasi} = \frac{\mathbf{48 + 36}}{\mathbf{48 + 36 + 7 + 9}} = \mathbf{0,84\ X\ 100} = \mathbf{84}\%$$

Perhitungan *Precision* :

$$\textbf{Precision Hoaks} = \frac{\mathbf{48}}{\mathbf{48 + 7}} = \mathbf{0,8415\ X\ 100} = \mathbf{84,15}\%$$

Perhitungan *Recall* :

$$\textbf{Recall} = \frac{\mathbf{48}}{\mathbf{48 + 7}} = \mathbf{0,84\ X\ 100} = \mathbf{84}\%$$

Perhitungan *F-measure* :

$$\mathrm{F-measure} = 2\ x\ \frac{0{,}8415\ x\ 0{,}84}{0{,}8415 + 0{,}84}\ \mathrm{X}\ 100 = 84{,}04\%$$

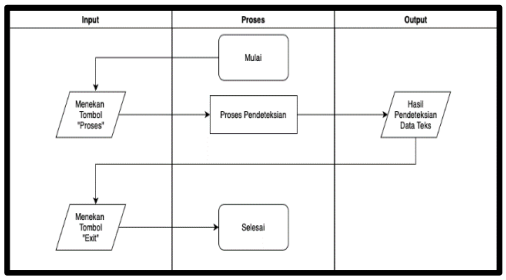
Selanjutnya model akan diuji dengan lima data uji berita teks COVID-19 baru. Pada Tabel 1 dijabarkan hasil pendeteksian pada lima data uji baru.

Tabel 1 Lima Data Uji Berita COVID-19

Data Uji Baru	***Kategori Berita***	***Hasil Prediksi***
Vaksin COVID-19 Dapat Mengubah DNA Manusia.	*Hoaks*	*Hoaks*
Kunyah Daun Sirih Untuk Mengatasi Covid-19.	*Hoaks*	*Hoaks*
Salah Satu Cara Untuk Menangkis Virus Covid-19 Dengan Air Garam dan Air Hangat.	*Hoaks*	*Hoaks*
Pasien yang dirujuk dari puskesmas Kelapa Gading terbanyak terjadi pada tanggal 21 Juli sebanyak 6 orang dan rujukan hanya dilakukan pada hari tertentu.	*Bukan Hoaks*	*Bukan Hoaks*
Menurut covid19.go.id, peta persebaran kasus positif aktif COVID-19 di sebagian wilayah Jakarta saat ini berwarna merah dan kuning, bukan warna merah dan hitam. Update terakhir dari tanggal 13 September 2020 dengan 4 warna zona yang digunakan yaitu warna hijau, kuning, oranye, dan hijau.	*Bukan Hoaks*	*Bukan Hoaks*

Hasil pendeteksian berhasil dilakukan karena model dapat mengkategorikan lima data uji sesuai dengan kategori berita dengan benar.

Langkah terakhir yaitu tahap pembuatan rancangan tampilan aplikasi dan alur pemakaian aplikasi. Tahap alur pemakaian aplikasi, *user* mengisi satu data uji berita COVID-19 kedalam format .csv kemudian *user* menekan tombol "proses" untuk melakukan klasifikasi berita uji menggunakan model yang sudah dibuat pada inkremental satu. Hasil klasifikasi akan muncul berupa akurasi model klasifikasi, prediksi teks tersebut hoaks atau bukan hoaks, dan menampilkan data uji yang diklasifikasi. Untuk keluar dari aplikasi tekan tombol "exit". Berikut



ditampilakan Gambar 4 alur pemakaian aplikasi dan Gambar 5 hasil tampilan aplikasi.

Gambar 4 Alur Pemakaian Aplikasi

Gambar 5 Hasil Tampilan Aplikasi

IV. SIMPULAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut**:**

1. Pengembangan aplikasi pendeteksian berita teks hoaks dan dan bukan hoaks berhasil dilakukan. Pembuatan model klasifikasi teks menggunakan algoritma *Naïve Bayes Bernoulli* dengan pembobotan *TF-IDF*. Dataset berjumlah 499 data yang didapatkan dengan proses *webscraping* melalui *website* turnbackhoax.id. Terdapat dua tombol pada tampilan aplikasi, yaitu tombol proses untuk menampilkan akurasi model, data uji baru yang akan diprediksi dan hasil prediksi.
   Kemudian tombol exit untuk menghentikan tampilan aplikasi. Tahap pengujian dilakukan dengan memprediksi lima data uji baru dan hasilnya aplikasi dapat mengklasifikasi dengan benar.

2. Didapatkan nilai akurasi sebesar 84%, *precision* 84,15%, *recall* 84%, dan *f-measure* 84,04%. Dari hasil yang diperoleh dapat disimpulkan bahwa model dapat melakukan klasifikasi dengan hasil yang cukup baik.

B. Saran

Berdasarkan kekurangan yang ada, dibutuhkan beberapa saran untuk pengembangan pada penelitian selanjutnya. Saran tersebut antara lain:

1. Menambah *dataset* dengan data berita COVID-19 periode waktu terbaru.

2. Menambah proses N-gram pada pembuatan model klasifikasi untuk menambah akurasi model dan memasukkan *form input* untuk mempermudah dalam memasukkan data uji secara langsung.

DAFTAR RUJUKAN

[1] M. Pane, "Covid-19," Alodokter, 2021. https://www.alodokter.com/covid-19 (accessed Nov. 20, 2020).

[2] K. Luxina, "Masa Pandemi Corona, Kominfo Temukan 474 Isu Hoax di Facebook-Youtube *Detiknews.com*, 2020. https://news.detik.com/berita/d-4969636/masa-pandemi-corona-kominfo-temukan-474-isu-hoax-di-facebook-youtube.

[3] M. Sing, M. Bhatt, H. Bedi, and U. Mishra, "Performance of Bernoillis's Naive Bayes Classifier in the detection of Fake News," vol. 3, p. 1, 2020.

[4] F. Fanesya and R. C. Wihandika, "Deteksi Emosi pada Twitter Menggunakan Metode Naive Bayes dan Kombinasi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, p. 3, 2019.

[5] D. R. Hidayat, "Penerapan Metode Multinomial Naive Bayes Dalam Pengembangan Aplikasi Klasifikasi Terhadap Cuitan Pada Media Sosial Twitter," Kalbis Insitute, 2019.