Deteksi Hoaks Pada Berita Berbahasa Indonesia Seputar COVID-19

Amanda Tabitha Bulan Panjaitan[1] dan Ibnu Santoso[2]
Program Studi Komputasi Statistik, Politeknik Statistika STIS [1,2]
Jl. Otto Iskandardinata No. 64C, Jatinegara, Jakarta, 13330
E-mail : 221709539@stis.ac.id[1], ibnu@stis. ac.id[2]

Abstract **-- The development of increasingly advanced technology certainly brings many conveniences for its users, but on the other hand it also accelerates the spread of fake news on the internet. Fake news or known as hoax is misleading and dangerous information because it misleads human perceptions by conveying false information as truth. Hoax itself can aim to influence readers with false information so that readers take action according to the contents of the hoax. Therefore, we need an intelligent system that is able to classify a story quickly that spreads via the internet so as not to mislead its readers. This research begins by doing news scraping that has been labeled as hoax or valid. The data set is divided into two types, which are training data and test data. Pre-processing is carried out starting from case folding, tokenization, filtering and stemming. In this study, a comparison is made of the effect of feature engineering. From the results of accuracy, it can be seen that the application of engineering features can improve the accuracy of five classification methods. The random forest method with the application of engineering features produces an accuracy rate of 96.05%.**

Key Words**: *web scraping, data mining, classification, feature engineering, hoax***

Abstrak **-- Perkembangan teknologi yang semakin maju tentu mendatangkan banyak kemudahan bagi para penggunanya namun di lain sisi juga mempercepat penyebaran berita bohong pada internet. Berita bohong atau dikenal dengan hoaks adalah informasi sesat dan berbahaya karena menyesatkan persepsi manusia dengan menyampaikan informasi palsu sebagai kebenaran. Hoaks sendiri dapat bertujuan untuk mempengaruhi pembaca dengan informasi palsu sehingga pembaca mengambil tindakan sesuai dengan isi hoaks. Oleh karena itu, diperlukan sistem cerdas yang mampu mengklasifikasi sebuah berita dengan cepat yang menyebar melalui internet agar tidak menyesatkan para pembacanya. Penelitian ini dimulai dengan melakukan *scraping* berita yang sudah diberi kategori hoaks atau valid. *Dataset* tersebut dibagi dua menjadi data latih dan data uji. Dilakukan *pre-processing* mulai dari *case folding, tokenizing, filtering* dan *stemming*. Pada penelitian ini dilakukan perbandingan terhadap pengaruh penerapan *feature engineering*. Dari hasil akurasi, dapat dilihat bahwa dengan diterapkannya *feature engineering* mampu meningkatkan akurasi kelima metode klasifikasi. Metode *random forest* dengan penerapan *feature engineering* menghasilkan tingkat akurasi sebesar 96,05%.**

Kata Kunci**: *web scraping, data mining, klasifikasi, feature engineering, hoaks***

I. PENDAHULUAN

1.1. Latar Belakang

Semakin berkembangnya zaman, semakin berkembang pula teknologi yang ada. Teknologi mengalami perkembangan dari waktu ke waktu dan tingkat kecanggihannya pun semakin tinggi. Perkembangan teknologi ini erat kaitannya dengan frekuensi penggunaan internet yang semakin tinggi. Era internet zaman ini mampu menghadirkan berbagai kemudahan yang dapat memenuhi kebutuhan masyarakat akan informasi maupun pemanfaatan untuk kepentingan sosial ekonomi. Akan tetapi ada pula dampak lain kehadiran internet, yakni membuka ruang lebar bagi kehadiran informasi atau berita-berita bohong tentang suatu peristiwa yang meresahkan publik dikenal sebagai hoaks. Data yang disampaikan oleh Rudiantara, Menteri Komunikasi dan Informatika (Menkominfo) Republik Indonesia, menunjukkan bahwa penyebaran hoaks dan ujaran kebencian diindikasikan berasal dari 800 ribu situs di Indonesia[1].

Salah satu hoaks yang juga banyak beredar melalui media sosial adalah hoaks seputar Covid-19. Dilansir dari ScienceAlert, Kamis (13/8/2020), dalam studi ini, sekelompok peneliti penyakit menular internasional menganalisis berbagai media sosial dan situs berita untuk mengetahui bagaimana misinformasi terkait Covid-19 yang menyebar di internet. Hasil penelitian mereka menemukan sekitar 2.300 laporan hoaks dan teori konspirasi Covid-19 dalam 25 bahasa di 87 negara. Dalam publikasi *The American Journal of Tropical Medicine and Hygiene* mengungkapkan bahwa misinformasi terkait Covid-19 ini telah menelan setidaknya 800 korban jiwa di seluruh dunia[2].

Adapula bentuk partisipasi pemerintah untuk mengantisipasi atau meminimalkan jumlah berita bohong yang beredar di masyarakat adalah dengan meluncurkan laman TurnBackHoax.id. TurnBackHoax bekerja sama dengan Masyarakat Anti Fitnah dan *Hoax* Indonesia (Mafindo) untuk membantah kabar burung, mengklarifikasi informasi yang simpang siur dan memberikan penjelasan. Metode identitas atau klasifikasi yang dilakukan pada situs tersebut masih dilakukan secara manual, sehingga jika informasi semakin berkembang akan kesulitan dikarenakan

informasi yang masuk semakin banyak. Oleh karena itu diperlukan suatu metode untuk mengklasifikasi berita secara cepat. *Data mining* merupakan proses ataupun kegiatan untuk mengumpulkan data yang berukuran besar kemudian mengekstraksi data tersebut menjadi informasi-informasi yang nantinya dapat digunakan. Salah satu tugas yang dapat dilakukan dengan *data mining* adalah pengklasifikasian. Metode klasifikasi ini mampu memberikan label sebuah berita sebagai hoaks atau valid berdasarkan ekstrasi informasi dari berita tersebut. Dalam penelitian ini kan dilakukan klasifikasi menggunakan model pendekatan *data mining* sehingga klasifikasi dapat dilakukan oleh sistem secara otomatis.

1.2. Tujuan Penelitian

Berdasarkan rumusan masalah diatas, tujuan dilakukannya penelitian ini adalah:

1. Membangun model klasifikasi untuk memprediksi kebenaran suatu berita yang dipengaruhi oleh karakteristik berita tersebut dan mengukur kebaikan model.
2. Menggunakan model untuk klasifikasi secara otomatis.
3. Mengukur pengaruh penerapan *feature engineering* pada kebaikan model.

1.3. Penelitian Terkait

Pada penelitian klasifikasi hoaks dengan metode KNN oleh Prasetyo, Indriati, dan Adikara mampu menghasilkan akurasi sebesar 94,12%. Penelitian ini menggunakan 51 berita telah dilabeli oleh pakar, 67 dilabeli oleh tim hoax buster dan 52 berita sisanya dilabeli oleh situs portal berita itu sendiri[3]. Adapula penelitian yang membandingkan tiga jenis algoritma untuk mendeteksi tumor payudara. Penelitian ini membandingkan teknik klasifikasi *Support Vector Machine, Extreme Gradient Boosting* dan *Multilayer Perceptron*[17]. Algoritma MLP menghasilkan akurasi terbaik namun memiliki kompleksitas yang tinggi. Oleh karena itu dipilih SVM/XGBoost karena memberikan akurasi yang tidak jauh berbeda namun lebih sederhana.

Penelitian pengaruh penerapan *feature engineering* oleh Rawat dan Khemchandani menyatakan bahwa proses ini memang memakan waktu yang lama namun sangat baik untuk mengefektifkan pembelajaran mesin pada model. Membangun model klasifikasi yang efisien untuk masalah klasifikasi dengan beberapa fitur relevan tambahan bersama dengan kumpulan data yang berbeda dan ukuran sampel yang berbeda adalah penting. Tugas utamanya adalah membuat fitur tambahan yang relevan dari kumpulan fitur yang ada, lalu pemilihan fitur untuk membuang fitur yang berlebihan, tidak relevan, atau sangat berkorelasi[18].

II. METODOLOGI PENELITIAN

2.1. Pengumpulan Data

Data dikumpulkan melalui *web scraping* dengan kata kunci “Covid-19”,”Corona” dan “Pandemi” pada situs Turnbackhoax.id dan Detik.com dari tanggal 18 November hingga 20 November. Dari hasil pengumpulan data didapatkan 1262 berita yang kemudian disimpan dalam ekstensi CSV.

Adapula informasi yang diambil dari *web scraping* yakni judul dan isi artikel. Data fakta diperoleh dari situs https://www.detik.com. Judul berita diambil dari *text* yang berada pada tag <h1> dengan class: 'detail__title' dan artikel berita diambil pada tag <p> dengan class: 'detail__body-text'. Data yang diambil kemudian disimpan dengan nama DataDetik dengan ekstensi CSV. Data hoaks diperoleh dari situs https://turnbackhoax.id. Judul berita diambil dari *text* yang berada pada tag <h1> dengan class: 'entry-title' dan artikel berita diambil pada tag <p> dengan class: 'entry-content mh-clearfix'. Data yang diambil kemudian disimpan dengan nama DataTurnBackHoax dengan ekstensi CSV.

2.2. Definisi

2.2.1. *Web Scraping*

Web Scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman web dalam bahasa markup seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain[4]. Hasil dari *web scraping* dapat dimanfaatkan kembali oleh sistem lain dan dianalisis lebih lanjut [16].

2.2.2. *Preprocessing*

Sebelum dilakukan analisa terhadap data latih dan data uji, perlu dilakukan *preprocessing*. Adapula tujuannya yakni untuk mengubah data input mentah menjadi format yang sesuai untuk analisis selanjutnya. Adapun tahapan *preprocessing* berdasarkan, yaitu: *case* folding, *tokenizing*, *filtering*, dan *stemming*[5].

1. *Case Folding*

Case folding adalah tahapan proses mengubah semua huruf dalam teks dokumen menjadi huruf kecil.

2. *Tokenizing*

Tokenizing adalah tahap pemotongan teks input menjadi kata, istilah, symbol, tanda baca, atau elemen lain yang memiliki arti yang disebut token[6]. Pada proses, token yang merupakan tanda baca yang dianggap tidak perlu seperti titik (.), koma (,), tanda seru (!), tanda tanya(?) dan lain-lain akan dihapus.

Tokenizing adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya sehingga menjadi kata-kata tunggal yang kemudian dapat dilakukan analisa. Kata, angka, simbol, tanda baca dan entitas penting lainnya pada suatu teks dapat dianggap sebagai token.

3. *Filtering*

Filtering merupakan proses pemilihan kata-kata penting dari hasil token yaitu kata-kata yang dapat digunakan untuk mewakili dokumen. *Filtering* dapat dilakukan dengan menggunakan *stopword removal*. Proses *stopword removal* pada dataset disesuaikan dengan koleksi *stopword list* yang dimiliki. Pada penelitian ini menggunakan *stopword list* kelas *StopWordRemoverFactory* dari library *Sastrawi*.

4. *Stemming*

Stemming merupakan suatu proses mengubah ekstraksi sebuah kata menjadi bentuk dasar dengan cara menghilangkan imbuhan[7]. Penerapan proses *stemming* dalam setiap bahasa berbeda-beda tergantung pada morfologi dari setiap bahasa tersebut. Tujuan utama dari proses ini adalah untuk mengetahui makna dari sebuah kata walaupun sudah mempunyai bentuk yang berbeda.

5. *Term Weighting*

Term weighting merupakan proses penghitungan bobot tiap *term* yang dicari pada setiap dokumen sehingga dapat diketahui ketersediaan dan kemiripan suatu *term* di dalam dokumen[15].

Term weighting yang paling populer adalah *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) [8]. Metode algoritma ini berguna untuk menghitung bobot suatu kata muncul pada dokumen. TF menentukan berapa seringnya kata muncul dalam sebuah dokumen. Jadi, semakin banyak frekuensi kemunculan dari kata tersebut, semakin besar pula nanti nilainya. Berbeda dengan TF, IDF merupakan sebuah perhitungan dari bagaimana *term* didistribusikan secara luas pada koleksi dokumen yang bersangkutan. *Term* yang paling jarang muncul akan memiliki bobot tertinggi dan yang paling sering muncul akan memiliki bobot paling rendah. Nilai bobot TF-IDF dapat dihitung dengan menggunakan persamaan (3).

$$TFt,d = \{1 + log\ log\ (ft,d), \quad ft,d > 0 \quad 0, \quad x = 0 \qquad (1)$$

$$IDFt = log\left(\frac{N}{dft}\right) \qquad (2)$$

$$TFIDFt,d = TFt,d \times IDFt \qquad (3)$$

2.2.3.*Feature Engineering*

Feature Engineering adalah proses memilih dan mengubah variabel saat membuat model prediktif menggunakan pembelajaran mesin atau pemodelan statistik. Proses ini melibatkan kombinasi analisis data, penerapan aturan praktis, dan penilaian. *Feature Engineering* juga merupakan teknik yang paling penting untuk mencapai hasil yang baik pada tugas prediksi[9]. Namun di sisi lain, teknik ini sulit untuk dipelajari dan dikuasai karena data yang berbeda membutuhkan teknik yang berbeda pula.

Feature pertama yang digunakan adalah melihat apakah penulisan huruf awal pada judul dan isi artikel menggunakan huruf kapital atau huruf kecil. Asumsinya, teks bukan hoaks biasanya mengikuti ejaan yang disempurnakan, termasuk penggunaan huruf kapital pada kata pertama dalam suatu kalimat. Adapula *feature* lainnya yang menghitung kemunculan tanda tanya dan tanda perintah pada judul dan isi artikel.

Kemudian dihitung pula kemirpan judul dan isi artikel dengan metode *cosine similarity*. Adapula fungsi *cosine similarity* ditulis dengan persamaan(4).

$$sim(x,y) = \frac{x.y}{||x||||y||'} \qquad (4)$$

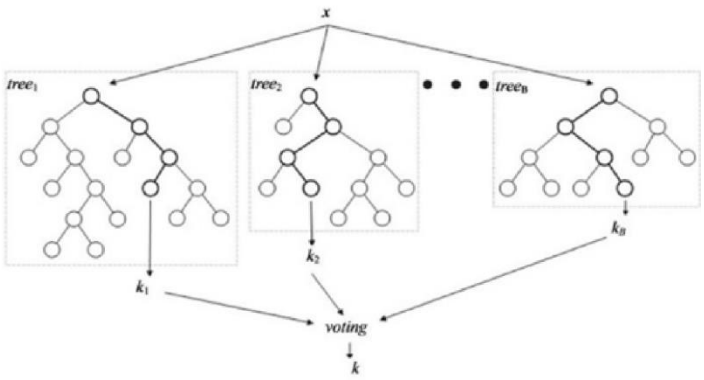
Dimana || x || adalah norma Euclidean dari vektor x = (x1, x2, ..., xp), didefinisikan sebagai x12 + x22 + ⋯ + xp2. Demikian pula, || y || adalah norma Euclidean dari vektor y. Pengukuran tersebut menghitung cosinus sudut antara vektor x dan y. Nilai cosinus mendekati 1 mengindikasikan semakin kecil sudut yang terbentuk, artinya semakin besar kecocokan antara kedua vector. Sebaliknya, apabila nilai cosinus mendekati 0, maka sudut yang terbentuk sebesar 90 derajat, yang artinya tidak memiliki kecocokan.

2.2.4. Metode Klasifikasi

Dalam penelitian ini menggunakan lima jenis metode klasifikasi, antara lain:

1. *Random Forest*

Metode *random forest* bekerja dengan cara mengkombinasikan masing – masing tree yang baik ke dalam satu model. Metode ini merupakan pengembangan dari metode *Classification and Regression Tree* (CART), yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection*[5]. *Random Forest* adalah metode klasifikasi yang terdiri dari kumpulan pengklasifikasi pohon terstruktur {h(x,Θk), k=1, ...} dimana {Θk} merupakan vektor acak terdistribusi yang identik dan independen dimana masing-masing pohon melemparkan unit suara untuk kelas paling populer diinput[10].

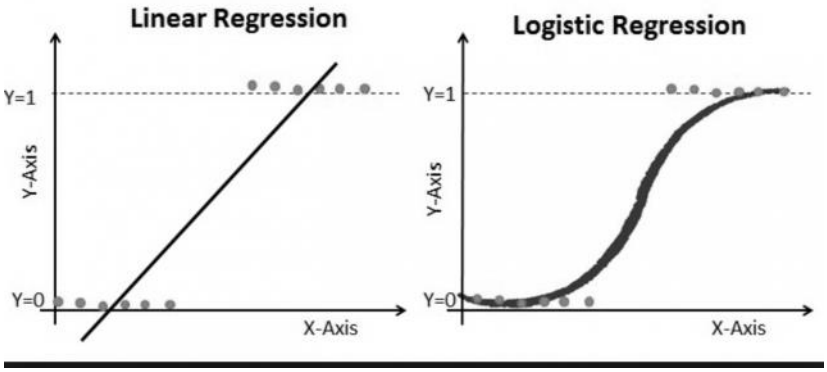


Gambar 1. *Random Forest Classifier*

Random forest melalui proses pengacakan yang tidak hanya dilakukan pada data sampel saja melainkan juga pada pengambilan variabel bebas sehingga pohon klasifikasi yang dibangkitkan akan memiliki ukuran dan bentuk yang berbeda-beda[11]. *Random forest* adalah klasifikasi yang terdiri dari beberapa pohon keputusan yang dibangun dengan menggunakan vektor acak.

Algoritma ini adalah pengembangan dari Algoritma C4.5 dengan menggunakan beberapa *decision tree*, dimana *training data* telah dilakukan pada setiap *decision tree* dengan menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut subset yang bersifat acak. Dan dalam perkembangannya, sejalan dengan bertambahnya *dataset*, maka *tree* pun ikut berkembang.

2. *Logistic Regression*

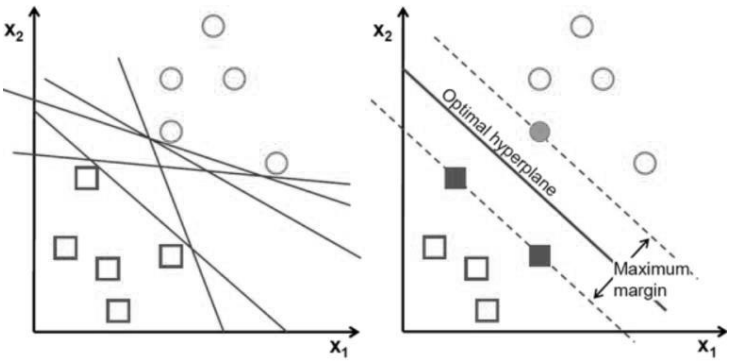


Gambar 2. *Linier Vs Logistic Regression*

Algoritma klasifikasi dengan *logistic* bekerja untuk mencari hubungan antara fitur (input) diskrit/kontinu dengan probabilitas hasil output diskrit tertentu. Adapun tujuan dari algoritma[12] ini yakni untuk memprediksi probabilitas terjadinya atau tidak terjadinya suatu kejadian berdasarkan nilai-nilai prediktor yang ada. Tujuan lainnya yakni untuk mengklasifikasikan subjek penelitian berdasarkan ambang (*threshold*) probabilitas.

3. *Support Vector Machine* (SVM)

SVM digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. *Hyperplane* adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas.

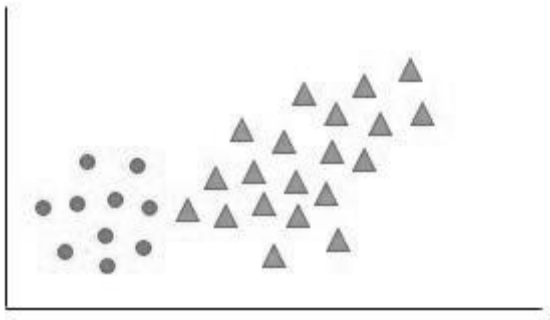


Gambar 3. *Support Vector Machine*

Hyperplane yang ditemukan posisinya berada ditengah-tengah antara dua kelas, artinya jarak antara *hyperplane* dengan objek-objek data berbeda dengan kelas yang berdekatan (terluar) yang diberi tanda bulat kosong dan positif. *Support vector* merupakan objek data terluar yang paling dekat dengan *hyperplane* disebut, objek ini sulit diklasifikasikan dikarenakan posisi yang hampir tumpang tindih (*overlap*) dengan kelas lain. Mengingat sifatnya yang kritis, hanya *support vector* inilah yang diperhitungkan untuk menemukan *hyperplane* yang paling optimal oleh SVM[13].

4. *Naïve Bayes*

Metode *Naive bayes* merupakan pengklasifikasi probabilitas sederhana yang didasarkan pada teori Bayes. Teori ini menyatakan bahwa peluang terjadinya suatu peristiwa dapat dihitung dengan mengalikan probabilitas intrinsik (dihitung dari data yang tersedia sekarang) dengan probabilitas bahwa hal serupa akan terjadi lagi di masa depan (berdasarkan pengetahuan yang terjadinya di masa lalu).



Gambar 4. *Naïve Bayes Classifier*

Rumus dibawah ini merupakan cara menghitung probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas). Di mana X merupakan data dengan *class* yang belum diketahui dan H merupakan hipotesis data merupakan suatu *class* spesifik.

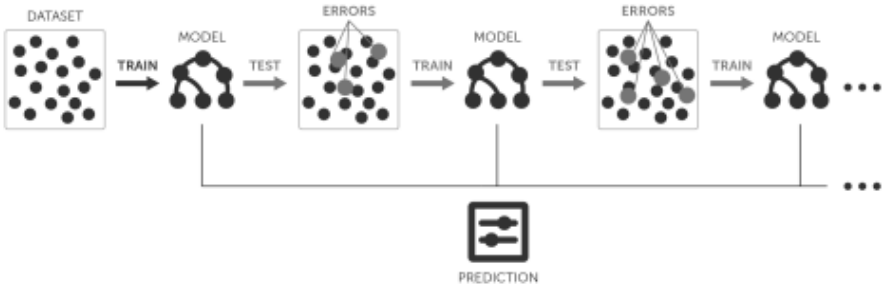
$$P(X) = P(X|H).P(H)P(X) \qquad (5)$$

Untuk menjelaskan metode *Naïve Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naïve Bayes* di atas disesuaikan sebagai berikut:

$$P(C|F1 \dots Fn) \qquad (6)$$
$$= \frac{P(C)P(F1 \dots Fn|C)P(F1 \dots Fn)}{P(F1 \dots Fn)}$$

Di mana Variabel C merepresentasikan kelas, sementara variabel F1 ... Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi.

5. *Gradient Boosting*



Gambar 4. *Gradient Boosting Classifier*

Gradient Boosting pertama kali diperkenalkan oleh J.H. Friedman. GBT mampu membangun *decision tree* berdasarkan peningkatan dalam struktur pohon pada pembelajaran yang lemah untuk memperbaiki kesalahan pohon dan mencegah terjadinya potensi *overfitting*. Dalam membangun *decision tree*, dapat dilakukan penambahan jumlah iterasi yang sangat konservatif yang dapat menghasilkan dan meningkatkan kinerja model yang lebih baik. GBT mampu memecahkan masalah dengan menyesuaikan pembelajaran lemah dengan gradien negatif dari fungsi kerugian (*loss function*) dan meningkatkan pohon (*trees*) dengan parameter yang mewakili variabel *split* yang dipasang pada setiap *node* terminal pohon[14].

2.2.5. Evaluasi Model

Sistem yang baik adalah sistem yang mampu mengklasifikasikan data secara tepat, namun tidak dapat dipungkiri bahwa kinerja suatu sistem klasifikasi tidak bisa benar secara sempurna. Oleh karena itu, diperlukan cara untuk mengukur kinerja sistem klasifikasi yang telah dibangun.

Pengukuran kinerja standar untuk masalah klasifikasi yang biasanya digunakan adalah akurasi. Namun, jika suatu *dataset* memiliki kelas dengan distribusi yang tidak seimbang, maka perlu digunakan nilai sensitivitas sebagai pengukuran tambahan.Dalam penelitian ini, sensitivitas merupakan ukuran yang mengukur seberapa baik sebuah model dalam memprediksi suatu artikel berita hoaks sebagai artikel berita hoaks. Dengan kata lain, sensitivitas adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Selain mengukur sensitivitas, perlu dilakukan pengukuran terhasap presisi dan nilai f1-score hasil klasifikasi. Presisi adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. F1-score adalah rata-rata dari presisi dan sensitivitas yang memperhitungkan kedua metrik tersebut dalam persamaan berikut:

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$sensitivitas = \frac{TP}{TP + FN} \quad (8)$$

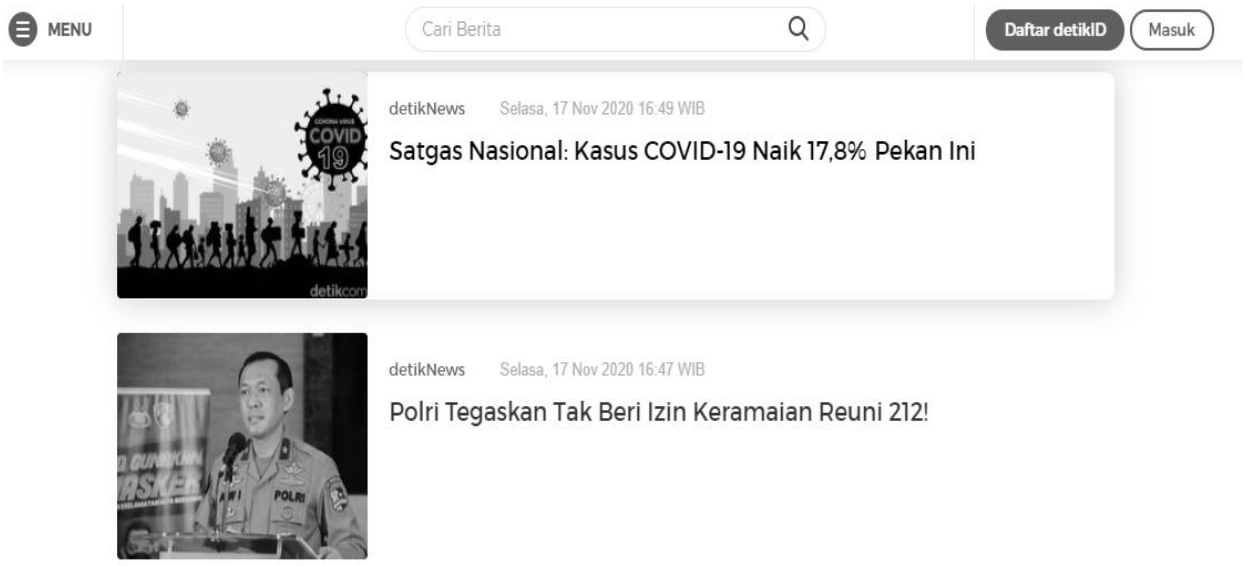
$$presisi = \frac{TP}{TP + FP} \quad (9)$$

$$f1 - score = \frac{2 * presisi * sensitivitas}{presisi + sensitivitas} \quad (10)$$

III. HASIL DAN PEMBAHASAN

Penelitian dimulai dengan *scraping* berita pada situs TurnBackHoax.id dan Detik.com. Adapula kata kunci yang digunakan untuk mendapatkan data berita yakni “Covid”, “Corona” dan “Pandemi”. Berikut merupakan tampilan kedua laman ketika dilakukan penelusuran.

Gambar 5. Hasil Pencarian dengan Kata Kunci pada Situs TurnBackHoax.id



Gambar 6. Hasil Pencarian dengan Kata Kunci pada Situs Detik.com

Setelah link berita di *scrap*, masing-masing berita akan dicari judul serta isi artikelnya. Adapun tampilan dari kedua situs berita tersebut seperti gambar yang di bawah ini.

FORUM | LAPOR HOAX | LAYANAN PUBLIK | MEDIA | RELAWAN | TENTANG KAMI

[SALAH] Pengumuman Beasiswa Covid-19 Rp4 Juta dari STMIK Royal Kisaran Melalui SMS

October 8, 2020 Bentang Febrylian Fitnah / Hasut / Hoax 0

Informasi palsu. Melalui media sosialnya, STMIK Royal Kisaran menyatakan pesan tersebut adalah hoaks. Pihak kampus mengimbau agar selalu mengecek segala bentuk informasi di situs dan media sosial resmi.Selengkapnya terdapat di penjelasan!

KATEGORI: FABRICATED CONTENT/KONTEN PALSU

===

SUMBER: PESAN BERANTAI WHATSAPP dan SMS

===

Gambar 7. Tampilan Artikel Berita Situs TurnBackHoax.id

MENU | Cari Berita

detikNews > Berita

Satgas Nasional: Kasus COVID-19 Naik 17,8% Pekan Ini

Yulida Medistiara - detikNews

Selasa, 17 Nov 2020 16:49 WIB

0 komentar SHARE

Ilustrasi (Edi Wahyono/detikcom)

Jakarta - Juru bicara Satgas Penanganan COVID-19 Wiku Adisasmito meng-update kasus positif COVID-19 mingguan. Pada pekan ini, Wiku mengatakan perkembangan kasus COVID-19 di tingkat nasional mengalami peningkatan 17,8 persen.

"Perkembangan kasus positif mingguan. Di minggu ini terlihat perkembangan kasus COVID-19 di tingkat nasional cenderung ke arah yang kurang baik dibandingkan pekan sebelumnya. Pada level nasional, di pekan ini terjadi peningkatan kasus positif sebesar 17,8 persen," kata Wiku dalam YouTube Setpres, Selasa (17/11/2020).

Ia mengatakan pekan ini peningkatan kasus positif secara signifikan dari biasanya. Sebab, selama ini hanya di kisaran 5-8 persen. Karena itu, Wiku meminta masyarakat memperhatikan lagi dan mematuhi protokol kesehatan memakai masker, mencuci tangan, menjaga jarak, dan menghindari kerumunan.

Gambar 8. Tampilan Artikel Berita Situs Detik.com

Setelah judul dan isi artikel berita didapat, dilanjutkan dengan tahap *preprocessing.* Tahap ini dimulai dengan melakukan *case folding*, *tokenizing*, *filtering*, dan *stemming.* Kemudian dilakukan penghapusan pada kata yang kemunculannya kurang dari lima kali. Berikut merupakan hasil dari sampel *preprocessing* teks dibandingkan dengan teks asli :

Tabel 1. Tabel Gabungan Judul dan Isi Artikel Berita Asli

Judul + Isi Artikel
Kanada Sudah Terbebas dari Covid-19 Informasi tersebut menyesatkan. Tidak benar jika Kanada sudah terbebas dari virus corona atau Covid-19. Merujuk pada kanal resmi WHO, data menunjukkan pada 10 Agustus 2020
Campaign “Jaga Jogja” MAFINDO Yogyakarta Dokumentasi Campaign “Jaga Jogja” MAFINDO Yogyakarta pada hari Sabtu-Minggu, 01-02 Agustus 2020 di Tugu Yogyakarta bersama 19 organisasi yang ada di D.I.Yogyakarta. Campaign “Jaga Jogja”

Tabel 2. Tabel Gabungan Judul dan Isi Artikel Berita Setelah *Preprocessing*

Judul + Isi Artikel
kanada bebas covid informasi sebut sesat benar kanada bebas virus corona covid rujuk kanal resmi data tunjuk agustus
campaign jaga jogja mafindo yogyakarta dokumentasi campaign jaga jogja mafindo yogyakarta hari sabtu minggu agustus tugu yogyakarta sama organisasi d yogyakarta campaign jaga jogja

Pada penelitian ini digunakan beberapa metode klasifikasi dengan *preprocessing* yange berbeda pula. Pada *dataset* yang sama dilakukan dua jenis pengujian, yakni dengan menerapkan dan tidak menerapkan *feature engineering*. Adapun hasil akurasi dari berbagai teknik dapat dilihat pada tabel berikut:

Tabel 3. Evaluasi Model

Metode	Dengan Feature Engineering					Tanpa Feature Engineering				
	RF LR SVM	LR	SVM	NB	GBT	RF	LR	SVM	NB	GBT
Akurasi	**96.05**	92.09	91.53	90.96	90.40	93.79	90.96	91.53	90.40	89.27
Presisi	**92.31**	91.67	90.59	87.78	85.26	89.25	89.53	91.57	86.81	82.83
Sensitivitas	**100**	91.67	91.67	94.05	96.43	98.81	91.67	90.48	94.05	97.62
F1 Score	**96**	91.67	91.12	90.8	90.5	93.79	90.59	91.02	90.29	89.62

Dari Tabel 3 dapat dilihat bahwa dengan menerapkan *feature engineering*, akurasi kelima metode yang digunakan mengalami kenaikan. Metode *random forest* pada kedua kelompok (diterapkan dan tidak diterapkan *feature engineering*) menghasilkan akurasi tertinggi, namun dengan penerapan *feature engineering* mampu menghasilkan tingkat akurasi lebih baik dibandingkan tidak terapkan, yakni sebesar 96.05%. Setelah *radom forest*, metode klasifikasi dengan *logistic regression* menduduki posisi akurasi kedua tertinggi, kemudian SVM. Naïve Bayes dan yang terkahir merupakan *gradient boosting*.

Jika melihat nilai evaluasi model pada algoritma SVM, penerapan *feature engineering* tidak memberikan dampak positif seperti pada algoritma lainnya. Hal ini dapat disebabkan oleh ketidaksesuaian *feature* yang digunakan pada algoritma ini.

IV. KESIMPULAN

Dari hasil dan analisis hasil uji coba yang telah dijelaskan, dapat disimpulkan bahwa berdasarkan evaluasi model keseluruhan, model terbaik dihasilkan oleh algoritma *random forest* dengan menerapkan *feature engineering*. Model ini menghasilkan akurasi, presisi, sensitivitas dan f1-score secara berturut-turut sebesar 96.05%, 92.31%, 100%, dan 96%.

Dari kelima algoritma yang digunakan, berdasarkan nilai evaluasi model, *random forest* merupakan algoritma yang paling tepat digunakan untuk mengklasifikasi berita hoaks berbahasa Indonesia.

Kesimpulan lain adalah, penerapan *feature engineering* mampu meningkatkan nilai dari evaluasi model. Tahapan ini tentu memerlukan pengetahuan yang baik terhadap konteks pemodelan yang sedang dihadapi serta kreativitas dalam menentukan fitur yang sesuai. Apabila fitur yang digunakan kurang tepat, maka akan menghasilkan evaluasi model yang tidak baik pula.

V. DAFTAR PUSTAKA

[1] Kominfo. (2017, Desember 12). *Ada 800.000 Situs Penyebar Hoax di Indonesia*. Dipetik Oktober 3, 2020, dari https://kominfo.go.id/content/detail/12008/ada-800000-situs-penyebar-hoax-di-indonesia/0/sorotan_media

[2] Forbes. (2020, August 23). *Report: More Than 800 Deaths And 5,800 Hospitalizations Globally May Have Resulted From COVID-19 Misinformation Online*. Dipetik Oktober 3, 2020, dari https://www.forbes.com/sites/markhall/2020/08/23/coronavirus-misinformation/#dc3c9f01684e

[3] Prasetyo, A. R., Indriati, & Adikara, P. P. (2018). *Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, II(12), 7466-7473.

[4] Turland, M. (2010). *php|architect's Guide to Web Scraping.* Los Angeles: Marco Tabini & Associates, Inc.

[5] Triawati, C. (2009). *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia.* Bandung: Institut Teknologi Telkom.

[6] Vijayarani, S., & Janani, R. (2016). *Text Mining: Open Source Tokenization Tools, an Analysis*. Advanced Computational Intelligence: An International Journal, *III*, 37-47.

[7] Narulita, L. F. (2018). Pengaruh Proses Stemming Pada Kinerja Analisa Sentimen Pada Review Buku. *Jurnal Hasil Penelitian LPPM Untag Surabaya, III*(1), 55- 59.

[8] Ren, F., & Sohrab, M. G. (2013). *Class-indexing-based term weighting for automatic text classification*. Inf. Sci., *236*, 109-125.

[9] Nurhikmat, T. (2018). *Implementasi Deep Learning Untuk Image Classification Menggunakan AlgoritmaConvutional Neural Network (CNN) Pada Citra Wayang Golek.* Yogyakarta: Universitas Islam Indonesia.

[10] Bramer, M. (2007). *Principles of Data Mining London.* Springer Clark.

[11] Liaw, A., & Wiener, M. (2002). Classification and Regression by Randomforest. *II*.

[12] Rosadi, D. (2011). *Analisis Ekonometrika dan Runtun Waktu Terapan dengan.* Yogyakarta: Andi Offset.

[13] Samsudiney. (2019, July). *Penjelasan Sederhana tentang Apa Itu SVM?* Diambil kembali dari https://medium.com/@samsudiney/penjelasansederhana-.

[14] Friedman, J. (2014*). Greedy Function Approximation: A Gradient Boosting*. Ann. Stat., *29*(5), 1189–1232.

[15] Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., et al. (2017). *Modified frequency-based term weighting schemes for text classification*. Applied Soft Computing Journal *58*, 193-206.

[16] Rahmatullah, A. & Gunawan, R. (2020). *Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar*. Indonesian Journal of Information Systems, II(2), 95-104.

[17] Handayani, A., Jamal, A., Septiandri, A. A. (2017). *Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara*. JNTETI, VI(4), 394-403.

[18] Rawat, Tara & Khemchandani, Vineeta. (2019). *Feature Engineering (FE) Tools and Techniques for Better Classification Performance*. IJIET, VIII(2), 169-179.