JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 7
No. 2
Agustus
2021

Perbandingan Klasifikasi dengan Pendekatan Pembelajaran Mesin untuk Mengidentifikasi *Tweet* Hoaks di Media Sosial *Twitter*

Shanto Moyrano Tambunan[#1], Yessica Nataliani[#2], Elizabeth Sri Lestari[#3]

[#]*Departemen Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana*
Jl. Diponegoro No. 52-60, Salatiga 50711
[1]682017148@student.uksw.edu
[2]yessica.nataliani@uksw.edu
[3]elizabeth@uksw.edu

Abstrak— **Perkembangan teknologi tidak luput dari dampak negatif, salah satunya hoaks. Twitter menjadi salah satu media sosial yang paling aktif digunakan sebagai pertukaran informasi, komunikasi, dan hiburan. Oleh karena itu pengguna Twitter dapat menyebarkan berita atau hoaks dengan mudah. Penelitian ini bertujuan mengidentifikasi *tweet* yang berisi informasi hoaks maupun valid menggunakan pembelajaran mesin. Algoritma yang digunakan adalah *Stochastic Gradient Descent, Naïve Bayes, Random Forest,* dan *Rocchio*. Keempat algoritma tersebut dibandingkan untuk kemudian dicari hasil terbaik dalam mengidentifikasi dan memverifikasi *tweet* di Twitter yang berisi hoaks atau informasi valid secara otomatis. Kata kunci yang digunakan adalah Corona, Mutasi Corona, PSBB, Dana Bansos, Dana Otsus, Utang Pemerintah, dan Sekolah Tatap Muka sebanyak 898 *tweet*. Data dikelompokkan berdasarkan kelas hoaks dan valid lalu diolah menjadi *dataset* dengan melewati tahap pra-proses hingga pembobotan kata dengan TF-IDF. Hasil pengujian menunjukkan algoritma *Stochastic Gradient Descent* merupakan algoritma terbaik dengan hasil akurasi rata-rata sebesar 84.92%. Pengujian lanjutan dilakukan dengan menghitung nilai presisi, *recall*, dan F-1. Hasil presisi terbaik sebesar 82.95% pada algoritma *Naïve Bayes*, sedangkan hasil *recall* dan F-1 terbaik didapat dari algoritma *Stochastic Gradient Descent* sebesar 85.05% dan 82.42%.**

Kata kunci*— Klasifikasi, Hoaks, *Tweet*, Pembelajaran Mesin, *Random Forest, Naïve Bayes, Stochastic Gradient Descent, Rocchio

I. PENDAHULUAN

Kemajuan teknologi tidak bisa dihindari dalam kehidupan manusia. Teknologi akan terus berkembang dan berinovasi untuk memberikan banyak manfaat bagi masyarakat. Namun efek dari kemajuan teknologi tidak akan terlepas dari dampak negatif. Seiring bertambah banyaknya informasi berita *online* yang tersebar secara luas, kualitas berita yang tersebar pun berkurang.

Informasi merupakan hasil pengolahan data menjadi sesuatu yang berguna bagi penerimanya. Informasi berfungsi untuk menambah wawasan karena dapat menggambarkan suatu permasalahan sehingga penerima dapat mengambil keputusan lebih cepat. Informasi juga dapat digunakan sebagai sarana membantu orang lain. Dalam penyampaian informasi, inti dari informasi sangatlah penting antara lain waktu, ruang atau tempat, serta hubungan keterkaitan antar situasi sehingga tidak menimbulkan kesalahtafsiran. Sumber informasi berasal dari data yang sesuai dengan kenyataan atau berdasar kejadian. Seperti halnya berita yang disebarkan lewat media sosial [1].

Teknologi informasi merupakan sebuah pengolahan dan pendistribusian informasi dalam bentuk digital ataupun elektronik. Teknologi informasi secara sederhana sebagai ilmu berbasis komputer di bidang informasi dan perkembangannya sangat pesat. Adanya perkembangan teknologi tersebut, membuat segala aktivitas di dunia nyata dapat dilakukan melalui dunia maya. Akses untuk internet menjadi lebih mudah dan efisien, mulai dari menjual produk, pesan singkat, hingga mencari informasi.

Menurut survei yang dilakukan oleh Assosiasi Penyelenggara Jasa Internet Indonesia [2] dalam periode 2019/2020-kuartal II, terdapat 196,7 juta pengguna internet, dalam arti sama dengan 73.7% penduduk Indonesia. Bahkan dari lima negara, Indonesia menjadi salah satu negara yang memiliki pengguna internet terbanyak. Banyaknya aktivitas terkait berita di media sosial dapat menjadi masalah akan validnya informasi.

Kabar bohong atau hoaks masih menjadi permasalahan besar di Indonesia, yang dapat menimbulkan kerugian baik bagi seseorang, kelompok, organisasi, ataupun pemerintah. Twitter menjadi salah satu sosial media yang paling sering digunakan, dimana tiap kicauannya (*posting*-an) berformat 140 kata. Hal tersebut membuat informasi menjadi mudah disebarkan dan informasi terbaru yang viral dapat dilihat melalui menu *trending*-nya.

Submitted 13-06-2021; Revised 08-07-2021; Accepted 12-07-2021

Hoaks merupakan tindakan memanipulasi data atau informasi yang dapat mempengaruhi opini pembaca, sehingga data atau informasi yang terindikasi hoaks tidak sesuai dengan fakta yang ada. Meskipun pemerintah Indonesia sudah melakukan banyak tindakan pencegahan hoaks, tapi masih banyak berita bohong yang tersebar di media social, yang dapat berdampak buruk dan merugikan banyak pihak. Survei yang dirilis oleh Masyarakat Telematika Indonesia (Mastel) tentang informasi palsu bahwa media sosial menjadi saluran penyebaran paling tinggi dengan persentase 87.5% [3].

Informasi hoaks di media sosial kebanyakan topik yang sedang viral. Beberapa topik yang sedang viral di Twitter beberapa bulan terakhir yang diangkat pada penelitian ini antara lain kesehatan, pendidikan, dan ekonomi. Oleh karena itu untuk menghindari dampak buruk tersebut, penelitian ini akan membantu mengetahui seberapa besar pembelajaran mesin (*machine learning*) mengidentifikasi kemungkinan hoaks di media sosial Twitter.

Penelitian yang dikembangkan adalah untuk melakukan perbandingan klasifikasi dengan pendekatan pembelajaran mesin untuk mengidentifikasi *tweet* hoaks di media sosial Twitter. Data dikumpulkan dengan metode *Web Crawling* menggunakan bahasa pemrograman Python dari tanggal 1 Maret 2021 hingga 15 April 2021. Algoritma yang digunakan adalah *Stochastic Gradient Descent* (SGD), *Naïve Bayes*, *Random Forest* dan *Rocchio*. Keempat algoritma tersebut diambil karena memiliki kemampuan *supervised learning*, yakni data dikumpulkan lalu dilakukan *pre-processing*, dan nantinya akan dilatih dengan pembelajaran mesin untuk dapat memprediksi dan mengklasifikasi suatu berita merupakan hoaks atau bukan.

Stochastic Gradient Descent dipilih menjadi metode yang dibandingkan karena metode ini merupakan salah satu metode klasifikasi yang memiliki performa yang baik untuk data yang berdimensi besar. SGD juga mempunyai akurasi yang tinggi, sehingga menjadikannya lebih presisi [4]. Sementara itu, algoritma *Naïve Bayes* juga dibandingkan karena metode tersebut merupakan metode klasifikasi yang cukup sederhana, yang menerapkan teorema Bayes, sehingga semua fitur dianggap tidak saling berhubungan. Dengan kata lain, jika ada fitur tertentu dari kelas, fitur tersebut tidak memiliki hubungan dengan fitur yang lain [5].

Random Forest menjadi pilihan lain perbandingan karena akurasinya dapat ditingkatkan dengan metode pemilihan data secara acak, yang dapat memanggil anak simpul dari setiap *node* (simpul atasnya). Hasil akumulasi dari setiap simpul pohon dipilih yang paling banyak muncul [6]. Semakin banyak pohon akan mempengaruhi tingkat akurasi, sehingga metode ini sangat baik dalam menangani data yang banyak dan dapat menyeimbangkan kesalahan (*error*). Algoritma *Rocchio* dipilih karena memiliki kinerja yang lebih baik dari *Naïve Bayes* pada penelitian sebelumnya, yaitu dengan melakukan perhitungan batas kelas menggunakan *centroid* dalam menetapkan hasil batas-batasnya [7].

II. LANDASAN TEORI

Klasifikasi berita atau informasi mengenai hoaks telah dilakukan pada beberapa penelitian. Salah satunya tentang pengembangan penyaringan hoaks berbahasa Indonesia, yang berdasar pada gambaran vektor dalam *Term Frequency* serta *Document Frequency* dengan teknik klasifikasinya. Penelitian tersebut menggunakan algoritma *Support Vector Machine* (SVM) serta SGD. Pada klasifikasi SGD yang menggunakan Huber, yang telah dimodifikasi, diperoleh akurasi tertinggi dengan persentase modifikasi 86%, berasal dari 100 data hoaks dan 100 situs web bukan hoaks, dengan pemilihan acak sebagai proses pelatihan di luar *dataset* [4].

Pada penelitian analisis sentimen data Twitter Bahasa Indonesia yang ditulis oleh Pantouw (2017) untuk mengetahui informasi dari data *tweet*, diklasifikasikan menjadi sentimen positif, negatif, serta netral. Penelitian tersebut dilakukan dengan metode klasifikasi *Multinomial Naïve Bayes* dan *Rocchio* dengan akurasi tertinggi dipereoleh *Rocchio* sebesar 96.28%. Nilai tersebut didapat setelah melakukan pelabelan ulang dan pemotongan yang akhirnya mengalami peningkatan [5].

Selain itu penelitian tentang eksperimen pada sistem klasifikasi berita hoaks berbahasa Indonesia berbasis pembelajaran mesin membahas tentang klasifikasi berita hoaks berbahasa Indonesia menggunakan pembelajaran mesin. Dengan menggunakan algoritma *Naïve Bayes*, *Support Vector Machine* (SVM) dan C4.5. Dari penelitian tersebut disimpulkan bahwa dalam klasifikasi teks, algoritma C4.5 dan SVM memiliki keakuratan yang cukup tinggi tanpa seleksi fitur. Namun pada hasil seleksi fitur, algoritma *Naïve Bayes* menjadi salah satu yang paling unggul [8].

Dalam penelitian lain dikembangkan analisis dan deteksi konten berita hoaks di Indonesia berdasarkan pembelajaran mesin. Penelitian ini menggunakan lima teknik klasifikasi. Akurasi tertinggi diperoleh dari pengklasifikasi *Random Forest* sebesar 76.46% setelah menggunakan presisi, *recall*, F-1, dan akurasi. Data yang diambil sebanyak 251 artikel berita online yang terdiri dari 151 artikel bukan hoaks dan 100 artikel hoaks [9].

Penelitian lain tentang identifikasi berita hoaks di media sosial Twitter dijelaskan tentang bagaimana cara mengidentifikasi berita hoaks melalui *tweet*. Identifikasi dilakukan menggunakan klasifikasi *Decision Tree* C4.5 dengan perbandingan seleksi fitur dari pembobotan, *Term Frequency-Inverse Document Frequency* (TF-IDF), serta *n*-gram. Dari penelitian tersebut didapatkan hasil bahwa seleksi fitur TF-IDF memiliki dampak yang signifikan, sehingga memungkinkan sistem untuk tidak menggunakan bobot data yang terlalu banyak pada penggunaan fitur yang banyak muncul data, sehingga menjadikannya lebih akurat dalam menentukan topik yang dibahas [10].

A. Berita

Informasi yang dapat menarik pembaca yang dapat disampaikan melalui beberapa media disebut dengan berita. Media yang dapat digunakan untuk menyebarkan

berita antara lain koran, televisi, internet, hingga jejaring sosial [11].

B. Hoaks

Berita yang dimanipulasi secara sengaja merupakan sebuah hoaks. Tujuan dari hoaks itu sendiri biasanya untuk memberikan berita yang salah dan bersifat menyesatkan [11]. Selain itu, hoaks juga bisa digunakan menjadi bahan untuk lelucon hingga menyebarkan ujaran kebencian ke satu pihak atau kelompok tertentu. Hal tersebut dapat mempengaruhi dan memberi dampak bagi penerimanya [12].

C. Media Sosial

Media sosial adalah media yang dapat dipakai oleh semua masyarakat dalam menunjukan ekspresi atau opini secara publik. Opini yang dipublikasi dapat berupa evaluasi (*review*), sentimen, maupun ungkapan emosi. Twitter adalah salah satu contoh media sosial yang banyak digunakan oleh masyarakat [13].

D. Twitter

Twitter ialah salah satu media sosial yang memungkinkan para penggunanya untuk membagikan informasi dengan batas maksimal 140 karakter [14].

E. Algoritma Pembelajaran Mesin

Algoritma yang digunakan dalam penelitian ini dijelaskan sebagai berikut:

1) Stochastic Gradient Descent (SGD): merupakan sebuah pendekatan sederhana dan efisien dalam melakukan klasifikasi secara linier menggunakan pembelajaran diskriminatif. Metode ini berupa algoritma optimasi iteratif (ulang) yang berguna untuk mencari titik fungsi minimum yang dapat diturunkan. Pada awal proses algoritma dimulai dengan melakukan penebakan. Kesalahan penebakan diperbaiki selama terjadi pengulangan tebakan menggunakan aturan gradien (turunan) dari fungsi yang akan diminimalkan. SGD memiliki kemampuan belajar lebih cepat dalam melakukan pelatihan klasifikasi. Selain itu, berdasarkan ukuran *dataset* latih tidak terbatas waktu pelaksanaannya [15].

2) Naïve Bayes: Klasifikasi *Naïve Bayes* mempunyai probabilitas sederhana dengan penerapan teori Bayes. Thomas Bayes adalah ilmuan Inggris yang menemukan metode ini yang digunakan untuk memprediksi peluang suatu peristiwa berdasarkan peristiwa yang pernah terjadi. Keuntungan pada klasifikasi ini yakni dibutuhkan hanyalah sejumlah data pelatihan yang kecil dalam mengestimasi skala yang dibutuhkan. *Naïve Bayes* dalam kebanyakan kondisi bekerja sangat baik pada implementasi di dunia nyata yang kompleks dibandingkan dengan yang diharapkan [9].

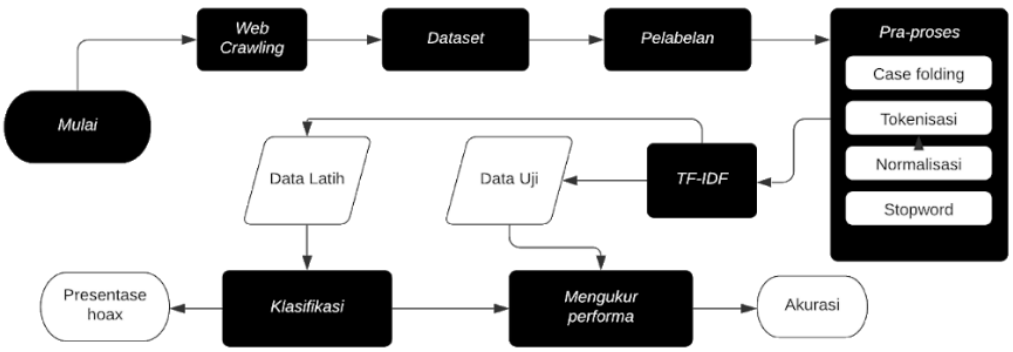
3) Random Forest (RF): dilakukan dengan membangun banyak pohon klasifikasi data secara acak dan terdistribusi sama rata. Dari setiap pohon dipilih nilai yang paling sering muncul dalam kelasnya. Kesalahan generalisasi dalam penggolongan pohon tergantung pada keakuratan setiap pohon di RF serta korelasinya antar mereka [6].

Akurasi dapat ditingkatkan oleh RF karena terdapat pemilihan acak dalam menghidupkan anak simpul untuk setiap node (simpul di atasnya) serta hasil klasifikasi diakumulasikan dari setiap pohon. Setelah itu, hasil yang paling banyak muncul akan dipilih. Tingkat akurasi klasifikasi dipengaruhi oleh jumlah pohon yang dibentuk. Semakin banyaknya pohon, hasil yang didapat semakin akurat. RF juga dapat menangani input variabel yang besar serta menyeimbangkan *unbalanced dataset* yang mengalami error [9].

4) Rocchio: Teknik Rocchio berperan untuk mendapatkan batas antar kelas. Teknik yang digunakan adalah dengan menerapkan batasan dalam bentuk *centroid*. *Centroid* itu sendiri merupakan sebuah rata-rata dari semua vektor [16].

III. METODE PENELITIAN

Tahapan yang dilakukan dalam penelitian ini diilustrasikan secara sederhana pada Gambar. 1 diagram alir penelitian.



Gambar. 1 Diagram alir penelitian

A. *Web Crawling*

Pengambilan data dilakukan dengan melakukan pemindaian (*scanning*) pada sebuah halaman web atau situs. Mencari atau merayapi data berupa informasi dari sebuah halaman adalah peran dari *crawler web* [17]. *Tweet* dari Twitter dipindai dengan bantuan akses Twitter *developer* dan dengan bahasa pemrograman Python. Gambar. 2 memperlihatkan hasil *web crawling* yang disimpan dalam bentuk Excel untuk nantinya diolah menjadi *dataset*.

B. *Dataset*

Tahap awal penelitian dilakukan dengan mengumpulkan data. Data tersebut didapat dari *tweet* yang mengandung informasi atau berita melalui Twitter dengan data yang didapatkan sebanyak 898 *tweet*. Topik *tweet* yang dikumpulkan mulai dari kesehatan, pemerintahan, hingga pendidikan. Dalam pencarian *tweet* yang mengandung topik tersebut digunakan kata kunci antara lain Corona, Mutasi Corona, PSBB, Dana Bansos, Dana Otsus, Utang Pemerintah, dan Sekolah Tatap Muka. *Tweet* yang telah didapat kemudian dicek kebenarannya secara manual. Pembagian bersifat subjektif dengan kriteria penilaian hoaks, berdasarkan *tweet* yang diambil dengan rentang waktu tertentu. Setelah itu data diberi label menjadi valid dan hoaks. Data tersebut dikumpulkan lalu dibagi menjadi data latih dan data uji.

username	tweet	replies	retweets	likes	link
detikcom	Pasokan vaksin Corona (COVID-19) yang diberikan China ke Palestina menjadi pasokan yang terbanyak sejauh ini. Berapa jumlahnya? https://t.co/QFdBmi0K1M	2	9	7	https://twitter.com/detikcom/status/1377046880660688906
meongempus	Presiden Trump. HCQ [f] Obat & mencegah Corona. Era Presiden Trump Tidak ada yang meninggal karena vaksin. Presiden Trump Cerdas, sebelum ada Covid 19 dan setelah ada Covid19 pegang Data kematian ada atau tidak covid berpengaruh. 😇	0	0	0	https://twitter.com/MeongEmpus/status/1376905330777464837
tribunkaltim	Data Terbaru Virus Corona di Balikpapan, Jadwal Hingga 14 April 2021 Lansia Dapat Vaksin Covid-19 https://t.co/O1ZdeX7kpy #coronavirus #Corona #COVID19 #COVID #Indonesia #Indonesian #Balikpapan #Sinovac #vaksincorona #Kalimantan #KalimantanTimur #coronamaatregelen #Jokowi	0	0	0	https://twitter.com/tribunkaltim/status/1376866720502059008
eradotid	Presiden Pakistan Arif Alvi pada Senin, (29/3/2021), menyatakan diri positif COVID-19 meski telah menerima dosis pertama vaksin corona. https://t.co/KtdYo5Tcf1	0	0	0	https://twitter.com/eradotid/status/1376760604
rakyatmerdeka99	Bantu Atasi Covid, TNI AD Gelar Vaksinasi Tahap Dua #TNIAD #Vaksin #VaksinasiNasional #Vaksinasi #SatgasCovid19 #Nasional #Corona #coronavirus #COVID19 #COVID—19 #COVID__19 #Covid_19 #COVID #RakyatMerdeka #RMid https://t.co/iZ5XRJu6z2	0	2	2	https://twitter.com/RakyatMerdeka99/status/1376722763054899200

Gambar. 2 Hasil pemindaian *tweet*

C. *Pelabelan*

Pelabelan dilakukan secara manual untuk menentukan hoaks atau tidaknya suatu *tweet*. Pelabelan fakta dan hoaks tersebut dilakukan mengikuti Tabel I, yang berfungsi sebagai referensi proses pelabelan *dataset* yang telah dikumpulkan [10]. Pelabelan pada penelitian ini menggunakan atribut pada Twitter berupa *username*, *link url*, *reply*, dan *hashtag* yang menjadi bukti pendukung sebagai pemilihan hoaks atau validnya informasi pada *tweet*. Atribut provokasi dan sentimen menjadi penilaian terhadap *tweet*.

TABEL I
PENILAIAN *TWEET* BERISI HOAKS

Atribut	Keterangan
Username	Melihat akun dari profil, asli, palsu, samaran.
Tweet	Kalimat yang tersusun dinilai dari informasi yang terkandung, apakah fakta atau hoaks.
Link Url	*Tweet* mengandung *link* gambar atau *url website*, dapat dinilai apakah berisi fakta atau hoaks.
Reply	Jumlah balasan komentar.
Hashtag	Kerelevanan atau keterkaitan *tweet* dengan *hashtag*-nya dapat dinilai apakah merupakan fakta atau hoaks.
Provokasi	Kalimat tweet yang mengandung kebencian, permusuhan dan provokasi.
Sentimen	Kalimat *tweet* yang mengandung ke arah negatif ataupun positif, sehingga tidak terkesan berat sebelah.

Pelabelan dilakukan secara manual dengan tabel penilaian seperti pada Tabel I [10]. Sesuai dengan atribut pada Tabel I, penulis menentukan *tweet* sebagai atribut utama. Lalu diurutkan atribut pendukung, dari *Username* untuk menilai akun tervertifikasi, asli, bodong atau samaran. *Link* Url berisi gambar atau website yang dapat dicek kebenarannya. *Hastag*, menilai hubungannya dengan isi *tweet*. Jika atribut diatas terpenuhi, sisa atribut pendukung tidak diperlukan. Namun jika tidak, makan atribut *reply* diperlukan untuk melihat balasan apakah mengandung informasi tambahan. dan dari 898 *tweet* berkurang menjadi 842 *tweet*. Pengelompokan *dataset* dibagi menjadi dua yaitu hoaks dan valid. Kriteria pada Tabel I digunakan sebagai acuan dalam penentuan hoaks dan validnya suatu data. Berdasarkan hasil pelabelan didapatkan sejumlah data hoaks dan valid seperti terlihat pada Tabel II.

TABEL II
JUMLAH DATA VALID DAN HOAKS

Label	Jumlah	Persentase
Valid	483	57.36%
Hoaks	359	42.64%

D. *Pra-proses Data*

Tahap pra-proses berfungsi untuk mengubah hasil data mentah yang tidak terstruktur menjadi data yang terstruktur. Beberapa tahap yang dilalui dalam tahap ini antara lain *case folding*, tokenisasi, normalisasi, *filtering*, penghilangan *stopwords*, *stemming*, dan pembobotan kata (TF-IDF). Selanjutnya, data siap dibagi menjadi data latih dan data uji.

1) Case Folding: Data mentah yang telah dikumpulkan selanjutnya diproses melalui tahap *case folding* atau *transform case* [15]. *Case folding* yakni tahap dimana semua huruf pada teks diubah menjadi huruf kecil mulai dari karakter ‘a’ hingga ‘z’ [18]. Tabel III merupakan perbandingan sebelum dan setelah dilakukan tahap *case folding*.

TABEL III
PERBANDINGAN SEBELUM DAN SETELAH TAHAP *CASE FOLDING*

***Tweet* (sebelum)**	***Tweet* (setelah)**	**Label**
Data Terbaru Virus Corona di Balikpapan, Jadwal Hingga 14 April 2021 Lansia Dapat Vaksin Covid-19 https://t.co/O1ZdeX7kpy #coronavirus #Corona #COVID19 #COVID #Indonesia #Indonesian #Balikpapan #Sinovac #vaksincorona #Kalimantan #KalimantanTimur #coronamaatregelen #Jokowi	data terbaru virus corona di balikpapan, jadwal hingga 14 april 2021 lansia dapat vaksin covid-19 https://t.co/o1zdex7kpy #coronavirus #corona #covid19 #covid #indonesia #indonesian #balikpapan #sinovac #vaksincorona #kalimantan #kalimantantimur #coronamaatregelen #jokowi	Valid

2) Tokenisasi: Ekstraksi dari teks menjadi kata hingga dikumpulkan bagian terkecil dari dokumen adalah tokenisasi [18]. Kata-kata atau istilah disebut sebagai token. Saat proses tokenisasi, karakter yang nilainya seperti angka, tanda baca, dan karakter selain huruf alfabet akan dianggap sebagai pemisah kata (delimiter) dan tidak punya pengaruh akan teks atau kalimat utamanya. Tokenisasi juga dapat diartikan sebagai penguraian kalimat menjadi kata-kata serta menghilangkan tanda seperti titik (.), koma (,), hingga spasi dan juga karakter lain yang ada dalam sebuah kalimat [14].

Proses tokenisasi dalam penelitian ini menghilangkan (1) *Mention* ke pengguna lain (@), *hashtag*, angka; (2) Tautan url, simbol, emotikon, *tweet* duplikat; (3) Tanda baca titik (.), koma (,), tanya (?), dan sebagainya. Jika *tweet* mengandung tanda-tanda tersebut, maka akan melalui satu sampai tiga tahap proses tokenisasi. Hasil tokenisasi nantinya memuat tabel baru. Tabel IV merupakan perbandingan sebelum dan setelah dilakukan tahap tokenisasi.

TABEL IV
PERBANDINGAN SEBELUM DAN SETELAH TAHAP TOKENISASI

***Tweet* (sebelum)**	***Tweet* (setelah)**	**Label**
rusia klaim indonesia registrasi obat covid-19 buatannya, 4 hari sembuh **sekalipun avifavir sebagai obat anti-covid, obat ini tdk diberikan kpd orang yg sehat. avifavir bkn vaksin, melainkan obat utk menyembuhkan pasien yg terinfeksi virus corona https://t.co/j8atwvfcyq	rusia klaim indonesia registrasi obat covid buatannya hari sembuh sekalipun avifavir sebagai obat anti covid obat ini tdk diberikan kpd orang yg sehat avifavir bkn vaksin melainkan obat utk menyembuhkan pasien yg terinfeksi virus corona	*Hoax*

3) Normalisasi: digunakan sebagai pengoreksi kata singkat atau tidak jelas. Lata-kata tersebut dicocokkan dengan kamus. Selain itu, normalisasi berfungsi mengganti kata yang tidak baku menjadi baku [10]. Di media sosial khususnya *Twitter*, sebuah *tweet* seringkali mengandung kata gaul atau kata yang tidak baku. Tabel V merupakan perbandingan sebelum dan setelah dilakukan tahap normalisasi.

TABEL V
CONTOH PERUBAHAN KATA SEBELUM DAN SETELAH NORMALISASI

Sebelum	**Setelah**
gua	saya
sy	saya
yg	yang
km	kamu
gk	tidak
g	tidak
lu	kamu
ayooo	ayo
kmrn	kemarin
smpai	sampai
dmn	dimana

4) *Penghilangan Stopword:* kumpulan kata yang paling sering muncul dalam dokumen namun tidak mempresentasikan apapun. Oleh karena itu, diperlukan penghilangan *stopword* yang terdeteksi supaya tidak mempengaruhi pembobotan TD-IDF [19]. Penghilangan *stopword* bertujuan untuk menyaring kata yang memiliki arti dan nilai yang lebih. Tabel VII merupakan perbandingan sebelum dan setelah dilakukan tahap penghilangan stopwords.

TABEL VI
PERBANDINGAN SEBELUM DAN SETELAH PENGHILANGAN *STOPWORDS*

***Tweet* (sebelum)**	***Tweet* (setelah)**	**Label**
bangun dulu soal jadi apa nantinya itu belakangan di aceh setelah dana otsus melimpah banyak sekali proyek terbengkalai setelah dibangun	bangun aceh dana otsus melimpah proyek terbengkalai dibangun	*Hoax*

5) Stemming: Hasil yang didapat dari penghilangan *stopword* disempurnakan kembali pada proses *stemming* [15]. Tahap stemming bertujuan untuk menemukan kata dasar dengan menghilangkan imbuhan (afiks) mulai dari awalan, sisipan, hingga akhiran, serta kombinasi dari awalan dan akhiran (sufiks) dari sebuah kata turunan. Proses ini dilakukan dengan mengubah bentuk sebuah kata menjadi kata dasar berdasarkan Bahasa Indonesia yang baik dan benar [19]. *Stemming* digunakan untuk mengamati pengaruh kata dasar dalam klasifikasi berita hoaks [8]. Tabel VII merupakan perbandingan sebelum dan setelah dilakukan tahap *stemming.*

TABEL VII
PERBANDINGAN SEBELUM DAN SETELAH TAHAP *STEMMING*

***Tweet* (sebelum)**	***Tweet* (setelah)**	**Label**
kepala dinas kesehatan kabupaten jayapura khairul lie mengatakan program vaksinasi covid kabupaten jayapura papua telah menggunakan dosis vaksin pon papua dorong umkm	kepala dinas sehat kabupaten jayapura khairul lie program vaksinasi covid kabupaten jayapura papua dosis vaksin pon papua dorong umkm	Valid

F. Seleksi Fitur dengan TF-IDF

Langkah selanjutnya adalah seleksi fitur yang dilakukan dengan *Term Frequency-Inverse Document Frequency* (TF-IDF) yakni menghitung bobot setiap kata pada *tweet*, baik dokumen uji maupun latih. Langkah ini bertujuan untuk memaparkan pentingnya peran kata dalam data menggunakan fitur gabungan unigram maupun bigram. Dengan begitu, didapatkan jarak antar dokumen yang merupakan hasil pembobotan dengan menggunakan vektor [18]. Gambar. 3 memperlihatkan hasil TF-IDF yang mentransformasi data berbentuk teks ke dalam bentuk vektor.

```
(0, 1380)    0.38027973651551783
(0, 338)     0.35584675920230047
(0, 360)     0.19266199236674536
(0, 358)     0.14515023262962792
(0, 2082)    0.19540151577383982
(0, 1408)    0.7952304196011589
(1, 1380)    0.38027973651551783
(1, 338)     0.35584675920230047
(1, 360)     0.19266199236674536
(1, 358)     0.14515023262962792
(1, 2082)    0.19540151577383982
(1, 1408)    0.7952304196011589
(2, 1440)    0.18660010273282862
(2, 1154)    0.18660010273282862
(2, 386)     0.1653534023967889
(2, 1426)    0.19344001249465262
(2, 335)     0.2022581898700495
(2, 1973)    0.16858296657336688
(2, 509)     0.1624319062325245
(2, 324)     0.1441067020607491
(2, 1332)    0.18101148953400975
(2, 654)     0.21468671283069235
(2, 2014)    0.644060138492077
(2, 1528)    0.392019266262085
(2, 360)     0.29400886741027094
(806, 909)   0.6496273543327079
(807, 58)    0.43399865760829237
(807, 1808)  0.43399865760829237
(807, 1244)  0.43399865760829237
(807, 524)   0.43399865760829237
(807, 1678)  0.36948647851530053
(807, 817)   0.2249682099120874
(807, 1237)  0.14127786386413677
(807, 1924)  0.14102243403428477
(807, 1713)  0.1400105202516809
(808, 455)   0.3419965382232675
(808, 315)   0.30815048474475937
(808, 1237)  0.10488379508090205
(808, 1628)  0.32219786614093743
(808, 868)   0.30815048474475937
(808, 27)    0.32496898276436675
(808, 1457)  0.274304431266251126
(808, 1658)  0.25450575918392115
(808, 1924)  0.1046941655862382
(808, 381)   0.26855314058009916
(808, 1713)  0.20788585435252346
(808, 1147)  0.17928654644113654
(808, 1375)  0.22956238227195117
(808, 617)   0.206612324309234960
(808, 2023)  0.2808247068766609
```

Gambar. 3 Hasil vektor pembobotan dengan TF-IDF

G. Klasifikasi

Klasifikasi merupakan sebuah analisis data dengan mengekstrak suatu model untuk mempresentasikan kelas data. Model yang disusun mencakup pengklasifikasian serta prediksi kategori kelas label. Teknik klasifikasi dapat diimplementasikan dalam banyak bidang, salah satunya adalah deteksi penipuan [11].

Setelah tahap TF-IDF maka akan dilakukan klasifikasi dengan algoritma yang digunakan, yaitu SGD, *Naïve Bayes*, *Random Forest* dan *Rocchio*. Keempat algoritma tersebut akan dibandingkan performanya dan dievaluasi kinerjanya dengan mengukur presisi, *recall*, nilai F-1, dan akurasinya.

IV. HASIL DAN PEMBAHASAN

Setelah *dataset* dipra-proses hingga diseleksi fiturnya dengan TF-IDF, selanjutnya dibagi menjadi data uji dan data latih. Total dari 842 *tweet* yang telah diberi label, selanjutnya dipra-proses dan menyisakan 810 *tweet* menjadi suatu informasi singkat dengan kalimat yang baku. Perbandingan yang digunakan untuk setiap klasifikasi dibagi menjadi data latih dan uji dengan perbadingan 60:40, 70:30, 80:20, 90:10. Tabel VIII merupakan tabel pembagiannya.

Semua proses perhitungan dalam penelitian ini menggunakan bahasa pemrograman Python, dengan antarmuka Jupyter Notebook dan Visual Studio Code. Rumus perhitungan atau *library* dapat diakses secara *open source* dan dapat digunakan sesuai kepentingan para *developer*.

TABEL VIII
PEMBAGIAN JUMLAH *TWEET* UNTUK DATA LATIH DAN UJI

Dataset Tweet	**Pembagian Jumlah *Tweet* (810)**			
	60:40	**70:30**	**80:20**	**90:10**
Latih	468	567	648	729
Uji	324	243	162	81

Hasil pembagian digunakan untuk mengetahui performa tiap algoritma klasifikasi dan selanjutnya dievaluasi melalui perhitungan nilai presisi, *recall*, F-1, dan akurasi. Tabel IX merupakan pembagian data latih dan data uji.

TABEL IX
CONFUSION MATRIX

Faktual	**Prediksi**	
	Valid	**Hoaks**
Valid	*True positive (TP)*	*False negative (FN)*
Hoaks	*False positive (FP)*	*True negative (TN)*

Saat melakukan perhitungan, prediksi pada *True Positive* (TP) akan menunjukkan hasil faktual yang valid sama seperti yang dimuat di *dataset*. *False Positive* (FP) terjadi ketika prediksi ternyata tidak hoaks tapi dimuat sebagai *tweet* hoaks. *True Negative* (TN) dinyatakan saat prediksi benar sebagai hoaks. *False Negative (*FN) adalah saat *tweet* hoaks tapi dimuat sebagai valid. Tabel IX adalah tabel kombinasi yang nantinya digunakan untuk memprediksi nilai dari akurasi, presisi, *recall*, dan F-1.

A. Akurasi

Akurasi diartikan sebagai tingkat kedekatan nilai dari hasil prediksi yang benar dari nilai aktual [20]. Dari Tabel IX, *confusion matrix* dapat digunakan untuk menghitung persen akurasi dari model klasifikasi dengan Rumus (1).

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (1)$$

TABEL X
HASIL AKURASI DARI TIAP ALGORITMA KLASIFIKASI

Algoritma	**Akurasi (%)**			
	60:40	**70:30**	**80:20**	**90:10**
SGD	83.95	84.77	84.56	86.41
Naïve Bayes	81.48	83.12	82.09	81.48
Random Forest	80.55	80.24	81.48	82.71
Rocchio	74.07	73.25	78.39	72.83

Berdasarkan Tabel X dapat dilihat bahwa algoritma terbaik pada perhitungan akurasi adalah SGD dengan data latih dan uji 90:10 sebesar 86.41%. SGD juga tetap memiliki akurasi terbaik dibandingkan algoritma lain dengan posisi tertinggi kedua 84.77% pada perbandingan 70:30, dan ketiga 84.56% pada perbandingan 80:20. Pada hasil algoritma *Naïve bayes* dan *Random Forest* nilai akurasi-nya tidak berbeda jauh. Hasil rata-rata masing-masing algoritma mulai dari SGD, *Naïve Bayes*, *Random Forest*, *Rocchio* secara berurutan adalah 84.87%, 82.02%, 81.25%, 74.64%.

B. Presisi

Presisi adalah perbandingan untuk mengukur keakuratan seluruh hasil *dataset* dan mengetahui hasil klasifikasi kategori yang sebenarnya. Presisi dapat juga diartikan sebagai kecocokan antara sebuah pertanyaan dengan jawaban atau informasi yang diberikan [21]. Presisi merupakan hasil dari perhitungan yang benar (TP) dibagi dengan jumlah data yang terindentifikasi oleh sistem. Rumus presisi diberikan pada Rumus (2).

$$\text{Presisi} = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

TABEL XI
HASIL PRESISI DARI TIAP ALGORITMA KLASIFIKASI

Algoritma	Presisi (%)			
	60:40	**70:30**	**80:20**	**90:10**
SGD	78.41	75.67	76.71	91.17
Naïve Bayes	83.33	85.33	85.71	77.41
Random Forest	78.57	72.64	86.95	75.00
Rocchio	69.23	74.01	76.74	71.42

Sama halnya dengan akurasi, pada perhitungan presisi, seperti pada Tabel XI, algoritma SGD dengan perbandingan 90:10 masih memiliki hasil terbaik. Namun, jika dilihat dari perbandingan data latih dan data uji, presisi pada SGD tergolong tidak stabil. Algoritma *Naive Bayes* menghasilkan nilai presisi yang lebih stabil dibandingkan SGD, dengan presisi tertinggi sebesar 85.71% pada perbandignan 80:20. Selain itu *Random Forest* juga memiliki hasil yang cukup baik, yaitu sebesar 86.95%, namun juga tidak stabil seperti SGD. Hasil rata-rata masing-masing algoritma mulai dari SGD, *Naive Bayes*, *Random Forest*, *Rocchio* secara berurutan adalah 80.49%, 82.95%, 78.29%, 72.85%.

C. Recall

Tingkat kesuksesan sebuah sistem dalam mendeteksi suatu kelompok menjadi sebuah parameter *recall*. *Recall* merupakan perhitungan dari pembagian antara jumlah data yang benar dengan jumlah yang seharusnya [20]. Rumus *recall* diberikan pada Rumus (3).

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

TABEL XII
HASIL *RECALL* DARI TIAP ALGORITMA KLASIFIKASI

Algoritma	*Recall* (%)			
	60:40	**70:30**	**80:20**	**90:10**
SGD	83.84	89.36	87.50	79.48
Naïve Bayes	65.38	68.08	65.62	75.00
Random Forest	77.46	81.91	74.07	84.37
Rocchio	79.59	78.33	81.48	81.39

Berdasarkan Tabel XII, SGD menjadi algoritma terbaik dengan pencapaian tertinggi 89.36% pada data latih dan uji 70:30. Posisi kedua juga tidak jauh berbeda sebesar 87.50% pada perbandingan 80:20, lalu diikuti 83.84% pada perbandingan 60:40. Uniknya, *Rocchio* memiliki nilai terbaiknya pada perhitungan *recall* sebesar 81.48% dan stabil. Hasil rata-rata masing-masing algoritma mulai dari SGD, *Naive Bayes*, *Random Forest*, *Rocchio* secara berurutan adalah 85.05%, 68.52%, 79.45%, 80.20%.

D. F-1

Nilai F-1 atau yang disebut juga dengan *harmonic mean*, digambarkan sebagai pengaruh relatih dari sebuah presisi dan *recall*. F-1 sendiri dapat menunjukkan performa akan suatu metode yang dipakai [21]. Rumus F-1 diberikan pada Rumus (4).

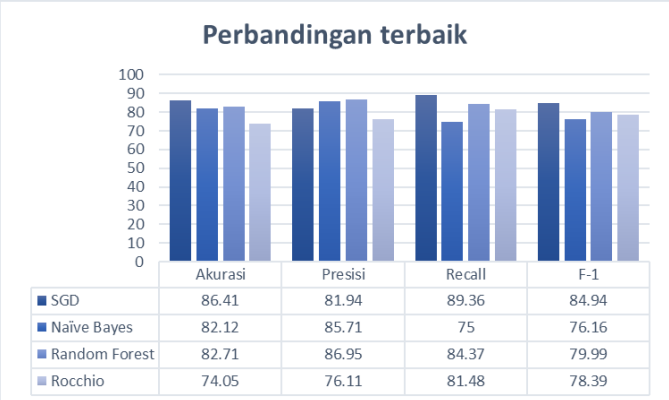
$$\text{F-1} = \frac{2 \times \text{recall} \times \text{presisi}}{\text{recall} + \text{presisi}} \times 100\% \quad (4)$$

TABEL XIII
HASIL F-1 DARI TIAP ALGORITMA KLASIFIKASI

Algoritma	Akurasi (%)			
	60:40	**70:30**	**80:20**	**90:10**
SGD	81.04	81.95	81.75	84.93
Naïve Bayes	73.27	75.73	74.33	76.19
Random Forest	78.01	77.00	79.99	79.41
Rocchio	78.39	76.11	79.04	76.08

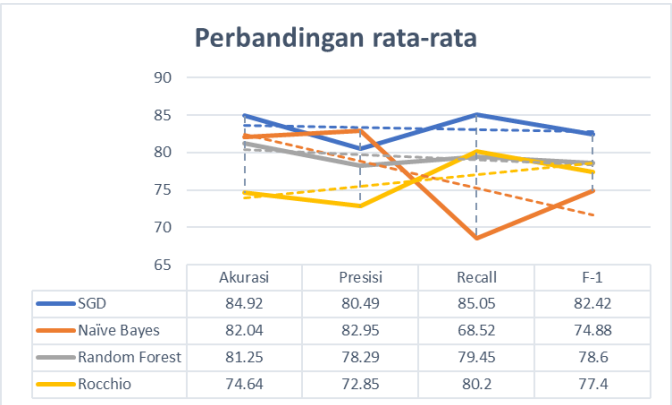
Pada perhitungan F-1 pada Tabel XIII, SGD dominan menjadi algoritma klasifikasi terbaik dengan perolehan nilai 84.93%, 81.95%, 81.75, dan 81.04% pada perbandingan 90:10, 70:30, 80:20, dan 60:40. Jika melihat algoritma lain, sebenarnya *Random Forest* juga tidak kalah jauh karena memiliki nilai hampir 80%. Hasil rata-rata masing-masing algoritma SGD, *Naive Bayes*, *Random Forest*, *Rocchio* secara berurutan adalah 82.41%, 74.88%, 78.60%, 76.32%.

Berdasarkan Gambar. 4 dan Gambar. 5 dapat dilihat bahwa persentase keempat algoritma dari akurasi, presisi, *recall*, hingga F-1 mengalami fluktuasi. SGD memiliki rata-rata yang cukup stabil namun mengalami fluktuasi dengan selisih yang cukup besar. Hal tersebut menunjukkan bahwa hasil penelitian ini sesuai dengan grafik iterasi pengujian model SGD dalam 10 kali percobaan (*Cross-Validation* 10) dalam [15], dimana SGD mengalami fluktuasi namun cukup stabil dengan selisih yang cukup besar dan terkadang mendapat nilai tertingginya.



	Akurasi	Presisi	Recall	F-1
SGD	86.41	81.94	89.36	84.94
Naïve Bayes	82.12	85.71	75	76.16
Random Forest	82.71	86.95	84.37	79.99
Rocchio	74.05	76.11	81.48	78.39

Gambar. 4. Perbandingan terbaik dari akurasi, presisi, *recall*, dan F-1



	Akurasi	Presisi	Recall	F-1
SGD	84.92	80.49	85.05	82.42
Naïve Bayes	82.04	82.95	68.52	74.88
Random Forest	81.25	78.29	79.45	78.6
Rocchio	74.64	72.85	80.2	77.4

Gambar. 5 Perbandingan rata-rata dari akurasi, presisi, *recall*, dan F-1

Naïve Bayes memiliki fluktuasi paling besar dengan perubahan persentase yang cukup drastis. Secara keseluruhan persentase terendah ditemukan pada algoritma *Rocchio*, namun cenderung naik diikuti oleh *Random Forest* yang cukup stabil dengan fluktuasi selisih yang kecil. Pada penelitian [7], nilai akurasi *Rocchio* menunjukkan hasil klasifikasi yang lebih baik jika dibandingkan dengan *Naive Bayes*. Hal tersebut dikarenakan total kesalahan klasifikasi data uji pada *Rocchio* lebih rendah dibanding *Naive Bayes*. Pada *Naive Bayes* dapat terjadi kasus *underflow* atau rendahnya perolehan nilai peluang mengakibatkan klasifikasi ini sulit untuk membaca data. Hal tersebut menjadi alasan *Naive Bayes* memiliki nilai akurasi yang rendah.

Random Forest adalah algoritma yang dapat mengembangkan hasil klasifikasinya dengan menerapkan fitur acak untuk setiap data yang dipanggil, sehingga dapat meningkatkan akurasinya [9]. Metode yang digunakan dengan membangun pohon keputusan ini memiliki hasil fluktuasi penurunan terbesar dalam perbandingan rata-rata. Hal tersebut dikarenakan *Random Forest* adalah salah satu algoritma klasifikasi yang dapat membuat banyak data sekaligus dengan sistem acak, yang dapat meningkatkan akurasi juga mengalami penurunan.

V. KESIMPULAN

Dari hasil dan pembahasan didapatkan kesimpulan bahwa identifikasi *tweet* dari Twitter yang mengandung berita atau informasi hoaks dapat diklasifikasi dengan pembelajaran mesin. Pada penelitian ini digunakan berita tentang kesehatan dengan kata kunci perpanjang PSBB, mutasi corona, vaksin, covid-19; pendidikan dengan kata kunci sekolah dan sekolah tatap muka; pemerintahan dengan kata kunci dana otsus, dana bansos, utang pemerintah. Total setelah pra-proses berjumlah 810 *tweet* dibagi menjadi dua label antara lain valid dan hoaks. Pada label valid ditemukan sebanyak 483 *tweet*, sedangkan pada label hoaks sebanyak 359 *tweet*.

Berita hoaks dan valid dapat diklasiikasi dengan pembelajaran mesin menggunakan empat algoritma klasifikasi yang berbeda, yaitu SGD, *Naive Bayes*, *Random Forest*, *Rocchio* serta menggunakan tahapan pra-proses (*case folding*, tokenisasi, normalisasi, penghilangan *stopwords*, *stemming*), dan seleksi fitur TF-IDF.

Dari hasil perbandingan dapat disimpulkan bahwa algoritma SGD memiliki akurasi, *recall*, dan F-1 terbaik yaitu masing-masing sebesar 86.41%, 89.36%, dan 84.39%, sedangkan hasil terbaik pada presisi diperoleh dari algoritma *Random Forest*, yaitu sebesar 86.95%. Sementara itu, untuk perbandingan rata-rata akurasi, *recall*, dan F-1 terbaik ditemukan pada algoritma SGD dengan hasil secara berturut sebesar 84.92%, 85.05%, dan 82.42%. sedangkan pada perbandingan rata-rata presisi tertinggi diperoleh algoritma *Naive Bayes* sebesar 82.95%.

Dengan hasil presentase yang disajikan, maka SGD menjadi algoritma terbaik dalam mengidentifikasi *tweet* hoaks di Twitter, karena memiliki kemampuan yang baik pada data berdimensi tinggi dan ingatan yang tinggi juga dapat memperbaiki hasil dari ingatan sebelumnya. *Random Forest* merupakan algoritma terbaik kedua yang dapat meningkatkan akurasi dengan memanggil simpul pohon yang lainnya dan lebih baik dalam menangani data yang besar. Hasil *Random Forest* dinilai cukup baik, karena menghasilkan fluktuasi yang stabil pada penelitian ini. Namun akan lebih baik jika data yang dimiliki lebih banyak.

Naive Bayes dengan teorema Bayes-nya yang menerapkan klasifikasi sederhana dan semua fitur dianggap tidak saling berhubungan, ternyata tidak cukup baik menangani klasifikasi pada penelitian ini sehingga terjadi fluktuasi presentase penurunan yang cukup besar dari hasil presisi 82.95% menjadi 68.52% pada *recall* dan 74.88% pada F-1. Pada posisi terakhir algoritma *Rocchio* walaupun memiliki kinerja yang lebih baik daripada *Naïve Bayes* pada penelitian lain, dengan melakukan perhitungan batas kelas menggunakan *centroid* untuk mendapatkan hasil batas-batasnya ternyata memiliki selisih yang cukup jauh dengan *Naïve Bayes* pada penelitian ini. Walaupun memiliki performa terburuk, tetapi *Rocchio* mengalami fluktasi kenaikan yang cukup baik pada *recall* dan F-1. Hal ini dikarenakan *Rocchio* hanya melakukan pendekatan dalam mencari kemiripan dengan *centroid* kelas sehingga melupakan data lain yang berbeda pada penelitian ini.

Kesimpulan akhir didapatkan bahwa perubahan persentase yang paling stabil ditemukan pada algoritma *Random Forest* dengan selisih perubahan kecil dan SGD yang memiliki selisih cukup besar. Fluktuasi terbesar ditemukan pada algoritma *Naïve Bayes* dengan penurunan dan *Rocchio* dengan kenaikan yang cukup drastis.

REFERENSI

[1] J. Hutahaean, *Konsep Sistem Informasi*, 1st ed. Yogyakarta: Deepublish, 2014.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "Laporan Survei Internet APJII 2019 - 2020 [Q2]," 2020. https://www.apjii.or.id/survei.

[3] Masyarakat Telematika Indonesia (Mastel), "Hasil Survey Wabah Hoax Nasional 2019," 2019. https://mastel.id/hasil-survey-wabah-hoax-nasional-2019/.

[4] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," *Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017*, vol. 2018-Janua, pp. 45–49, 2017, doi: 10.1109/ICITACEE.2017.8257673.

[5] J. C. W. Pantouw, "Perbandingan klasifikasi rocchio dan multinomial naïve bayes pada analisis sentimen data twitter bahasa indonesia," *Dep. Ilmu Komput. Fak. Mat. Dan Ilmu Pengetah. Alam Inst. Pertan. Bogor 2017*, 2017.

[6] M. D. Nugraha, J. A. Utama, and S. Sulistiani, "Implementasi Metode Random Forest Dalam Memprediksi Peristiwa Flare," *Pros. Semin. Nas. Fis.*, pp. 258–263, 2018.

[7] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *J. Ilmu Komput. Agri-Informatika*, vol. 5, pp. 1–10, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jika.

[8] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Cybermatika*, vol. 3, no. 2, pp. 1–8, 2015, [Online]. Available: https://www.mendeley.com/import/.

[9] T. Trisna Astono Putri, H. S. Warra, I. Yanti Sitepu, and M. Sihombing, "Analysis and Detection of Hoax Contents in Indonesian News Based on Machine Learning," *J. Informatics Pelita Nusant.*, vol. 4, no. 1, pp. 19–26, 2019, [Online]. Available: http://e-jurnal.pelitanusantara.ac.id/index.php/JIPN/article/view/489/291.

[10] B. Irena and Erwin Budi Setiawan, "Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 711–716, 2020, doi: 10.29207/resti.v4i4.2125.

[11] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo J. Inf. Technol.*, vol. 1, no. 1, p. 1, 2019, doi: 10.21580/wjit.2019.1.1.3915.

[12] C. Juditha, "Hoax Communication Interactivity in Social Media and Anticipation (Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya)," *J. Pekommas*, vol. 3, no. 1, p. 31, 2018, doi: 10.30818/jpkm.2018.2030104.

[13] B. Liu, *Sentiment Analysis and Opinion Mining*, no. April. 2012.

[14] B. Kurniawan, M. A. Fauzi, and A. W. Widodo, "Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 10, pp. 1193–1200, 2017.

[15] R. Umar, I. Riadi, and P. Purwono, "Klasifikasi Kinerja Programmer pada Aktivitas Media Sosial dengan Metode Stochastic Gradient Descent," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 5, no. 2, p. 55, 2020, doi: 10.31328/jointecs.v5i2.1324.

[16] F. R. Lumbanraja, J. I. Komputer, and F. U. Lampung, "Sistem Pencarian Data Teks dengan Menggunakan Metode Klasifikasi Rocchio (Studi Kasus : Dokumen Teks Skripsi)," 2013.

[17] R. R. Muzad Aad Miqdad, "Korpus Berita Daring Bahasa Indonesia Dengan Depth First Focused Crawling," *Sentrinov (Seminar Nas. Terap. Ris. Inov.*, vol. 2, no. 1, pp. 11–20, 2016.

[18] F. N. Rozi and D. H. Sulistyawati, "Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor Dan Pembobotan Menggunakan Tf-Idf," *Konvergensi*, vol. 15, no. 1, 2019, doi: 10.30996/konv.v15i1.2828.

[19] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *J. Mantik Penusa*, vol. 2, no. 1, pp. 35–40, 2018.

[20] C. A. Ul Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 1–6, 2018, doi: 10.23919/IConAC.2018.8748995.

[21] A. A. Puspitasari, E. Santoso, and Indriati, "Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved k-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, pp. 486–492, 2018.