Jurnal Kajian dan Terapan Matematika
Volume 8, Edisi 3, Bulan November 2022, 199-209
http://journal.student.uny.ac.id/ojs/index.php/jktm

Klasifikasi informasi hoaks pada media sosial twitter menggunakan algoritma random forest berbasis particle swarm optimization

Classification of hoax information on social media twitter using random forest algorithm based on particle swarm optimization

Taqiyudin Muhammad Khalil, Prodi Matematika FMIPA UNY
Sahid *, Prodi Matematika FMIPA UNY
*e-mail: sahid@uny.ac.id

Abstrak
Twitter merupakan salah satu media sosial yang memiliki peran penting sebagai sarana penyebaran informasi dan berkomunikasi. Maraknya penyebaran informasi hoaks memberikan banyak dampak buruk kepada penggunanya. Untuk mengetahui apakah informasi tersebut hoaks atau bukan diperlukan klasifikasi. Penelitian ini melakukan klasifikasi informasi hoaks dengan menggunakan algoritma Random Forest (RF) berbasis Particle Swarm Optimization (PSO). Data penelitian diambil menggunakan Twitter Application Programming Interface (API) pada rentang waktu November 2021 sampai Februari 2022. Data sampel terdiri atas tiga kategori informasi, yaitu akurat (100 sampel), hoaks yang menyesatkan (50 sampel), dan hoaks yang salah (50 sampel). Data sampel diklasifikasikan setelah melalui tahap text mining, pra proses data, eksplorasi data, pembobotan kata, pembagian data training dan data testing dalam dua perbandingan, yaitu 70% data training dan 30% data testing, dan 80% data training dan 20% data testing. Algoritma yang menghasilkan klasifikasi terbaik adalah RF berbasis PSO dengan tingkat akurasi sebesar 73%, presisi sebesar 66%, dan recall sebesar 65%. Selanjutnya diperoleh beberapa karakteristik informasi hoaks pada data sampel, di antaranya: memiliki jumlah komentar, likes, dan retweet yang tinggi. Topik yang paling sering terpapar hoaks pada data sampel adalah politik (24 sampel), kriminal dan kesehatan (21 sampel), dan terkini (20 sampel). Kata-kata pada data sampel tweet yang paling sering terpapar hoaks adalah "sebar", "hati", dan "buzzer".
Kata kunci: Twitter, Informasi Hoaks, Klasifikasi, Random Forest, Particle Swarm Optimization.

Abstract
Twitter is one of the social media that has an important role as a means of disseminating information and communicating. The spread of hoax information has a lot of bad effects on its users. To find out whether the information is a hoax or not, a classification is needed. This study classified hoax information using the Random Forest (RF) algorithm based on Particle Swarm Optimization (PSO). The research data was taken using the Twitter Application Programming Interface (API) in the period November 2021 to February 2022. The sample data consisted of three categories of information, namely accurate (100 samples), misleading hoaxes (50 samples), and false hoaxes (50 samples). The sample data is classified after going through the stages of text mining, data preprocessing, data exploration, word weighting, distribution of training data and testing data in two comparisons, namely 70% training data and 30% testing data, and 80% training data and 20% testing data. . The algorithm that produces the best classification is RF-based PSO with an accuracy rate of 73%, precision of 66%, and recall of 65%. Furthermore, several characteristics of hoax information were obtained in the sample data, including: having a high number of comments, likes, and retweets. The topics most frequently exposed to hoaxes in the sample data are politics (24 samples), crime and health (21 samples), and recent (20 samples). The words in the sample tweet data that were most frequently exposed to hoaxes were "spread", "heart", and "buzzer".
***Keywords**: Twitter, Hoax Information, Classification, Random Forest, Particle Swarm Optimization.*

PENDAHULUAN

Pada era disrupsi teknologi digital, media sosial memiliki peran penting dalam berbagi informasi dan berkomunikasi. Salah satu media sosial yang digunakan sebagai sarana untuk saling bertukar informasi dan berkomunikasi adalah Twitter. Dalam penyebaran informasi, Twitter memiliki lebih dari 200 topik dengan 48 bahasa yang tersedia. Berdasarkan hasil survei pada situs Hootsuite (2022), Twitter menempati peringkat ke tujuh sebagai media sosial terfavorit di dunia dengan jumlah pengguna aktif harian sebanyak 429 juta. Dilansir dalam situs Statista, (2021b), Indonesia menempati peringkat ke enam dunia dengan jumlah pengguna Twitter aktif sebanyak 16.32 juta. Cara menyebarkan informasi pada media sosial Twitter yaitu dengan membuat *tweet* yang berisi tentang tulisan mengenai topik yang diinginkan.

Tweet semakin mudah diakses oleh pengguna dan cepat tersebar yang berujung viral dikarenakan kehadiran fitur *comment, share, retweet, like, hashtag, tag* dan *trending topic* pada media sosial Twitter. Kemudahan tersebut tidak hanya memberikan dampak positif, tetapi juga memberikan dampak negatif. Salah satu dampak negatif dari kemudahan akses tersebut adalah perkembangan kualitas informasi yang disebarluaskan oleh pengguna semakin berkurang. Informasi tersebut di dalamnya termasuk informasi yang mengandung unsur kepalsuan atau hoaks. Hoaks merupakan informasi yang dimanipulasi dengan sengaja dan bertujuan untuk memberikan pemahaman yang salah. Penyebaran informasi hoaks di Twitter pada kalangan masyarakat dapat menyebabkan efek negatif, seperti kepanikan, hilangnya kepercayaan, memburuknya hubungan sosial, kerusakan psikologis dan materiil, kerugian waktu dan ekonomi, dan sebagainya.

Maraknya penyebaran informasi hoaks yang telah terjadi dengan beragam bahasa dan bentuk konten yang tersebar di Twitter dan tidak adanya pengelompokkan informasi menyebabkan kesulitan untuk mengetahui keakuratan informasi tersebut. Teks pada informasi *tweet* di media sosial Twitter banyak mengandung kata-kata yang tidak relevan dan bersifat redudansi yang perlu diolah sehingga kata-kata tersebut menjadi suatu informasi yang berguna. Pengolahan teks pada *tweet* di media sosial Twitter dapat menggunakan *text mining*. *Text mining* merupakan proses mendapatkan informasi baru dengan kualitas yang tinggi berdasarkan berbagai sumber teks yang dilakukan oleh komputer (Rodiyansyah & Winarko, 2012). Informasi baru yang telah diolah menggunakan *text mining* dapat digunakan untuk berbagai analisis, salah satunya adalah klasifikasi.

Klasifikasi merupakan proses untuk mengatur dan memberi arti informasi yang berguna untuk menentukan atau menetapkan kesesuaian peristiwa, gagasan, barang, dan orang. Klasifikasi bertujuan untuk mengelompokkan suatu informasi berdasarkan persamaan dan ciri-ciri yang dimiliki ke dalam kelas-kelas yang sesuai. Secara umum klasifikasi harus melewati beberapa tahap, yaitu pengambilan data, persiapan set data, pra-proses data, pembagian set data, pelatihan data, penggunaan algoritma, evaluasi model, dan pengujian.

Penelitian mengenai klasifikasi informasi hoaks sebelumnya dilakukan oleh Nurhayati dan Pasaribu (2020) dengan menggunakan algoritma klasifikasi *Levenshtein Distance* (LD). Hasil penelitian ini memiliki tingkat akurasi sebesar 80% dari data sebanyak 40 berita. Penelitian tentang klasifikasi informasi hoaks berikutnya dilakukan oleh Prasetijo et al (2017). Klasifikasi dilakukan dengan menggunakan algoritma *Support Vector Machine* (SVM) dan *Stochastic Gradieent Descent* (SGD). Penelitian ini menghasilkan persentase dari tingkat akurasi sebesar 86% dari 200 informasi. Penelitian serupa juga dilakukan oleh Afriza dan Adisantoso (2018) menggunakan algoritma *Rocchio*. Akurasi hasil klasifikasi informasi hoaks dengan 600 informasi adalah 82.6%.

Algoritma yang digunakan pada penelitian ini adalah *Random Forest* (RF) berbasis *Particle Swarm Optimization* (PSO). RF merupakan salah satu algoritma yang menggunakan

metode *ensemble* pada klasifikasi dalam jumlah besar. Metode *ensemble* merupakan metode untuk meningkatkan akurasi klasifikasi dengan cara menggunakan beberapa algoritma dalam pencarian solusi terbaik (Jian et al., 2012: 377-379). RF dikenal mampu mencapai akurasi yang baik tanpa melakukan pencarian yang banyak pada parameter data *training* melalui penggabungan pohon (*tree*). Kelemahan algoritma ini adalah pembelajaran dapat berjalan secara lambat, tergantung pada parameter dan atribut yang digunakan. Algoritma ini juga tidak dapat memperbaiki model yang dihasilkan secara berulang.

Penggunaan seleksi fitur PSO dapat memperbaiki atribut pada set data dan mengambil keputusan berdasarkan parameter optimal dari pengklasifikasian yang telah dilakukan sebelumnya sehingga meningkatkan hasil akurasi (Amir et al., 2020). Penggunaan PSO dapat dikombinasikan dengan algoritma klasifikasi lain, seperti algoritma SVM, *K-Nearest Neighbor* (KNN), *Naïve Bayes* (NB), dan RF. Penggunaan PSO dapat meningkatkan nilai akurasi terhadap algoritma SVM sehingga menunjukkan nilai akurasi menjadi 99.6% (Amalia et al., 2017).

Atas dasar tersebut, penelitian ini akan menerapkan algoritma PSO untuk klasifikasi informasi hoaks di media sosial Twitter menggunakan algoritma RF berdasarkan tweet yang disebarluaskan oleh pengguna untuk memberikan tingkat akurasi dan pengambilan keputusan yang maksimal dalam pengklasifikasian.

METODE PENELITIAN

Deskripsi Jenis dan Data Penelitian

Jenis penelitian yang digunakan adalah penelitian campuran. Penelitian campuran didefinisikan oleh Johnson dan Onwuegbuzie (2007) sebagai pendekatan penelitian yang menggabungkan pendekatan kualitatif dan kuantitatif. Pendekatan ini menyertakan asumsi-asumsi filosofis, penggabungan kedua pendekatan kualitatif dan kuantitatif dalam satu penelitian, dan aplikasi pendekatan tersebut.

Data penelitian yang digunakan adalah data primer, yaitu pengambilan sampel data *tweet* di Twitter secara langsung dan dikumpulkan sebagai set data berjumlah 240 sampel tweet yang berisi 8 atribut yaitu topik informasi, asal informasi, jumlah komentar, jumlah likes, jumlah retweet, jumlah hashtag, jumlah tag, dan isi teks pada setiap sampel yang dikelompokkan menjadi tiga, yaitu informasi akurat sejumlah 100 sampel tweet, informasi hoaks yang salah sejumlah 50 sampel tweet, dan informasi hoaks yang menyesatkan sejumlah 50 sampel tweet, serta informasi acak sejumlah 40 sampel tweet untuk pengujian algoritma.

Tahapan Penelitian

1. Pengumpulan Data
   Pengumpulan data dilakukan secara langsung pada penyebaran *tweet* di Twitter dengan menggunakan Twitter API bahasa pemrograman Python.
2. Pelabelan Data
   Pelabelan data terdiri atas tiga label kelas, yaitu label informasi akurat, label informasi hoaks yang salah, dan label informasi hoaks yang menyesatkan.
3. *Text Mining*
   Teks pada sampel data *tweet* dilakukan tahap *text mining* untuk mendapatkan informasi baru yang berguna pada proses klasifikasi. Terdapat beberapa tahapan proses dalam *text mining*, yaitu *case folding*, *tokenizing*, *filtering* (*stopwords removal*), normalisasi, dan *stemming*.
4. Pra Proses Data
   Pra proses data merupakan proses untuk menghilangkan beberapa permasalahan pada data yang dapat mengganggu saat pemrosesan data. Terdapat beberapa tahapan proses dalam pra proses data, yaitu *encoding* data, dan *scaling* data.
5. Tahap Eksplorasi Data

Eksplorasi data digunakan untuk memahami komponen dan isi penyusun data tergantung pada konteks dan tipe model data untuk mendapatkan suatu informasi yang penting. Pada penelitian ini, pra proses data yang digunakan adalah *encoding* data, dan *scaling* data.

6. Pembobotan TF-IDF
   Pembobotan TF-IDF digunakan untuk mengevaluasi pentingnya kata-kata dalam teks. Metode ini menggabungkan dua konsep perhitungan bobot, yaitu TF atau frekuensi kemunculan kata dan IDF atau jumlah kalimat yang mengandung kata tersebut.
7. Pembagian Data
   Pembagian data dilakukan untuk menentukan data yang dijadikan data training dan data yang dijadikan data testing untuk proses pemodelan klasifikasi. Pembagian data dilakukan menjadi data *training* dan data *testing* dengan perbandingan 70:30 dan 80:20.
8. Pemodelan Klasifikasi *Random Forest*
   Data yang telah siap digunakan selanjutnya diproses menggunakan klasifikasi *Random Forest*.
9. Pemodelan Seleksi Fitur *Particle Swarm Optimization*
   Data yang digunakan pemodelan klasifikasi selanjutnya dilakukan tahap seleksi fitur. Tahap ini bertujuan untuk mengukur dan mengevaluasi tingkat relevansi dan redudansi pada setiap atribut yang digunakan pada klasifikasi sebelumnya.
10. Hasil Pengujian Pemodelan Klasifikasi
    Pemodelan yang telah dirancang selanjutnya dilakukan pengujian pada data yang dimiliki.
11. Evaluasi dan Analisis Hasil Pemodelan Klasifikasi.
    Evaluasi performansi dilakukan untuk menguji hasil dari model klasifikasi yang digunakan dengan mengukur nilai performansi dari model klasifikasi yang telah dibuat.

HASIL DAN PEMBAHASAN

Pengambilan sampel data pada informasi *tweet* di Twitter menggunakan Twitter *Application Programming Interface* (API) pada *library tweepy* bahasa pemrograman Python dan menggunakan fungsi *api.search_tweets* yang berisi 8 atribut yaitu topik informasi, asal informasi, jumlah komentar, jumlah *likes*, jumlah *retweet*, jumlah *hashtag*, jumlah *tag*, dan isi teks pada setiap sampel. Atribut topik informasi merupakan atribut kategorik yang terdiri atas politik, kriminal, terkini, kesehatan, olahraga, pertahanan dan keamanan, hiburan, ekonomi, dan lingkungan. Asal informasi merupakan atribut kategorik yang terdiri atas nama media *online*, dan pengguna. Jumlah komentar, jumlah *likes*, jumlah *retweet*, jumlah *hashtag*, jumlah *tag* merupakan jumlah pada setiap satu sampel tweet. Data yang diperoleh kemudian disimpan ke dalam format file csv

Setelah proses pengambilan data dan verifikasi data, kemudian dilakukan penyaringan data dan pelabelan data. Penyaringan data dilakukan untuk mengumpulkan data yang hanya berisi informasi berupa narasi teks. Data yang dianggap tidak relevan dan memiliki format selain narasi teks seperti video, foto, dan suara dihapus. Data tweet yang tidak relevan tersebut berisi antara lain *retweet* yang mengandung teks yang sama, *tweet* mengenai informasi iklan produk, balasan *tweet* dari akun resmi, dan *tweet* yang mempromosikan atau mengandung *endorsement* suatu layanan.

Proses penyaringan data menyisakan sebanyak 200 sampel *tweet*. Data tersebut selanjutnya dilakukan proses pelabelan data, yaitu pelabelan pada topik informasi, asal informasi, jumlah komentar, jumlah *likes*, jumlah *retweet*, jumlah *hashtag*, jumlah *tag*. Label informasi akurat sejumlah 100 sampel *tweet*, label informasi hoaks yang salah sejumlah 50 sampel *tweet*, dan label informasi hoaks yang menyesatkan sejumlah 50 sampel *tweet*.

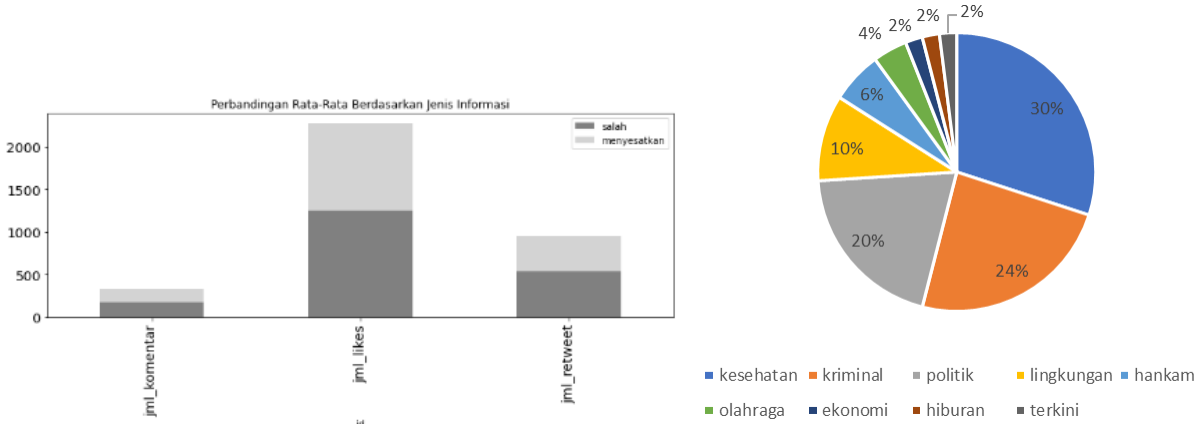
Teks pada set data selanjutnya dilakukan tahap *case folding* untuk mengubah semua karakter yang terdapat dalam dokumen teks menjadi huruf kecil, tahap *tokenizing* teks untuk pemecahan perkata serta menghilangkan tab, garis baru, *backslash*, emotikon, tautan, angka,

tanda baca, *whitespace*, dan satu karakter, tahapan *filtering* untuk mengambil kata-kata yang dianggap penting untuk proses selanjutnya, tahap normalisasi kata untuk menyeragamkan kata-kata dengan makna sama, namun menggunakan penulisan yang berbeda, dan tahapan *stemming* untuk menghilangkan imbuhan depan dan imbuhan belakang kembali ke bentuk dasarnya. Contoh hasil *text mining* disajikan pada Tabel 2.

Tabel 1. Contoh Hasil Text Mining

Sebelum	Sesudah
Selamat memperingati HARI KESAKTIAN …	[selamat, ingat, sakti, pancasila, …
Bukannya itu vlog Deddy tahun lalu. dan itu la...	[vlog, deddy, lab, kelar, dumdum, ...
Brigjen TNI Cahyo Ingatkan Bahaya Komunis Gaya...	[brigjen, tni, cahyo, ingat, bahaya, …
pinjol ilegal hajar bravo @DivHumas_Polri	[sayang, mui, haram, hadap, vaksin, …
Sangat di sayangkan MUI @MUIPusat TIDAK TEGAS ...	[gambargambar, muncul, media, …

Tahap ini dilakukan untuk menghilangkan beberapa permasalahan pada data yang dapat mengganggu saat pemrosesan data. Permasalahan tersebut di antaranya adalah data menggunakan format yang tidak konsisten. Pada penelitian ini, pra proses data yang digunakan adalah *encoding* data, dan *scaling* data. *Encoding* data merupakan proses mengubah objek non numerik pada atribut menjadi objek numerik sehingga dapat dilakukan pengolahan data secara matematis. *Scaling* data merupakan proses menyamakan rentang nilai pada atribut yang memiliki objek numerik sehingga tidak ada lagi satu atribut yang mendominasi atribut data lainnya seperti rentang nilai jumlah *tag* lebih sedikit dibandingkan dengan rentang nilai jumlah *likes*. Eksplorasi data yang digunakan pada penelitian ini adalah melihat *missing value*, statistik deskriptif, dan visualisasi dalam bentuk grafik dan *word cloud*.



Gambar 1. Perbandingan Rata-Rata Berdasarkan Jenis dan Topik Informasi

Word Cloud Informasi Hoaks Menyesatkan

Word Cloud Informasi Hoaks yang Salah

Gambar 2. Hasil Output Tampilan Word cloud

Berdasarkan Gambar 2 dan Gambar 3, karakteristik informasi hoaks pada data penelitian ini adalah memiliki jumlah komentar, jumlah *likes*, dan jumlah *retweet* yang tinggi. Topik yang paling sering terpapar hoaks adalah politik dengan 24 sampel, kriminal dan kesehatan dengan 21 sampel, dan terkini dengan 20 sampel; kata-kata pada *tweet* yang terpapar hoaks paling banyak menggunakan kata "sebar", "vaksin", "hati", dan "buzzer".

	kata	rank
92	indonesia	7.088335
101	jakarta	4.888581
60	dunia	4.588158
227	sebar	4.515029
111	jokowi	4.500245
...	...	...
238	situs	0.585099
74	hama	0.585099
241	spt	0.585099
46	desa	0.559062
231	siang	0.526866

Gambar 3. Frekuensi Kemunculan Kata Berdasarkan Peringkat 5 Besar Teratas dan Terbawah

Berdasarkan Gambar 4, kata "Jakarta" dan "Jakarta" merupakan kata yang memiliki frekuensi tertinggi, dan kata "desa" dan "siang" merupakan kata yang memiliki frekuensi terendah.

```
Sampel TFIDF ke-13

['orang', 'setrum', 'listrik', 'kisar', 'detik', 'sbg', 'tolong', 'telapak', 'kaki', 'darah', 'alir', 'tubuh', 'silah', 'seba
r', 'kepada', 'sahabat', 'kerabat']

                         TF            IDF           TF-IDF        Term

array position 34        0.166667      5.199705      0.866618      darah
array position 42        0.166667      5.605170      0.934195      detik
array position 115       0.166667      5.199705      0.866618      kepada
array position 146       0.166667      5.199705      0.866618      listrik
array position 176       0.166667      4.688879      0.781480      orang
array position 225       0.166667      3.813411      0.635568      sebar
```

Gambar 4. Sampel Hasil Pembobotan TF-IDF

Berdasarkan Gambar 5, kata "sebar" merupakan kata yang memiliki frekuensi tertinggi pada dokumen sehingga nilai TF-IDF yang dihasilkan rendah dibandingkan dengan kata "detik", "darah", dan "listrik" yang merupakan kata dengan frekuensi terendah pada dokumen sehingga nilai TF-IDF yang dihasilkan lebih tinggi.

1. Pembagian Data
   Pada penelitian ini akan digunakan dua tipe data, yaitu tipe I dan tipe II. Pada data tipe I data *training* dan data *testing* menggunakan perbandingan 70:30, sedangkan untuk data tipe II menggunakan perbandingan 80:20. Jumlah data *training* dan data *testing* yang digunakan untuk setiap data disajikan pada Tabel 3.

Tabel 2. Perbandingan Data Training dan Data Testing

Tipe Data	Perbandingan Data *Training* dan Data *Testing*	Data *Tweet*	
		Data *Training*	Data *Testing*
Tipe I	70:30	140	60
Tipe II	80:20	160	40

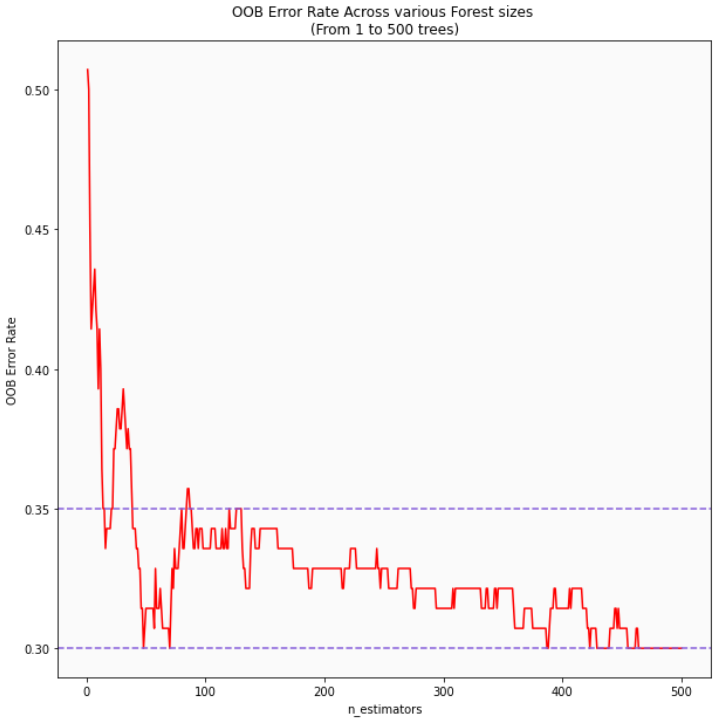
2. Klasifikasi *Random Forest*
   Tahapan-tahapan dalam melakukan klasifikasi menggunakan RF di antaranya adalah memilih jumlah pohon yang digunakan (n_estimators) dan jumlah atribut acak yang memisahkan setiap sampel (max_features) terbaik yang kemudian dicari rataan tingkat misklasifikasi, dan mencari atribut penting pada keputusan klasifikasi, serta menentukan tingkat akurasi, presisi, dan *recall* pada hasil prediksi klasifikasi dengan data *training* dan data *testing*.

Pemilihan Parameter Klasifikasi

Dalam proses klasifikasi menggunakan RF, hal yang perlu diperhatikan pada tahap awal adalah menentukan parameter yang digunakan seperti banyaknya atribut bebas yang digunakan dalam klasifikasi pada satu pohon, banyaknya pohon yang digunakan untuk klasifikasi, dan kriteria klasifikasi yang digunakan. Pengambilan atribut independen ini dilakukan secara acak. Jumlah atribut tersebut kemudian akan dilakukan percobaan sebanyak n pohon. Proses pemilihan jumlah atribut dan jumlah pohon terbaik untuk proses klasifikasi dapat menggunakan fungsi *GridSearchCV()* dari *library sklearn* untuk mendefinisikan parameter terbaik berdasarkan data yang digunakan. Berdasarkan pencarian tersebut diketahui bahwa model klasifikasi terbaik untuk penelitian ini menggunakan *entropy* sebagai kriteria penilaian, jumlah cabang pada pohon (max_depth) adalah 4, dan jumlah atribut yang digunakan (max_features) adalah $log2$ atau sebanyak 3 atribut.

Rataan Tingkat Misklasifikasi

Rataan tingkat misklasifikasi digunakan untuk menentukan berapa banyak pohon yang sesuai dengan pemodelan klasifikasi pada data yang digunakan. RF dapat menemukan tingkat misklasifikasi dengan menggunakan konsep *out-of-bag* (OOB). OOB menggunakan data yang tidak terlihat pada pembagian data dengan cara membagi matriks akurasi menggunakan metode validasi silang. Hasil *output* proses menemukan tingkat misklasifikasi disajikan pada Gambar 6.



Gambar 5. Rataan Tingkat Misklasifikasi

Berdasarkan *output* visualisasi rataan tingkat misklasifikasidiketahui bahwa persebaran nilai misklasifikasi memiliki pola di tingkat *error* 0.3 sampai 0.35. Pada penelitian ini diambil jumlah pohon sebanyak 500 dengan nilai rataan tingkat misklasifikasi sebesar 0.3.

Tingkat Kepentingan Atribut

RF memiliki dua tingkat kepentingan yang diberikan untuk setiap atribut di hutan acak. Tingkat pertama didasarkan pada seberapa besar akurasi menurun ketika atribut dikecualikan dan tingkat kedua didasarkan pada penurunan nilai kriteria ketika atribut yang dipilih sebagai pembagi sebuah cabang pohon. Setiap sampel pada pohon terdapat sampel yang tidak digunakan selama konstruksi model. Sampel ini digunakan untuk menghitung tingkat kepentingan atribut tertentu berdasarkan rataan tingkat misklasifikasi. Tingkat kepentingan atribut pada penelitian ini disajikan pada Gambar 7.

```
Feature ranking:
1. The feature 'asal_informasi' has a Mean Decrease in Impurity of 0.42810
2. The feature 'teks' has a Mean Decrease in Impurity of 0.12497
3. The feature 'jml_likes' has a Mean Decrease in Impurity of 0.11539
4. The feature 'jml_komentar' has a Mean Decrease in Impurity of 0.10256
5. The feature 'jml_retweet' has a Mean Decrease in Impurity of 0.10168
6. The feature 'topik_informasi' has a Mean Decrease in Impurity of 0.06022
7. The feature 'jml_hashtag' has a Mean Decrease in Impurity of 0.04452
8. The feature 'jml_tag' has a Mean Decrease in Impurity of 0.02257
```

Gambar 6. Tingkat Kepentingan Atribut

3. Seleksi Fitur *Particle Swarm Optimization*

   Tahap ini bertujuan untuk mengukur dan mengevaluasi tingkat relevansi dan redudansi pada setiap atribut yang digunakan pada klasifikasi sebelumnya. Tahapan-tahapan tersebut adalah sebagai berikut.

Inisiasi Seleksi Fitur *Particle Swarm Optimization*

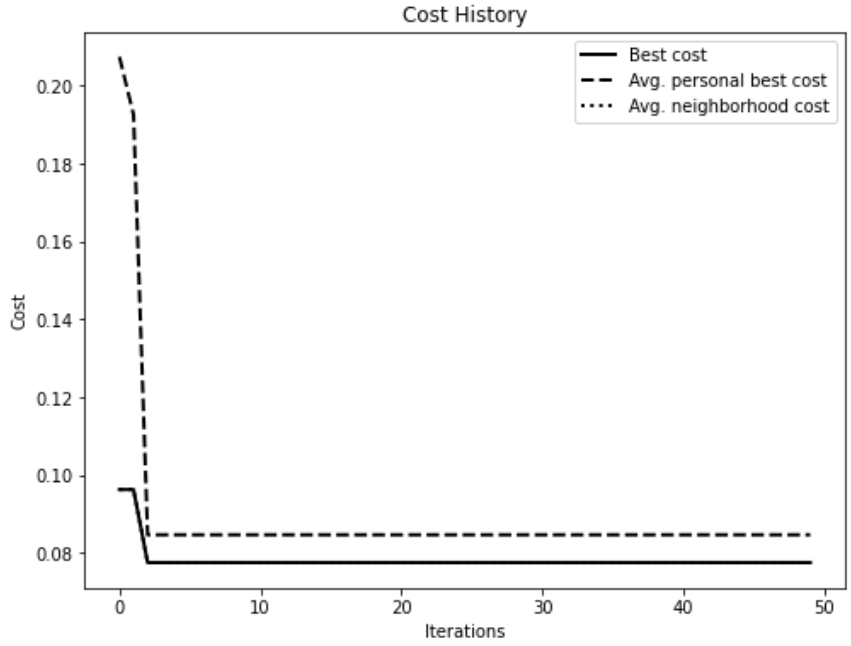
Seleksi fitur menggunakan algoritma *Particle Swarm Optimization* (PSO) digunakan untuk memilih atribut terbaik dari data dan mengurangi atribut yang tidak relevan untuk meningkatkan efektifitas dan efisiensi kinerja dari algoritma klasifikasi. PSO mengoptimasi

permasalahan klasifikasi dengan cara menggerakkan partikel berdasarkan posisi dan kecepatan setiap partikel (*swarm*). Partikel-partikel pada algoritma PSO merupakan cikal bakal untuk melakukan proses seleksi fitur, dimana pada partikel tersebut akan dilakukan inisialisasi sebanyak 8 atribut yang diproses. Pergerakan partikel dipengaruhi oleh nilai *fitness* terbaik pada *swarm* yang dilakukan pada setiap iterasi hingga mencapai iterasi maksimum. Iterasi maksimum merupakan pengujian untuk mengetahui nilai *fitness* terbaik pada iterasi tersebut. Iterasi yang akan diuji sebanyak 50 iterasi untuk mencapai iterasi maksimum. Pada penelitian ini parameter-parameter yang diujikan disajikan pada Tabel 4.

Tabel 3. Nilai Parameter Algoritma PSO

Parameter	Nilai
Ukuran populasi	200
Koefisien akselerasi 1 (c_1)	2
Koefisien akselerasi 2 (c_2)	2
Bobot Inersia (w)	0.9

Hasil pengujian iterasi maksimum didapatkan nilai *fitness function* yang berbeda. Nilai *fitness* awalnya adalah 0.40625 yang kemudian setelah dilakukan pengujian sebanyak 50 kali hingga mencapai iterasi maksimum didapatkan nilai *fitness* terbaik (*Pbest*) adalah 0.2555 dengan nilai *Gbest* adalah 0.2087. Perbandingan hasil pengujian nilai *fitness* dan nilai *Gbest* pada iterasi maksimum disajikan pada Gambar 4.33.



Gambar 7. Hasil Pengujian Iterasi Maksimum

Seleksi Fitur *Particle Swarm Optimization*

Setelah kondisi berhenti telah terpenuhi dan iterasi sudah mencapai iterasi maksimum, proses seleksi fitur dihentikan dan posisi partikel yang lama digantikan dengan posisi partikel yang baru berdasarkan nilai *fitness* terbaik yang didapatkan. Atribut yang didapatkan pada proses seleksi fitur PSO selanjutnya menggantikan atribut sebelumnya untuk digunakan pada proses pengujian klasifikasi.

4. Hasil Pengujian Pemodelan Klasifikasi

Pengujian hasil klasifikasi pada penelitian ini dibagi menjadi dua, yaitu pengujian hasil klasifikasi *Random Forest* tanpa *Particle Swarm Optimization* dan pengujian hasil klasifikasi *Random Forest* berbasis *Particle Swarm Optimization.*

5. Evaluasi dan Analisis Hasil Pemodelan Klasifikasi
   Evaluasi performansi dilakukan untuk menguji hasil dari model klasifikasi yang digunakan dengan mengukur nilai performansi dari model klasifikasi yang telah dibuat. Parameter pengujian yang digunakan untuk evaluasi yaitu akurasi, presisi, dan *recall*. Perhitungan nilai akurasi, presisi, dan *recall* dapat diperoleh dengan cara membandingkan kelas aktual dan kelas prediksi dari data *training* dan data *testing* menggunakan *confusion matrix*. Hasil perhitungan nilai akurasi, presisi, dan *recall* model klasifikasi *Random Forest* dan *Random Forest* berbasis *Particle Swarm Optimization* masing-masing tipe data disajikan pada Tabel 5.

Tabel 4. Evaluasi Hasil Pemodelan Klasifikasi

Tipe Data	Uji Data	Pemodelan	Hasil		
			Akurasi	Presisi	*Recall*
Tipe I	Data *Training*	RF	84%	69%	72%
		RF-PSO	81%	72%	73%
	Data *Testing*	RF	72%	63%	63%
		RF-PSO	73%	66%	66%
Tipe II	Data *Training*	RF	85%	67%	65%
		RF-PSO	83%	73%	72%
	Data *Testing*	RF	65%	58%	57%
		RF-PSO	73%	66%	65%

Berdasarkan Tabel 3 dapat diketahui bahwa hasil model klasifikasi terbaik yaitu pada data tipe II menggunakan *Random Forest* berbasis *Particle Swarm Optimization* dengan perbandingan data *training* dan data *testing* yaitu 80:20 dengan total 160 data *training* dan 40 data *testing*. Tingkat akurasi sebesar 73%, presisi sebesar 66% dan *recall* sebesar 65% dengan total kesalahan hasil prediksi sebanyak 11 data.

6. Pengujian Sampel dengan Label Acak

Pengujian pada sampel dengan label acak digunakan untuk mengetahui bagaimana performa model dalam memprediksi hasil klasifikasi pada sampel yang memiliki kelas acak. Sampel yang digunakan berjumlah 40 sampel. Hasil prediksi dengan menggunakan algoritma *Random Forest* berbasis *Particle Swarm Optimization* untuk sampel dengan label acak didapatkan hasil prediksi informasi hoaks yang menyesatkan sebanyak 14 sampel, prediksi informasi hoaks yang salah sebanyak 6 sampel, dan prediksi informasi akurat sebanyak 20 sampel.

SIMPULAN

Berdasarkan hasil pembahasan yang telah dipaparkan sebelumnya, penerapan klasifikasi *Random Forest* (RF) berbasis *Particle Swarm Optimization* (PSO) untuk menentukan informasi akurat dan hoaks pada media sosial Twitter dengan menggunakan bahasa pemrograman Python dapat disimpulkan sebagai berikut.

1. Hasil tingkat keakuratan klasifikasi algoritma RF berbasis PSO pada data penelitian ini memiliki peningkatan dibandingkan dengan RF tanpa PSO sebanyak 1% pada tipe data I dengan perbandingan data *training* dan data *testing* adalah 70:30 dari 72% menjadi 73% dan 8% pada tipe data II dengan perbandingan data *training* dan data *testing* adalah 80:20 dari 65% menjadi 73%.

2. Algortima dengan hasil klasifikasi terbaik pada data penelitian ini adalah RF berbasis PSO pada data tipe II dengan perbandingan data *training* dan data *testing* adalah 80:20 yang berjumlah 160 data *training* dan 40 data *testing*. Diperoleh tingkat akurasi sebesar 73%, presisi sebesar 66%, *recall* sebesar 65%, dan total kesalahan hasil prediksi sebanyak 11 sampel data.
3. Karakteristik informasi hoaks pada data penelitian ini adalah memiliki jumlah komentar, jumlah *likes*, dan jumlah *retweet* yang tinggi. Topik yang paling sering terpapar hoaks adalah politik dengan 24 sampel, kriminal dan kesehatan dengan 21 sampel, dan terkini dengan 20 sampel. Kata-kata pada *tweet* yang terpapar hoaks paling banyak menggunakan kata "sebar", "vaksin", "hati", dan "buzzer".

DAFTAR PUSTAKA

Afriza, A., & Adisantoso, J. (2018). Metode Klasifikasi Rocchio untuk Analisis Hoax. *Jurnal Ilmu Komputer dan Agri-Informatika*, *5*(1), 1–10. https://doi.org/10.29244/jika.5.1.1-10

Amalia, Hilda; Lestari, A. F. P., & Ari. (2017). Penerapan Metode SVM Berbasis PSO Untuk Penentuan Kebangkrutan Perusahaan. *Techno Nusa Mandiri*, *14*, 131–136.

Amir, R. F., Sobari, I. A., & Rousyati, R. (2020). Penerapan PSO Over Sampling Dan Adaboost Random Forest Untuk Memprediksi Cacat Software. *Indonesian Journal on Software Engineering (IJSE)*, *6*(2), 230–239. https://doi.org/10.31294/ijse.v6i2.9258

Hootsuite. (2022). Favourite Social Media Platforms in 2022 [Set data]. Diambil dari https://datareportal.com/social-media-users

Jian, Jiawei, P., Micheline, H., & Kamber. (2012). *Data Mining : Concepts and Techniques* (Third Edit). Waltham: Elsevier B.V.

Nurhayati, N., & Pasaribu, A. (2020). Perancangan Sistem Pendeteksi Berita Hoax Menggunakan Algoritma Levenshtein Distance Berbasis Php. *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, *19*(2), 74–84. https://doi.org/10.53513/jis.v19i2.2601

Prasetijo, A. B., Isnanto, R. R., Eridani, D., Soetrisno, Y. A. A., Arfan, M., & Sofwan, A. (2017). Hoax Detection System on Indonesian News Sites Based on Text Classification Using SVM and SGD. *Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017*, *2018-Janua*, 45–49. https://doi.org/10.1109/ICITACEE.2017.8257673

Rodiyansyah, S. F., & Winarko, E. (2012). Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, *6*(1), 91–100. https://doi.org/10.1163/ej.9789004182127.i-302.6

Statista. (2021). Social Media in Indonesia [Set data]. Diambil dari https://www.statista.com/topics/8306/social-media-in-indonesia