Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes

Fani Prasetya, Ferdiansyah*

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Bina Darma, Palembang, Indonesia
Email: 1,*ferdi@binadarma.ac.id
Email Penulis Korespondensi: ferdi@binadarma.ac.id
Submitted: **16/09/2022**; Accepted: **26/09/2022**; Published: **30/09/2022**

Abstrak– Penyebaran informasi yang begitu cepat seiring pesatnya perkembangan teknologi seiring masifnya kecepatan media elektronik dan internet. Tetapi penyebaran berita yang secara cepat ini tidak dapat menjamin informasi dan berita yang kita peroleh dapat di validasi dari sumber yang valid. Berdasarkan data yang dirilis oleh Kominfo pada akhir tahun 2021 terdapat 1773 berita hoax yang berhasil di klarifikasi dari berita-berita hoax tersebut. Lalu selama pandemi Covid-19 sendiri, ada beragam hoaks yang beredar di masyarakat. Sepanjang 2021, Kementerian Kominfo menemukan sebanyak 723 hoaks seputar Covid-19. Berdasarkan latar belakang diatas peneliti dan pada penelitian sebelumnya, telah membahas tentang deteksi hoax pada berbagai bidang. Seperti, deteksi penipuan pada gaya penulisan daring [1], klasifikasi berita hoax berbasis pembelajaran mesin[3] dan Penerapan Algoritma naïve bayes dan pso untuk klasifikasi berita hoax pada media sosial [4]. Dari sini peneliti mencoba melalukan eksperimen pada algoritma klasifikasi naïve bayes untuk mengklasifikasikan berita hoax covid 19. Berdasarkan hasil penelitian yang telah dilakukan yang model naïve bayes dan cross validation dapat melakukan klasifikasi berita hoax dengan baik, akurasi yang dihasilkan sebesar 86.3% dimana 80-90% masuk pada kriteria good classification. Data yang diprediksi salah juga tidak terlalu banyak dari total 300 dataset hanya 41 yang dinyatakan salah dalam pelabelan tidak sampai 2% dari keseluruhan total dataset, sehingga dapat disimpulkan model ini dapat digunakan sebagai referensi apabila ingin dilanjutkan pada model prediksi yang lebih kompleks lagi, misalnya model prediksi menggunanakan machine learning berbasis web.

Kata Kunci: Covid-19; Klasifikasi; Hoax; Naive Bayes

Abstract–The rapid dissemination of information along with the rapid development of technology along with the massive speed of electronic media and the internet. But the rapid spread of news cannot guarantee that the information and news that we get can be validated from valid sources. Based on data released by Kominfo at the end of 2021, there were 1773 hoax news that were successfully clarified from the hoax news. Then during the Covid-19 pandemic itself, there were various hoaxes circulating in the community. Throughout 2021, the Ministry of Communications and Informatics discovered as many as 723 hoaxes about Covid-19. Based on the background above, the researchers and previous studies have discussed hoax detection in various fields. Such as, fraud detection in online writing style [1], classification of hoax news based on machine learning [3] and the application of nave Bayes and PSO algorithms for classification of hoax news on social media [4]. From here the researchers tried to carry out experiments on the nave Bayes classification algorithm to classify hoax covid 19 news. Based on the results of research that has been done, the nave Bayes model and cross validation can classify hoax news well, the resulting accuracy is 86.3% where 80-90% included in the good classification criteria. The data that is predicted to be incorrect is also not too much from a total of 300 datasets, only 41 are declared incorrect in labeling less than 2% of the total dataset, so it can be concluded that this model can be used as a reference if you want to proceed to a more complex prediction model, for example the model prediction using web-based machine learning.

Keywords: Covid-19; Classification; Hoax; Naive Bayes

1. PENDAHULUAN

Penyebaran informasi yang begitu cepat seiring pesatnya perkembangan teknologi. Penyebaran media cetak seperti surat kabar, majalah dan sebagainya bersaing dengan masifnya kecepatan media elektronik dan internet. Internet berkembang sebagai media penyebar informasi yang cepat disertai dengan fasikitas pencarian dari media pencarian di internet seperti google, sehingga mendukung masifnya penyebaran informasi dan berita di internet. [5]

Tetapi penyebaran berita yang secara cepat ini tidak dapat menjamin informasi dan berita yang kita dapat dapat di validasi dari sumber yang valid. Sehingga banyak sekali berita-berita palsu dari sumber yang tidak terpercaya tersebar di internet. Beberapa tahun belakangan ini , dunia informasi di berbagai negara termasuk juga Indonesia diserang dengan maraknya berita-berita palsu atau hoax, hoax sendri berasal dari istilah (hocus to trick) yang dibuat dengan tujuan memanipulasi atau mengundang orang untuk melakukan suatu tindakan menggunakan ancaman atau penipuan.[6]

Berdasarkan data yang dirilis oleh Kominfo pada akhir tahun 2021 terdapat 1773 berita hoax yang berhasil di klarifikasi misinformasi dari berita-berita hoax tersebut. Lalu selama pandemi Covid-19 sendiri, ada beragam hoaks yang beredar di masyarakat. Sepanjang 2021, Kementerian Kominfo menemukan sebanyak 723 hoaks seputar Covid-19. [7]Tentu saja ini berdampak kurang baik dan menimbulkan banyak multipersepesi dkalangan masyarakat yang saat ini juga masyarakat Indonesia berada di peringkat 62 dari 70 negara yang memiliki tingkat literasi yang rendah.[8]

Berdasarkan latar belakang diatas peneliti dan Pada penelitian-penelitian sebelumnya, telah membahas tentang deteksi hoax pada berbagai bidang. Seperti, deteksi penipuan pada gaya penulisan daring [1]. Dari penelitian menunjukkan bahwa dengan menggunakan set fitur yang besar, dimungkinkan untuk membedakan

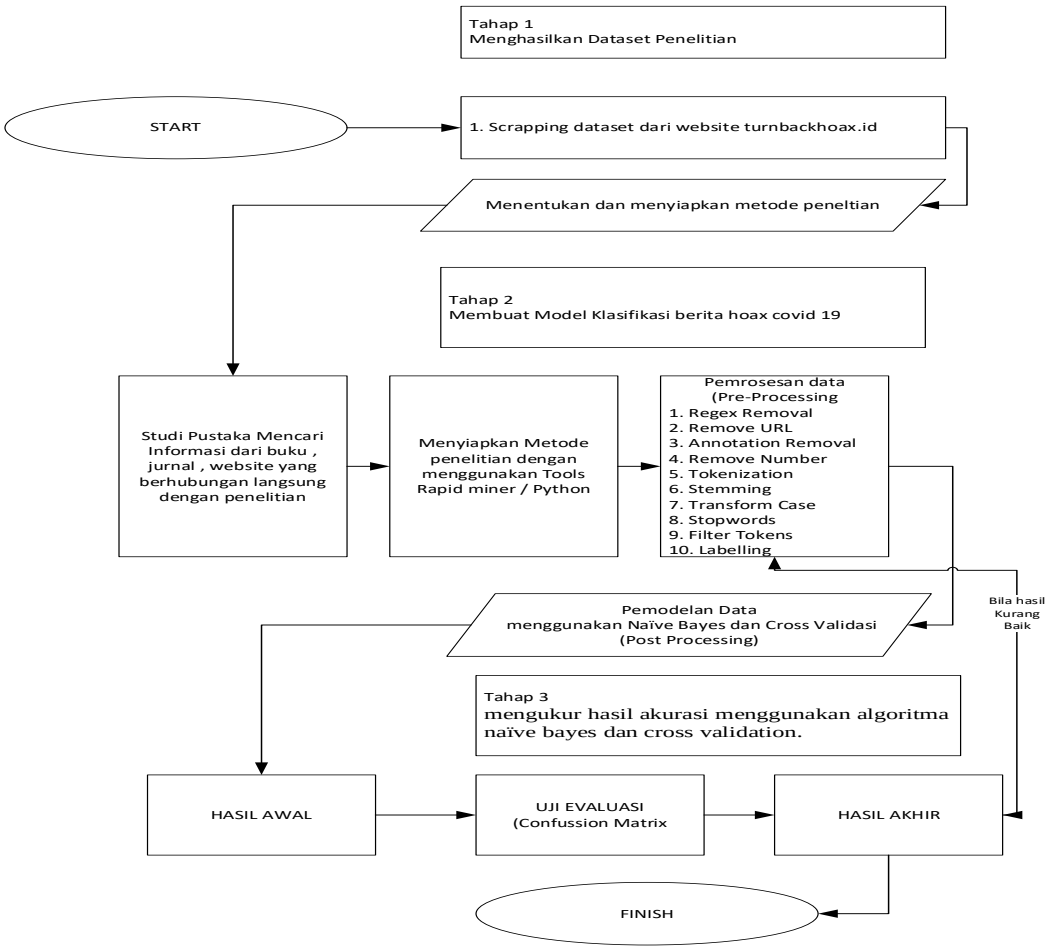
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

dokumen biasa dari dokumen palsu dengan akurasi 96,6% (F-measure). Klasifikasi ulasan asli dan palsu pada toko daring [2] penelitian ini menggunakan 10 supervised algoritma pembelajaran untuk menganalisis sejauh mana otentik dan

ulasan palsu dapat dibedakan berdasarkan empat linguistic petunjuk, yaitu keterpahaman, tingkat detail, gaya penulisan dan indikator kognisi, hasilnya secara umum menjanjikan. Pada penelitian klasifikasi berita hoax berbahasa Indonesia berbasis pembelajaran mesin[3] penelitian ini menggunakan eksraksi fitur, tanpa stemming dan stopword tetapi menggunakana ekstraksi bigram dan unigram akruasi terbaik dicapai oleh naïve bayes sebesar 91.36% dibanding SVM dan C4.5, pada penerapan algoritma Naïve bayes dan PSO untuk klasifikasi berita hoax pada media sosial [4] hasil yang di dapat adalah sebesar 85.19% sedangkan hasil yang di dapat tanpa PSO sebsar 74.67%. Dari sini peneliti mencoba melalukan eksperimen menggunakan algoritma klasifikasi Naïve bayes untuk mengklasifikasikan berita hoax yang secara spesifik menganalisa berita terkait covid 19 dan dari beberapa penelitian tersebut Naive bayes terlihat unggul dari algoritma lain sehingga sangat tepat digunakan untuk penelitian ini, agar hasilnya dapat di jadikan rekomendasi untuk mendeteksi berita hoax covid 19 secara Nasional.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian



Gambar 1. Tahapan Metode Penilitian

Berdasarkan gambar 1. Penelitian dilakukan melalui beberapa tahapan, yaitu;

1. Tahap 1 Menghasilkan Data penelian
   Langkah yang dilakukan adalah mengumpulkan data penelitian dengan teknik *scrapping* text di *website* turnbackhoax.id dan dari beberapa berita di media masa dan social media dari tahun 2019 sampai 2022 selama covid 19 berlangsung. Dataset ini tergolong dalam unstructured dataset dan termasuk dalam unsupervised learning dimana dataset harus diberikan label terlebih dahulu dalam hal ini akan terbagi dalam dua label hoax dan non hoax, dimana hal tersebut perlu dilakukan sebelum di proses kedalam machine learning untuk di preprocessing, dan di cek validasi akurasinya terhadap pelabelan tersebut. Sementara labelnya sendiri dilihat dari label yang sudah di anggap positif hoax dan bukan hoax oleh website turnbackhoax.id, mafindo (masyarakat anti fitnah Indonesia dan beberapa sumber lainnya. Untuk pelabelan sendiri dilakukan secara

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

manual dengan mengambil informasi dari sumber-sumber tadi dengan dua kelas pelabelan yaitu Positif Hoax dan Negatif Hoax.

2. Tahap 2 Menghasilkan model klasifikasi berita hoax covid 19, Prosesnya sebagai berikut :
   a. Studi Pustaka Mencari Informasi dari buku , jurnal , website yang berhubungan langsung dengan penelitian
   b. Menyiapkan Metode penelitian dengan menggunakan Tools Rapid miner/python
   c. Pemrosesan data dilakukan dengan beberapa cara penjelasan secara detail sudah di jelaskan pada bab sebelumnya, semua langkah ini tidak semua dilakukan, akan dilakukan secara ekspremintal agar mendapat hasil akurasi yang lebih baik, dikarenakan tidak ada standard khusus yag mengharuskan semua tahapan pre-processing dilakukan . tetapi tahapan secara umum sebagai berikut
      a) *Regex Removal*
      b) *Remove URL*
      c) *Annotation Removal*
      d) *Remove Number*
      e) *Tokenization*
      f) *Stemming*
      g) *Transform Case*
      h) *Stopwords*
      i) *Filter Tokens*
      j) *Labelling*
3. Tahap 3 Mengukur akurasi positif hoax dan negatif hoax terhadap berita menggunakan *Naïve bayes* dan cross validation. setelah tahapan diatas dilalui dan telah memiliki hasil awal, menggunakan *Naïve bayes* dan *Cross validation* dilanjutkan dengan mengevaluasi hasil menggunakan confussion matrix yang akan memberikan hasil akhir , tapi apabila hasil yang didapat kurang maksimal dapat mengulang proses kembali ketahan filtering atau preprocessing.
4. Selesai.

2.2 Text Mining

Text mining adalah ilmu yang bertujuan untuk memproses teks agar menjadi informasi yang diperoleh dari peramalan pola dan kecenderungan melalui pola statistik. Teks yang diolah bisa berupa teks terstruktur dan teks tidak terstruktur. Text mining mengacu pada information retrieval, data mining, machine learning, statistik dan komputasi [9]. Text mining bertujuan untuk menganalisis pendapat, sentimen, evaluasi, sikap, penilaian, emosi seseorang seihngga dapat diketahui apakah berkenaan dengan suatu topik, layanan, organisasi, individu, atau kegiatan tertentu [10].Penggunaan dari text mining dilakukan untuk klasterisasi, klasifikasi, information retrieval, dan information extraction [11].

2.3 Text Processing

Preprocessing teks adalah proses mengubah format data teks tidak terstruktur ke format terstruktur sesuai kebutuhan untuk mengekstrak informasi dalam proses penambangan. Teks yang diolah disebut istilah, dan setiap istilah dalam dokumen memiliki beberapa arti yang tidak ada artinya. Istilah yang tidak berarti dihapus dari dokumen dan istilah yang berarti tetap ada dan merupakan representasi dari dokumen. ([12])
Berikut ini proses penjelasan tahap *text preprocessing* :

1. Regex Removal : menghilangkan regular expression yang ada dalam teks.
2. Remove URL : menghapus URL yang terdapat dalam teks.
3. Annotation Removal : menghilangkan tanda @ (annotation) yang ada dalam teks.
4. Remove Number : menghapus angka yang terdapat dalam teks.
5. Tokenization : proses tokenisasi pada data teks adalah melakukan pemecahan sekumpulan kalimat menjadi potongan karakter atau kata-kata sesuai kebutuhan yang sering disebut token, sehingga menjadi kata yang memiliki arti tertentu.
6. Stemming : menghilangkan imbuhan yang terdapat pada masing-masing kata sehingga menjadi kata dasar, juga bertujuan untuk membersihkan suatu kata dengan pengejaan yang kurang tepat. Algoritma stemming untuk bahasa yang satu berbeda dengan algoritma stemming untuk bahasa lainnya.
7. Transform Case : mengubah semua huruf besar atau kapital dalam data menjadi huruf kecil ataupun sebaliknya. Hal ini dilakukan agar ketika masuk ke tahap pemodelan klasifikasi terdapat keseragaman huruf dan tidak terjadi kesalahan dalam proses tokenize, yang biasa digunakan adalah mengubah semua huruf menjadi huruf kecil (lower case).
8. Filter Stopwords : sebuah proses untuk menghilangkan kata-kata yang tidak mempunyai arti yang biasanya merupakan kata sambung, kata keterangan dan sebagainya pada hasil parsing sebuah dokumen teks dengan cara membandingkannya dengan stoplist berisi kata-kata yang terlalu sering muncul dalam dokumen-dokumen, belum tentu berguna dalam proses retrieval, kemungkinan besar tidak akan memberikan pengaruh prediksi. Kata-kata yang tidak berguna nantinya akan dibuang dan tidak dijadikan index term. Tahap ini

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

merupakan proses untuk melakukan filter terhadap kata-kata umum seperti "the", "a", "it", "they" dan lainnya, yang tidak diperlukan saat pemrosesan data.

9. Filter Tokens (By Length) : menghilangkann kata dengan panjang huruf tertentu. Misalnya minimal 2 karakter dan maksimal 25 karakter. Artinya kata yang panjangnya hanya 1 karakter dan lebih dari 25 karakter
10. Labelling : merupakan tahap dimana hasil dari tahapan sebelumnya akan dilakukan perhitungan terhadap polarity dari ulasan yang terambil, sehingga dapat menghasilkan dua kategori yaitu label positif dan negatif, untuk label netral (nilai = 0) tidak diproses.

2.4 Klasifikasi Naïve Bayes

Naive Bayes adalah klasifikasi paling sederhana dan paling umum digunakan. Naive Bayes menghitung probabilitas kelas berdasarkan distribusi kata dalam dokumen. NaiveBayes memiliki beberapa keunggulan, seperti kesederhanaan, kecepatan dan keakuratan. Namun, klasifikasi ini memiliki batasan penting dan tidak selalu memenuhi hipotesis independensi antar atribut. Dan hal ini berpengaruh pada tingkat akurasi klasifikasi [13] Adapun persamaan teorema naive bayes sebagai berikut

$$p(x|y) = \frac{\mathrm{p(y|x)p(x)}}{\mathrm{p(y)}} \quad (1)$$

Keterangan :
y = data dengan kelas yang belum di ketahui
x = hipotesis data y merupakan suatu klas spesifik
P(x|y) = probalitas hipotesis x berdasarkan kondisi y (*posteriori probability*)
P(x) = probabilitas hipotesis x (*prior probability)*
P(y|x) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = Probalitas dari y
Untuk menjelaskan teorema *Naive bayes* perlu di ketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apakah cocok bagi sample yang di analisis tersebut. Teorema *bayes* di atas di sesuaikan sebagai berikut

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (2)$$

2.5 Cross Validation

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran dataset. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. [14] Metode cross-validation digunakan untuk menghindari overlapping pada data testing

2.6 Confussion Matrix

Confusion Matrix merupakan metode yang dapat digunakan untuk menghitung keakuratan proses klasifikasi. Anda dapat menggunakan matriks konfusi untuk menganalisis seberapa baik pengklasifikasi mengenali record di kelas yang berbeda. Ini adalah tabel *Confusion Matrix* [15]

Tabel 1. Confusion Matrix

Aktual	**Prediksi**	
	Positif	**negatif**
Positif	TP	FN
Negatif	FP	TN

Keterangan dari *confusion matrix* tersebut sebagai berikut :

1. TP (True Positive) merupakan jumlah data dengan kelas aktual positif dan target kelas prediksi positif.
2. FN (False Negative) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas negatif.
3. FP (False Positive) merupakan jumlah data dengan kelas aktual positif dan target kelas prediksi negatif.
4. TN (True Negative) merupakan jumlah data dengan kelas aktual negatif dan target kelas prediksi negatif.

2.7 Akurasi

Akurasi adalah metode pengujian berdasarkan tingkat kedekatan antara nilai yang diharapkan dan nilai aktual. Mengetahui jumlah data yang diklasifikasikan dengan benar dapat digunakan untuk memverifikasi keakuratan hasil prediksi.

$$Accuracy = \frac{(\mathrm{TP+TN})}{\mathrm{TP+TN+FP+FN}} \quad (3)$$

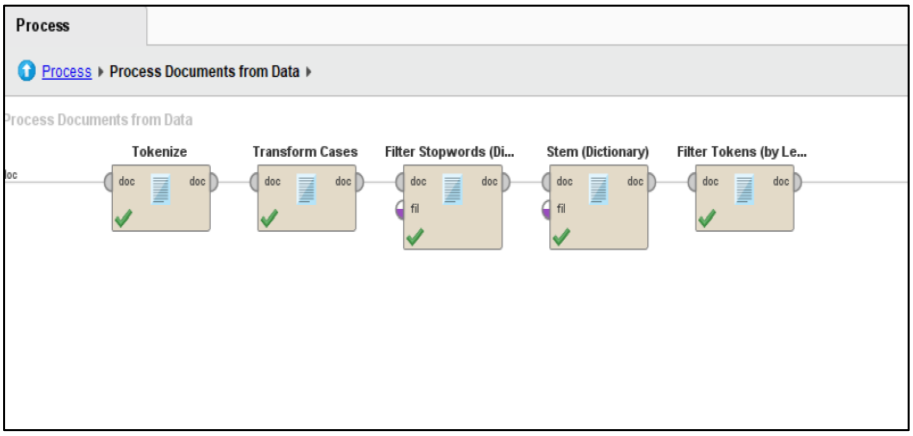
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

2.8 Dataset

Sumber data dari penelitian ini adalah data kualitatif , merupakan data sekunder, diambil dari dataset publik yang dapat diakses oleh masyarakat umum. Sumber berasal dari dua website informative https://turnbackhoax.id dan https://liputan6.com yang mana dua sumber berita tersebut menyediakan berita-berita terkait covid-19 secara faktual serta akurat, sementara turnbackhoax sendiri salah satu laman website independent pencari fakta terkait berita covid-19 yang dikelola oleh MAFINDO (Masyarakat anti fitnah/(hoax) Indonesia).
Dataset tersebut terdiri dari 300 baris , 250 baris merupakan data dengan kategori hoax dan 50 data sisanya merupan data dengan kategori fakta atau valid. Diberi label secara manual dengan melihat keterangan dan informasi yang diberikan pada sumber berita tersebut.

3. HASIL DAN PEMBAHASAN

Setelah data dikumpulkan dalam bentuk format excel xls. Langkah selanjutnya melakukan preprocessing data melalui 4 tahapan yaitu tokenization, transform cases, stopwords, stemming, dan terakhir tokenize by length.



Gambar 2. Preprocessing Documents

Seperti yang dijelaskan pada pembahasan 2. Text processing dimana tokenization adalah memecah sekumpulan kata menjadi karakter yang memiliki arti tertentu dan diberikan nilai contoh kata Achmad memiliki bobot 0.105 yang menandakan achmad memiliki nilai berdasarkan algoritma tokenize yang ada pada rapid miner.

abdul	about	acara	achmad	acid	acuan	acute	adegan	adenovirus	adhan	adhanom	adisasmito
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.207	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.105	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Gambar 3. Proses *Tokenization*

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

3.2 Stopwords

Teks sebelum dilakukan proses filter Stopwords.	Teks setelah dilakukan proses filter Stopwords.
WHO: “Saran penyakit Coronavirus (COVID-19) untuk umum: membantah Mitos. Virus COVID-19 dapat ditularkan di daerah dengan iklim panas dan lembabDari bukti sejauh ini, virus COVID-19 dapat ditularkan di SEMUA AREA, termasuk daerah dengan cuaca panas dan lembab. Apa pun iklimnya, lakukan tindakan perlindungan jika Anda tinggal di, atau bepergian ke area yang melaporkan COVID-19. Cara terbaik untuk melindungi diri dari COVID-19 adalah dengan sering membersihkan tangan. Dengan melakukan ini, Anda menghilangkan virus yang mungkin ada di tangan Anda dan menghindari infeksi yang dapat terjadi saat menyentuh mata, mulut, dan hidung Anda.	Who saran penyakit coronavirus covid 19 umum membantah mitos virus covid 19 ditularkan daerah iklim panas lembab bukti sejauh ini virus covid-19 ditularkan semua area, termasuk daerah dengan cuaca panas lembab. apa pun iklimnya, lakukan tindakan perlindungan jika anda tinggal bepergian ke area melaporkan covid 19 cara terbaik untuk melindungi diri covid 19 adalah sering membersihkan tangan. dengan melakukan ini anda menghilangkan virus yang mungkin tangan anda menghindari infeksi dapat terjadi saat menyentuh mata mulut hidung anda

3.3 Stemming

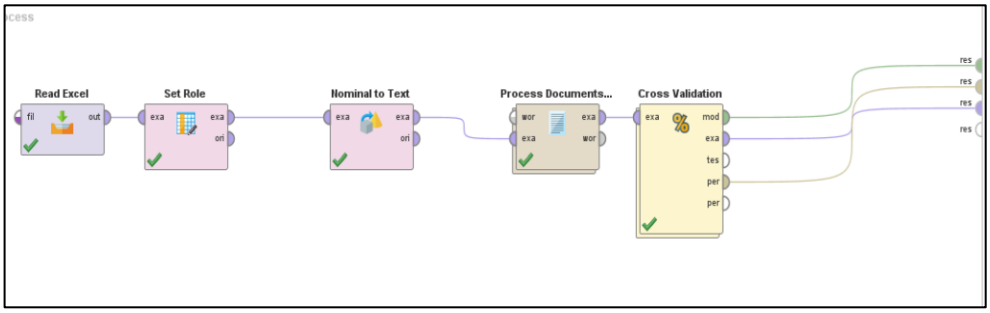
Proses ini dilakukan dengan menghilangkan kata imbuhan menjadi kata dasar dengan tujuan menghilangkan ejaan yang kurang tepat berikut filter stemming dalam imbuhan Bahasa Indonesia yang dihilangkan.

:-(.*)$:-*(lah|kah|tah|pun)$:-*(ku|mu|nya)$:(is|isme|isasi|i|kan|an)$:^(di|ke|se)$:-*(ku|mu|nya)$

Teks sebelum dilakukan proses Stemming	Teks setelah dilakukan proses filter Stemming
WHO: “Saran penyakit Coronavirus (COVID-19) untuk umum: membantah Mitos. Virus COVID-19 dapat ditularkan di daerah dengan iklim panas dan lembabDari bukti sejauh ini, virus COVID-19 dapat ditularkan di SEMUA AREA, termasuk daerah dengan cuaca panas dan lembab. Apa pun iklimnya, lakukan tindakan perlindungan jika Anda tinggal di, atau bepergian ke area yang melaporkan COVID-19. Cara terbaik untuk melindungi diri dari COVID-19 adalah dengan sering membersihkan tangan. Dengan melakukan ini, Anda menghilangkan virus yang mungkin ada di tangan Anda dan menghindari infeksi yang dapat terjadi saat menyentuh mata, mulut, dan hidung Anda.	Who saran penyakit corona virus covid 19 umum membantah mitos virus covid 19 tular daerah iklim panas lembab bukti sejauh ini virus covid-19 tular semua area, termasuk daerah dengan cuaca panas lembab. apa pun iklim, laku tindak perlindung jika anda tinggal berpergi area lapor covid 19 cara terbaik untuk lindungi diri covid 19 adalah sering membersih tangan dengan laku ini anda menghilang virus yang mungkin tangan anda menghindari infeksi dapat terjadi saat nyentuh mata mulut hidung anda.

Sedangkan fungsi transform case hanya mengubah huruf menjadi lowercase dalam hal ini diseragamkan lalu fungsi token by length fungsinya adalah agar penilaian berdasarkan jumlah minimal karakter dan maksimal jumlah karakter, minimal karakter yang di set adalah 4 karakter sedangkan maksimal 25 karakter per kata.

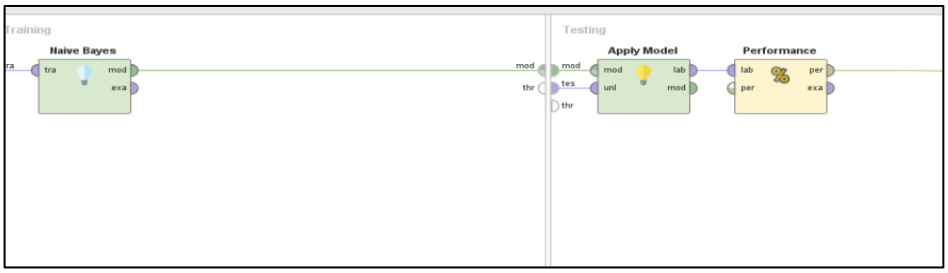
3.4 Model Naïve bayes



Gambar 4. Model eksperimen Naïve bayes

Proses diatas menggambarkan keseluruhan proses penggabungan antara prerprocessing pada tahapan sebelumnya dilanjutkan dengan pengujian dengan algoritma Naïve bayes dan cross validation, pada pengujian ini tidak menggunakan metode membagi data menjadi data latih dan data uji, karena sudah diwakili penggunaan fungsi cross validation dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852



Gambar 5. Didalam mode cross validation

Seperti yang dijelaskan sebelumnya algoritma yang digunakan adalah model naïve bayes yang dipadukan dengan cross validation di atas adalah penerapannya pada rapid miner.

Gambar 6. 10 folds validation

Gambar 6. merupakan keterangan set folds yang berarti pengujian mengunakan cross validation akan dilakukan selama 10 kali lalu kemudian dirata-ratakan untuk mendapatkan hasil validasi yang lebih baik.

accuracy: 86.33% +/- 5.97% (micro average: 86.33%)

	true VALID	true HOAX	class precision
pred. VALID	21	12	63.64%
pred. HOAX	29	238	89.14%
class recall	42.00%	95.20%	

Gambar 7. Hasil Akurasi klasifikasi berita hoax

Setelah melalui proses pengujian dengan menggunakan naïve bayes dan cross validation gambar 7. merupakan hasil dari confussion matrix yang menampilkan hasil sebenarnya dari dataset yang sudah dilatih kedalam model naïve baye. Seperti yang dijelaskan pada pembahasan dataset pada bagian metode dataset terdiri dari 250 berita hoax dan 50 berita yang valid, setelah dilakukan pengujian hasil yang didapatkan dan diklasifikasikan kembali oleh model ternyata ada yang berbeda. Berikut perhitungannya berdasarkan hitungan manual confussion matrix :

$$Accuracy = \frac{(238+21)}{238+21+29+12} = 86.33\%$$

Secara garis besar akurasi yang dihasilkan oleh model ini terbilang cukup baik diatas 85%, dari hasil pengujian didapatkan pergeseran ataupun perubahan kalimat yang sebelumnya dilabelkan hoax ternyata masuk ke kategori valid atau bukan hoax. Dari 250 berita hoax yang ada dalam dataset 238 dataset yang benar hoax true Hoax ada 238 sedangkan yang di prediksi hoax ternayata Valid ada 29, lalu untuk berita yang di prediksi valid dan benar valid (True Valid) ada 21, sedangkan untuk yang di prediksi valid ternyata hoax ada 12.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dijabarkan sebelumnya dapat diambil kesimpulan dari hasil penelitian ini adalah. Model penelitian yang diambil yaitu naïve bayes dan cross validation dapat melakukan klasifikasi berita hoax dengan baik, akurasi yang dihasilkan sebesar 86.3% dimana 80-90% masuk pada kriteria good classification. Data yang diprediksi salah pun tidak terlalu banyak dari total 300 dataset hanya 41 yang dinyatakan salah dalam pelabelan tidak sampai 2% dari keseluruhan total dataset, sehingga dapat disimpulkan model ini dapat digunakan sebagai referensi apabila ingin dilanjutkan pada model prediksi yang lebih kompleks lagi, misalnya dalam model

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 1, September 2022
e-ISSN 2685-998X
DOI 10.30865/json.v4i1.4852

prediksi menggunanakan machine learning berbasis web yang sedang trend skrang sebagai kelanjutan dari system pakar yang sebelumnya sudah ada, saran selanjutnya adalah menggunakan dataset yang lebih besar lagi dan dari berbagai macam bahasa, dan model preprocessing document yang lebih baik lagi, pada model ini peneliti hanya menggunakan spesifik preprocessing dalam Bahasa Indonesia, jika ingin mendapatkan hasil yang lebih kaya lagi dapat dilanjutkan dengan model algoritma deep learning yang lebih beragam metode pengujiannya.

REFERENCES

[1] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *2012 IEEE Symposium on Security and Privacy*, 2012, pp. 461–475.
[2] S. Banerjee, A. Y. K. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proceedings of the 9th international conference on ubiquitous information management and communication*, 2015, pp. 1–7.
[3] E. Rasywir and A. Purwarianti, "Eksperimen pada sistem klasifikasi berita hoax berbahasa Indonesia berbasis pembelajaran mesin," *J. Cybermatika*, vol. 3, no. 2, 2016.
[4] R. Wati and others, "Penerapan Algoritma Naive Bayes Dan Particle Swarm Optimization Untuk Klasifikasi Berita Hoax Pada Media Sosial," *JITK (Jurnal Ilmu Pengetah. Dan Teknol. Komputer)*, vol. 5, no. 2, pp. 159–164, 2020.
[5] L. Ishwara, *Catatan-catatan jurnalisme dasar*, vol. 1. Penerbit Buku Kompas, 2005.
[6] S. Kasman, "Sistem Verifikasi Menangkal Berita Hoax di Media Cetak," *J. Mimb. Kesejaht. Sos.*, vol. 2, no. 1, 2019.
[7] T. Kompas, "Data Sebaran Hoaks Sepanjang 2021, Terbanyak soal Pandemi Covid-19," 2022. [Online]. Available: https://www.kompas.com/tren/read/2022/01/03/163216365/data-sebaran-hoaks-sepanjang-2021-terbanyak-soal-pandemi-covid-19?page=all. [Accessed: 16-May-2022].
[8] Novrizaldi, "Tingkat Literasi Indonesia Memprihatinkan, Kemenko PMK Siapkan Peta Jalan Pembudayaan Literasi Nasional." [Online]. Available: https://www.kemenkopmk.go.id/tingkat-literasi-indonesia-memprihatinkan-kemenko-pmk-siapkan-peta-jalan-pembudayaan-literasi#:~:text=Berdasarkan survei yang dilakukan Program,yang memiliki tingkat literasi rendah. [Accessed: 16-May-2022].
[9] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques Third Edition [M]," *Morgan Kaufmann Ser. Data Manag. Syst.*, vol. 5, no. 4, pp. 83–124, 2011.
[10] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[11] M. W. Berry and J. Kogan, *Text mining: applications and theory*. John Wiley \& Sons, 2010.
[12] N. Herlinawati, Y. Yuliani, S. Faizah, W. Gata, and S. Samudi, "Analisis Sentimen Zoom Cloud Meetings di Play Store Menggunakan Na{\"\i}ve Bayes dan Support Vector Machine," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 2, pp. 293–298.
[13] L. L. Dhande and G. K. Patnaik, "Analyzing sentiment of movie review data using Naive Bayes neural classifier," *Int. J. Emerg. Trends \& Technol. Comput. Sci.*, vol. 3, no. 4, pp. 313–320, 2014.
[14] A. Wibowo, "10 FOLD-CROSS VALIDATION," 2017.
[15] S. Narkhede, "Understanding auc-roc curve," *Towar. Data Sci.*, vol. 26, pp. 220–227, 2018.