DETEKSI BERITA HOAX PADA WEBSITE TURNBACKHOAX DENGAN MENGGUNAKAN *MACHINE* LEARNING

SKRIPSI

Rahmawati
11160940000035

\

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UIN SYARIF HIDAYATULLAH JAKARTA
2021 M / 1442 M

PERNYATAAN

DENGAN INI SAYA MENYATAKAN BAHWA SKRIPSI INI BENAR-BENAR HASIL KARYA SENDIRI YANG BELUM PERNAH DIAJUKAN SEBAGAI SKRIPSI ATAU KARYA ILMIAH PADA PERGURUAN TINGGI ATAU LEMBAGA MANAPUN

Jakarta, 08 Februari 2022

Rahmawati

NIM. 11160940000035

PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan dibawah ini:

Nama : Rahmawati
NIM : 11160940000035
Program Studi : Matematika Fakultas Sains dan Teknologi

Demi pengembangan ilmu pengetahuan, saya menyetujui untuk memberikan **Hak Bebas Royalti Non-Eksklusif** (*Non-Exclusive Free Right*) kepada Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta atas karya ilmiah saya yang berjudul:

"DETEKSI BERITA HOAX PADA WEBSITE TURNBACKHOAX DENGAN MENGGUNAKAN *MACHINE* LEARNING"

beserta perangkat yang dipelukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini, Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta berhak menyimpan, mengalihmedia/formatkan mengelolanya dalam bentuk pangkalan data (*data base*), mendistribusikannya, dan menampilkan/mempublikasikannya di internet dan media lain untuk kepentingan akademis tanpa perlu meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Segala bentuk tuntuan hokum yang timbul atas pelanggaran Hak Cipta karya ilmiah ini menjadi tanggungjawab saya sebagai penulis. Demikian pernyataan ini yang saya buat dengan sebenarnya.

Jakarta, 08 Februari 2022

(Rahmawati)

PERSEMBAHAN DAN MOTTO

Puji Syukur atas Kehadirat Allah SWT dalam hidup, atas izin dan berkat-Nya penulis mampu menyelesaikan skripsi ini.

Sholawat beriringan salam tercurahkan kepada Nabi Muhammad SAW

Skripsi ini penulis persembahkan untuk kedua orangtua beserta keluarga yang telah memberikan doa dan dukungan yang begitu luar biasa.
Skripsi ini juga dipersembahkan untuk yang terkasih, teman-teman.

MOTTO

“Sesungguhnya Allah tidak akan merubah suatu kaum sehingga mereka merubah keadaan yang ada pada diri mereka sendiri”

KATA PENGANTAR

Assalamualaikum Wr. Wb.

Alhamdulillahirabbil ‘alamin, puji dan syukur penulis panjatkan kepada Allah SWT yang telah memberikan nikmat, berkat, rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan penelitian ini. Shalawat serta salam senantiasa penulis curahkan kepada junjungan Nabi Muhammad SAW beserta keluarga, para sahabat dan para pengikutnya.

Penulis menyelesaikan penelitian ini untuk memperoleh gelar sarjana Matematika. Dalam penyusunan, peneliti tidak luput dari kesulitan, hambatan dan rintangan. Namun, banyak pihak-pihak yang memberikan doa, dukungan, bantuan, motivasi dan memberikan semangat sehingga penelitian ini dapat terselesaikan. Oleh karena itu peneliti mengucapkan terima kasih kepada:

1. Bapak Ir. Nashrul Hakiem, S.Si., M.T., Ph.D, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta.
2. Ibu Dr. Summa’inna, M.Si., selaku Ketua Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta dan Ibu Irma Fauziah M.Sc., selaku Sekretaris Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.
3. Bapak Dr. Taufik Edy Sutanto, M.Sc.Tech., selaku pembimbing I yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan penulis dalam penyusunan skripsi ini hingga akhirnya dapat terselesaikan.
4. Ibu Nurmaleni, M.Si., selaku pembimbing II yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan penulis dalam penyusunan skripsi ini hingga akhirnya dapat terselesaikan.
5. Bapak Muhaza Liebenlito, M.Si selaku dosen penguji I dan dosen pembimbing Akademik yang telah memberikan banyak motivasi, dan arahan kepada penulis selama kuliah di Program Studi Matematika Fakultas Sains dan Teknologi UIN Jakarta.

6. Bapak M. Irvan Septiar Musti, M.Si selaku dosen penguji II yang telah memberikan banyak motivasi, dan arahan kepada penulis selama kuliah di Program Studi Matematika Fakultas Sains dan Teknologi UIN Jakarta.
7. Mama dan Bapak, Abang Ulhaq, Kak Lia yang tiada henti-hentinya memberikan doa, dukungan, semangat dan motivasi hingga peneliti akhirnya mampu menyelesaikan skripsi ini.
8. Angga Saputra dan Fajar Nur Aulia yang selalu menjadi tempat untuk berbagi keluh kesah selama penyusunan skripsi ini dan memberikan dukungan kepada penulis.
9. Inggi Putriana yang sudah direpotkan penulis untuk konsultasi skripsi ini hingga selesai.
10. Teman-teman yaitu Nadhila Farhana, Panji Reza, Badriatul Mursyidah, Hanny Nurrohmah, Indah Tri Nurlita, Danny Yoga, yang selalu hadir di segala suka dan duka selama kuliah hingga akhir.
11. Silma Novshienza dan Riana Indah yanh selalu memberikan dukungan kepada penulis
12. Teman-teman Matematika 2016 UIN Syarif Hidayatullah Jakarta yang tidak dapat penulis sebutkan satu persatu.
13. Seluruh pihak yang secara langsung maupun tidak langsung telah membantu, mendukung serta mendoakan penulis dalam penyelesaian skripsi ini. Meski tidak tertulis namun tidak mengurangi rasa cinta dan terima kasih dari penulis.

Penulis menyadari bahwa masih ada kesalahan dalam penyusunan skripsi ini. Maka dari itu penulis mengharapkan kritik dan saran yang membangun supaya menjadi bahan perbaikan bagi peneliti selanjutnya. Penulis juga berharap penelitian ini bermanfaat bagi siapapun yang membacanya.

Wassalamualaikum Wr.Wb

Jakarta,08 Februari 2022

Rahmawati

ABSTRAK

Rahmawati, Deteksi Berita *Hoax* Pada Website *Turnbackhoax* Dengan Menggunakan *Machine Learning*, dibawah bimbingan **Dr. Taufik Sutanto, M.ScTech dan Nurmaleni, M.si**

Penyebaran berita hoax terus meningkat setiap tahunnya, hal itu terjadi karena kurangnya tingkat integritas masyarakat dalam meninjau lebih lanjut berita yang beredar. Sebuah berita hoax biasanya menggunakan ancaman atau informasi yang menyesatkan untuk membuat mereka percaya hal-hal yang tidak nyata. Penelitian ini mendeteksi berita hoax dengan mengklasifikasikan berita menggunakan *machine learning* pada website Turnbackhoax dengan jumlah berita yang dianalisis sebanyak 4.551. Data dilakukan pelatihan untuk mendapatkan model yang optimal dan dapat memprediksi klasifikasi dengan *Machine Learning* yang digunakan yaitu *Random Forest Classifier,* Regresi Logistik dan *Support Vector Machine.* Model terbaik dalam penelitian ini yaitu *Support Vector Machine* dengan akurasi sebesar 83% dan recall pada kelas hoax sebesar 99%. Tujuan dari pekerjaan ini adalah untuk menemukan solusi yang dapat digunakan oleh pengguna untuk mendeteksi dan menyaring situs yang berisi informasi palsu dan menyesatkan.

Kata Kunci: *Cross Validation, Grid Search, Hoax Detection, Machine Learning, Support Vector Machine.*

ABSTRACT

Rahmawati, *Hoax News Detection on Turnbackhoax website Using Machine Learning,* under the guidance of **Dr. Taufik Sutanto, M.ScTech** and **Nurmaleni, M.si**

The increase in hoax news continues every year. This happened because public doesn't have the same level of integrity, in reviewing existing news. A hoax news usually used threats or misleading information to make them believe things that are not real. This research proposes an experiment using Machine Learning to detect hoax news on the Turnbackhoax website with a total of 4.551 news. Data will be trained to get an optimal model for predicting classification using *Machine Learning* used *Random Forest Classifier, Logistic Regression,* and *Support Vector Machine.* From this research, the results of the best model is *Support Vector Machine* obtained an accuracy is 83% and recall hoax class is 99%. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information.

Keywords: *Cross Validation, Grid Search, Hoax Detection, Machine Learning, Support Vector Machine.*

DAFTAR ISI

DAFTAR GAMBAR

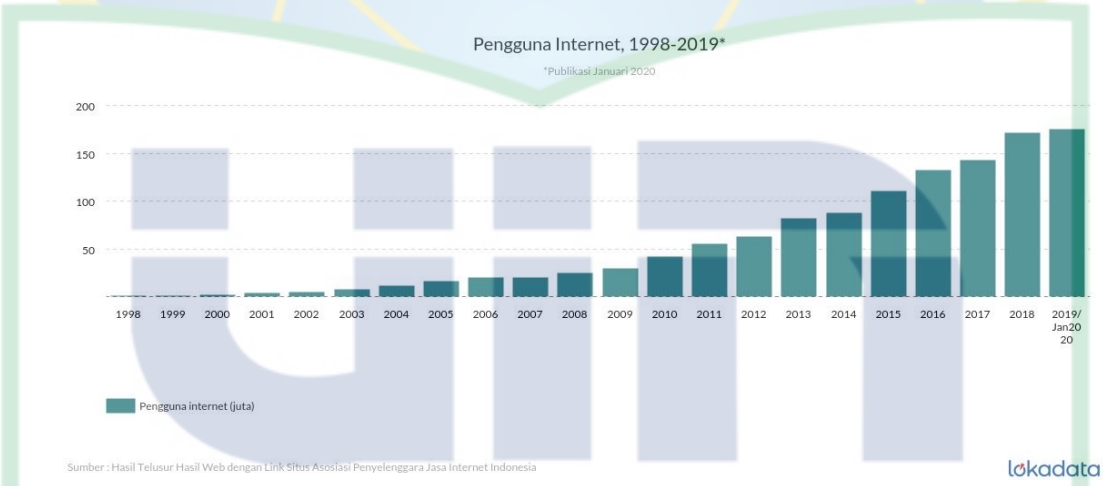
DAFTAR TABEL

BAB I
PENDAHULUAN

Pada bab ini, penulis akan menjelaskan mengenai gambaran umum pelaksanaan penelitian yang mencakup latar belakang, rumusan masalah, batasan masalah, tujuan penelitian dan manfaat penelitian. Hal tersebut akan dijelaskan secara berurutan pada bab ini.

1.1 Latar belakang

Perkembangan internet yang begitu pesat membawa banyak perubahaan dalam kehidupan, salah satunya dalam berkomunikasi tidak hanya dalam media komunikasi, penggunaan internet juga dapat berperan sebagai media penyebaran informasi. Informasi dapat tersebar melalui internet dengan cepat dan dapat dikonsumsi oleh masyarat dengan mudah.



Gambar 1.1. Pengguna Internet,1998-2019 (Sumber: https://lokadata.beritagar.id)

Berdasarkan Gambar 1.1, pertumbuhan pengguna aktif internet di Indonesia terus meningkat setiap tahunnya, penggunaan internet merupakan wadah yang tepat sebagai media komunikasi dan juga berperan sebagai media penyebaran informasi. Penyebaran informasi dengan mudah merupakan hal yang positif namun tidak semua informasi yang disebarkan di internet berupa fakta, banyaknya berita yang beredar di internet adalah berita yang tidak benar atau sering disebut *hoax*. Berita *hoax* merupakan informasi menyesatkan dan

manipulatif dengan menyebarkan informasi yang salah namun dianggap benar [1]. Informasi *hoax* dapat merugikan banyak pihak yang tidak bersalah dengan meyakini pembaca tentang kejadian yang tidak benar. Berdasarkan Surat Al-Hujurat ayat 6 dapat dipahami bahwa ayat ini menunjukan dengan jelas perlunya memeriksa denga teliti sebelum dipahami untuk mempercayai berita dan disebarkan kepada orang lain agar tidak menimbulkan musibah.

Allah SWT berfirman dalam Surat Al-Hujurat ayat 6:

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوٓاْ إِن جَآءَكُمْ فَاسِقٌۢ بِنَبَإٖ فَتَبَيَّنُوٓاْ أَن تُصِيبُواْ قَوْمَۢا بِجَهَٰلَةٖ فَتُصْبِحُواْ عَلَىٰ مَا فَعَلْتُمْ نَٰدِمِينَ

Artinya:

Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti, agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu

Setiap tahun semakin banyak berita *hoax* beredar di kalangan masyarakat, karena banyaknya berita *hoax* munculah tindakan-tindakan pencegahan berita *hoax* dan banyak media yang menyajikan layanan untuk melaporkan konten yang diduga mengandung unsur *hoax* dan SARA. Salah satu upaya yang dapat dilakukan yaitu dengan mengklasifikasikan berita *hoax dan non-hoax* menggunakan *machine learning*, dimana *machine learning* merupakan salah satu cabang dari kecerdasan buatan yang sedang berkembang pesat dan banyak digunakan untuk berbagai tujuan. Oleh karena itu, penggunaan *machine learning* untuk proses klasifikasi diharapkan dapat mempersingkat waktu serta memiliki keakuratan yang tinggi dalam hal akurasi. Sehingga kedepannya penyebaran berita *hoax* bisa berkurang. Dengan demikian, masyarakat tidak lagi dibuat bingung dengan informasi atau berita yang beredar. Karena tidak semua masyarakat memiliki tingkat inetegritas yang sama untuk meninjau kembali kebenaran berita yang beredar dengan melihat visualisasi keterkaitan kata pada wordcloud dan wordlink untuk memberikan kemudahan dalam melakukan interpretasi dan melakukan interpetasi langsung terkait berita *hoax*.

Penelitian klasifikasi berita *hoax* telah dilakukan sebelumnya oleh Muhammad Athaillah, dkk pada tahun 2020 dengan menggunakan data berita *hoax* yang berasal dari turnbackhoax dan berita *non-hoax* berasal dari kompas.com dan detik.com [2]. Pada penelitian tersebut menggunakan metode *Naïve Bayes Classifier* (NBC) untuk melakukan klasifikasi dengan data yang digunakan sebanyak 200 artikel berita. Dalam penelitian tersebut diperoleh akurasi sebesar 93%. Penelitian semacam ini telah dilakukan sebelumnya oleh Faizal Nur Rozi dan Dwi Harini Sulistyawati pada tahun 2019 melakukan klasifikasi berita *hoax* pilpres [3]. Dalam penelitiannya, metode klasifikasi yang digunakan adalah *Modified K-Nearest Neighbor* (KNN) dengan nilai akurasi sebesar 92,3%. Selain itu Diki Arisandi, dkk pada tahun 2020 juga melakuka penelitian klasifikasi berita *hoax* pada hasil pencarian berita online [4]. Pada penelitian tersebut, peneliti menggunakan metode *Decision Tree C4.5* dengan akurasi terbaik yang dihasilkan sebesar 80%.

Model klasifikasi seperti *Support Vector Machine, Random Forest,* dan Regresi Logistik merupakan beberapa model yang dapat digunakan dan tidak jarang ditemui untuk mengatasi masalah klasifikasi. Metode-metode berikut dapat menghasilkan performa yang cukup baik dalam meningkatkan hasil akurasi pada klasifikasi binary, seperti pada metode SVM dapat secara efisien melakukan klasifikasi linier ataupun non-linier menggunakan trik kernel [12]. Metode Random Forest dapat menangani kelas yang tidak terwakili dengan baik, serta memiliki efisiensi dari komputasi karena pencarian variabel acak dibatasi [12]. Metode Regresi Logistik dapat memahami hubungan antara variabel dependen dan variabel independen dengan memperkirakan probabilitas menggunakan perasamaan regresi logistik sehingga dapat memprediksi klasifikasi lebih akurat [7]. Oleh karena itu, penelitian klasifikasi berita hoax akan dilakukan terhadap tiga metode yaitu Random Forest, Regresi Logistik, dan *Support Vector Machine* (SVM).

1.2 Rumusan masalah

Berdasarkan latar belakang penelitian tersebut, maka perumusan masalah pada penelitain ini yaitu:

1. Bagaimana analisa perbandingan metode SVM, Random Forest dan Regresi Logistik pada klasifikasi berita *hoax*?
2. Apa saja informasi (*insight*) yang dapat diperoleh dari setiap kelas?

1.3 Batasan masalah

Batasan masalah pada penelitian ini sebagai berikut

1. Data yang digunakan dalam penelitan didapatkan melalui situs *turnbackhoax.id* yang telah memvalidasi berita *hoax* dan mengelompokan berita *hoax* pada priode tahun 2015-2020.
2. Data yang diklasifikasi merupakan data berbahasa Indonesia.
3. *Machine* learning yang digunakan dalam penelitian ini adalah *Random Forest Classifier*, Regresi logistik, dan *Support Vector Machine.*

1.4 Tujuan penelitian

1. Tujuan penelitian ini untuk mengetahui metode terbaik dalam mengklasifikasikan berita *hoax* dengan membandingkan tiga metode klasifikasi antara SVM, Random Forest dan Regresi logistik.
2. Melihat informasi apa saja yang dapat diperoleh dari setiap kelas.

1.5 Manfaat Penelitian

Melalui penelitian ini diharapkan mampu menjadi referensi dalam mendeteksi berita *hoax* sebagai salah satu fenomena yang berpotensi merusak jurnalisme. Selain itu, dapat ikut serta dalam ilmu pengetahuan sebegai referensi mengenai klasifikasi berita hoax pada beberapa *machine learning* dan dapat digunakan untuk pengembangan permodelan suatu system yang dapat mendeteksi nilai *hoax* sebuah berita.

BAB II
LANDASAN TEORI

Pada bab ini penulis akan menjelaskan definisi dan teori-teori yang digunakan sebagai landasan penelitian.

2.1 *Preprocessing*

Banyak data narasi berita yang diperoleh dari turnbackhoax.id masih menggunakan kata yang tidak terstruktur seperti menggunakan emotikon, symbol, singkatan, angka, dan link sehingga perlu digunakan teknik *preprocessing*. Berikut *preprocessing* yang dilakukan pada penelitian ini:

1. *Case Folding*

Case folding dibutuhkan untuk menyamakan bentuk pada kata, contohnya seperti yang dilakukan pada data narasi yang digunakan pada penelitian ini yaitu mengubah semua kata menjadi huruf kecil atau disebut *lowercase* [5].

2. Menghapus Simbol, Angka, dan Emoji

Berita yang ditulis terdapat banyak angka, emoji, dan tanda baca namun ketiganya tidak penting untuk diolah, sehingga perlu dihapus untuk mempermudah dalam pengolahan data [5].

3. Tokenisasi

Tokenisasi bertujuan untuk membagi kalimat menjadi beberapa bagian seperti kata-kata, frasa atau elemen bermakna yang lain [5].

4. *Lemmatization*

Lemmatization adalah proses transformasi yang dilakukan untuk menormalisasi suatu kata dan mengubah kata kedalam bentuk kata dasar, suatu proses yang bertujuan untuk menormalkan kata berdasarkan bentuk dasarnya yaitu bentuk lemma [7].

5. *Stopword*

Stopwrod merupakan kata-kata umum yang sering muncul yang tidak memiliki makna atau tidak memiliki pengaruh yang signifikan dalam kalimat, sehingga penghapusan stopword diperlukan dimana penulis melakukan proses

penghapusan berdasarkan daftar kata-kata stopword pada dataset. Contohnya adalah “yang”, “di”, “dari” dan sebagainya [5].

2.2 *Vector Space* Model

Umumnya model klasifikasi dan beberapa algoritma tidak dapat langsung memproses teks dokumen dalam bentuk aslinya, sedangkan komputer hanya mampu mengolah dokumen yang terstruktur dan dalam bentuk tabular. Oleh karena itu, dalam tahapan *preprocessing*, dokumen harus diubah supaya lebih mudah direpresentasikan [6]. Dokumen direpresentasikan sebagai fitur berbentuk vektor.

Pada tahap ini, penulis menggunakan fitur TF-IDF (*Term Frequenct Inverse Document Frequency*). Algoritma TF-IDF merupakan penggabungan dari dua buah aturan dalam pembentukan fiturnya, dimana masing-masing aturan menangkap salah satu dari dua intuisi yang disajikan berikut [7]:

1. *Term Frequency* (*tf*), didefinisikan sebagai banyaknya kata ‘x’ yang muncul pada ‘d’ dokumen, frekuensi dihitung sebagai hasil bagi banyaknya kata terhadap keseluruhan kata di dokumen tersebut atau secara matematis dapat dibentuk sebagai:

$$tf_{x,d} = count(x, d) \tag{2.1}$$

Dimana:

x : kata
d : dokumen
$tf(x, d)$: frekuensi banyaknya x yang muncul di d

2. *Inverse Document Frequency* (IDF), didefinisikan sebagai hasil bagi N/d_x dimana N merupakan banyaknya doumen yang tersedia dari suatu topik dan d_x merupakan banyaknya dokumen yang mengandung ‘x’ kata. Semakin sedikit dokumen yang tersedia, semakin besar bobot dari fiturnya

$$idf_x = \log\left(\frac{N}{d_x}\right) \tag{2.2}$$

Dimana:

N : total dokumen

d_x : dokumen yang mengandung x kata

idf_x : frekuensi banyaknya x kata yang muncul di dokumen d

Oleh karenanya algoritma TF-IDF merupakan kombinasi kedua intuisi tf dan idf, dan perhitungan bobot untuk fitur TF-IDF didefinisikan sebagai:

$$w_{x,d} = tf_{x,d} \times idf_x \tag{2.3}$$

Dimana:

$w_{x,d}$: besar bobot TF-IDF

2.3 *Machine Learning*

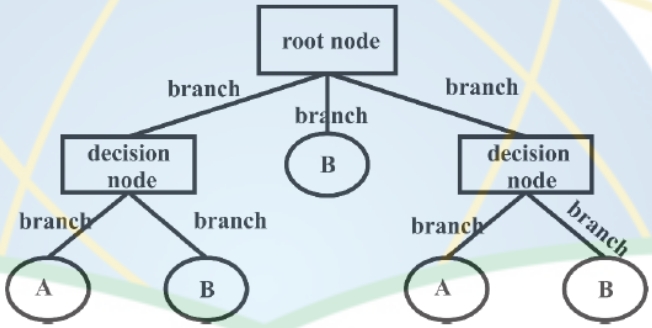
Pembelajaran mesin (*Machine Learning*) merupakan bagian dari *Artificial Intelligence* (AI) atau kecerdasan buatan yang digunakan untuk mengoptimalkan kinerja kerja manusia dalam melakukan sesuatu secara otomatis agar lebih efisien [8]. Pada *machine learning* terdapat proses pelatihan dan pembelajaran. Ada beberapa Teknik yang dimiliki oleh *machine learning*, namum secara luas *machine learning* memilki dua teknik dasar belajar, yaitu *supervised* dan *unsupervised learning. Supervised learning* merupakan teknik yang diterapkan pada pembelajaran mesin yang bisa menerima informasi yang sudah ada pada data dengan memberikan label tertentu yang digunakan oleh mesin untuk melakukan pengelompokan objek berdasarkan kriteria tertentu. *Unsupervised learning* merupakan teknik yang diterapkan pada pembelajaran mesin pada data yang tidak memiliki informasi yang bisa diterapkan, dalam teknik ini mesin melakukan penerkaan untuk membantu menemukan struktur atau pola tersembunyi pada data yang telah dipelajari dalam training.

2.4 *Decision Tree*

Pohon Keputusan merupakan sekumpulan simpul keputusan (*decision nodes*) berbentuk pohon yang menotasikan tes atribut, dihubungkan oleh cabang (*branches*) yang merepresentasikan hasil keluaran atribut, memanjang ke bawah

dari simpul akar (*root node*) hingga berakhir di simpul daun (*leaf node*) yang menggambarkan label kelas. Dimulai pada simpul akar, di mana menurut aturan diletakkan di paling atas dari struktur pohon, atribut diuji pada simpul keputusan (*decision nodes*), dengan setiap hasil memungkinkan untuk menghasilkan cabang. Setiap cabang kemudian mengarah ke simpul keputusan lain atau ke simpul daun terakhir (*terminal node/leaf node*) [23].

Berikut disajikan bentuk umum pohon keputusan berdasarkan uraian pada paragraf sebelumnya mengenai definisi dari pohon keputusan:



Gambar 2.1 Bentuk Umum Pohon Keputusan

Gambar 2.1 merupakan bentuk pohon umum dari pohon keputusan. Untuk setiap simpul keputusan, nantinya akan merepresentasikan sebuah tes pada atribut. Setiap cabang yang terbentuk merepresentasikan hasil keluaran atribut, "*A*" dan "*B*" pada simpul daun sebagai contoh merepresentasikan hasil klasifikasi dari kelas atributnya apakah atribut dilabeli kelas "*A*" atau kelas "*B*".

Algoritma pohon keputusan mulai berkembang pada akhir tahun 1970 dan awal tahun 1980, algoritma pohon keputusan *CHAID* pertama kali dipaparkan oleh Kass pada tahun 1980 [24], selanjutnya Quinlan (1983) memaparkan algoritma ID3 [25] lalu mengembangkan metode ini menjadi algoritma C4.5 (1993) [26] dan C5.0 (1998) [23]. Breiman dan ilmuwan lain juga memaparkan metode pohon keputusan lain yang diberi nama *CART* atau *Classification and Regression Tree* di tahun 1984 [27].

2.5 *Random Forest Classifier*

Random Forest merupakan pengembangan dari metode *classification and regrestion tree* (CART) untuk meningkatkan akurasi pada klasifikasi dengan cara mengkombinasikan metode klasifikasi [9]. *Random forest* merupakan salah satu metode *supervised learning* dimana kita sudah memiliki variable masukan (*input*) dan keluaran (*output*). Tujuannya memperkirakan fungsi pemetaannya, sehingga kita memiliki variabel inputan baru sehingga mesin dapat memprediksi output dari inputan tersebut [10].

Dalam *random forest* terdiri dari sekumpulan pohon keputusan sehingga dari sekumpulan pohon keputusan terbentuk hutan (*forest*) kemudian digunakan untuk mengklasifikasi data ke suatu kelas pada kumpulan pohon tersebut. Pada data training berukuran n dan 0 peubah penjelas , random forest dilakukan dengan cara [11]:

1. Bentuk dataset sebanyak *n* menggunakan teknik *boostrap.*
2. Bentuk sebanyak *n* pohon keputusan Tb pada masing-masing dataset menggunakan algoritma *CART.*
3. Bangun pohon Random Forest notasikan T_b dengan mengulang langkah-langkah berikut secara rekursif untuk setiap terminal node (daun) yang terbentuk hingga ukuran minimum simpul n_{min} tercapai.
   i. Pilih variable m secara acak dari variabel p.
   ii. Pilih variable terbaik diantara m.
   iii. Bagi simpul menjadi dua simpul yang lebih kecil.
4. Bentuk pohon *ensemble* dari { T_b} sebanyak *P* kali sehingga terbentuk *P* banyak pohon keputusan .
5. Misalkan $\hat{D}_b(x)$ adalah kelas hasil prediksi dari *b* pohon pada *Random Forest* yang terbentuk, maka lakukan $\hat{D}_{RfB} = majority\ vote\ \{\hat{D}_b(x)\}_{1B}$.

Dalam hal ini metode *random forest* cendrung lebih baik dengan metode sebelumnya, *random forest* juga dapat menangani kelas yang tidak terwakili dengan baik, serta memiliki efisiensi dari komputasi karena pencarian variabel acak dibatasi [9].

2.6 Regresi Logistik

Regresi logistik merupakan teknik analisis data dalam statistika yang bertujuan untuk mengetahui hubungan antar variabel dimana variabel responnya bersifat kategorik, baik nominal maupun ordinal dengan variabel penjelasnya dapat bersifat kategorik atau kontinu [13]. Regresi logistik dapat digunakan untuk menyelesaikan masalah klasifikasi dengan cara menghitung probabilitas. Klasifikasi yang berasal dari variabel respon biner dilakukan dengan cara menentukan titik potong. Titik potong yang dapat digunakan sebesar 0.5. Klasifikasi yang berada dari variabel respon biner dilakukan dengan menggunakan probabilitas dengan ketentuan berikut [14]:

$$Kategori = \begin{cases} 1, & P(Y=1) \geq 0.5 \\ 0, & untuk\ lainnya \end{cases}$$

Metode regresi logistik memiliki Teknik dan prosedur yang tidak jauh berbeda dengan metode regresi linear. Jika prosedur dalam estimasi parameter menggunakan *Ordinary Least Squares* (OLS), sedangkan estimasi parameter dalam regresi logistik menggunakan *Maximum Likelihood Estimator* (MLE). Untuk mencari persamaan logistiknya model yang digunakan adalah [14]:

$$P(Y=0) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}{1 + \exp\left(\beta_o \sum_{j=1}^{p} \beta_j x_j\right)} \quad (2.6)$$

$P(Y=1)$: peluang teradinya variabel respon

x_j : variabel predictor ke j

j : banyaknya variabel predictor

β : Intersep / koefisien regresi untuk setiap variabel predictor

Untuk mempermudah Interpretasi dan penduga parameter, peluang pada persamaan diatas.

Dilakukan transformasi logit sehingga didapatkan fungsi logit sebagai berikut

Dari persamaan (2.6) diperoleh $P(Y=1)$ sebagai berikut:

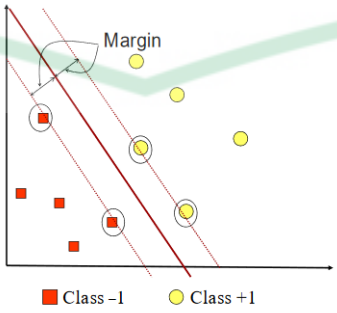
$$P(Y=1) = \frac{1}{1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)} \quad (2.7)$$

Model regresi logistik yang variabel responnya terdiri dari dua kategori di sebut dengan regresi logistik biner. Dua kategon tersebut yaitu sukses atau gagal, dengan menetapkan $P(Y=1)$ sebagai sukses dan $P(Y=0)$ sebagai gagal.

2.7 *Support Vector Machine*

Secara konseptual *Support Vector Machine* (SVM) adalah mentransformasikan data keruang yang berdimensi lebih tinggi dan menemukan *hyperline* terbaik. *Hyperline* merpakan fungsi pemisah yang terletak diantara dua macam set obect (*pattern*) dari dua kelas. Untuk menentukan *hyperline* terbaik yaitu dengan mengukur margin dari *hyperline* tersebut. *Margin* adalah jarak antara *hyperplane* dengan pattern terdekat dari masing-masing kelas. Pattern yang paling dekat dengan *hyperplane* ini disebut *support vector* [15].

Misal data latih dinyatakan dala (x_i, y_i) dimana $i = 1,2,3,\ldots,n$. $x_i = [x_{i1}, x_{i2}, \ldots, x_{ij}]$ adalah vector baris dari fitur ke-j dan y_i adalah label dari x_i yang didefinisikan sebagai $y_i \in \{+1, -1\}$. Diasumsikan kedua kelas dapat dipisah secara linier oleh *hyperplane*. Pada gambar 2.2, *hyperplane* ditunukan oleh garis lurus berwarna merah. Data yang berbentuk lingkaran dan berada diatas *hyperplane* adalah kelas +1 dan yang berbentuk persegi dan berada dibawah *hyperplane* adalah kelas -1.



Gambar 2.2. Contoh *Hyperplane* Dua Dimensi

Persamaan *hyperplane* didefinisikan sebagai berikut:

$$f(x) = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{2.8}$$

Dengan:

$\boldsymbol{w}$: parameter bobot

$\boldsymbol{x}$: Vector imput

b : bias

Hyperline terbaik merupakan *hyperplane* yang letaknya diantara dua set obyek dari dua kelas. Untuk itu, perlu menemukan *hyperplane* terbaik dengan mendapatkan nilai margin terbesar. Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplan*e dan titik terdekatnya. Dari persamaan 2.8 tersebut maka *hyperplane* pendukung dari kelas pattern yang memenuhi kelas -1 dapat dinyatakan dengan persamaan $\boldsymbol{w}\cdot\boldsymbol{xi}+b=-1$ dan pattern yang memenuhi kelas +1 dapat dinyatakan dengan persamaan $\boldsymbol{w}\cdot\boldsymbol{xi}+b=1$.

Sehingga margin dapat dihitung dengan mencari nilai tengah dari kedua jarak kelas seperti pada persamaan 2.9.

$$Margin = \frac{1}{2}\left(\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, x^{+}\right) - \left(\frac{w}{\|w\|}, x^{-}\right)\right) \tag{2.9}$$

$$= \frac{1}{2}\left(\left(\frac{1-b}{\|\boldsymbol{w}\|}\right) - \left(\frac{-1-b}{\|\boldsymbol{w}\|}\right)\right)$$

$$= \frac{1}{2}\left(\frac{1+1}{\|\boldsymbol{w}\|}\right)$$

$$= \frac{1}{\|\boldsymbol{w}\|}, \|\boldsymbol{w}\| \neq 0$$

Diberikan kasus sederhana, dimana digunakan dua kelas -1 dan 1, dengan sampel $X = \{x_i, y_i\}$ dimana $y_i = 1$ jika $x_i \in C_1$ dan $y_i = -1$ jika $x_i \in C_2$ kita akan mencari nilai $\boldsymbol{w}$ yang memenuhi:

$$\boldsymbol{w}\cdot\boldsymbol{x_i} + b \leq -1, jika\ y_i = -1 \tag{2.10}$$

$$\boldsymbol{w}\cdot\boldsymbol{x_i} + b \geq +1, jika\ y_i = +1 \tag{2.11}$$

Persamaan tersebut dapat dibentuk menjadi:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1 > 0, \forall\, 1 \le i \le n, i \in N \qquad (2.12)$$

Memaksimalkan nilai margin ekuivalen dengan meminimumkan $\|\boldsymbol{w}\|^2$. Maka pencarian hyperline terbaik dengan margin terbesar dapat ditulis menjadi masalah optimasi SVM sebagai berikut:

$$\max margin = \min \frac{1}{2}\, \|\boldsymbol{w}\|^2 \qquad (2.13)$$

Dengan kendala:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1 > 0, \forall\, 1 \le i \le n, i \in N$$

Masalah ini dapat diselesaikan dengan mengubah persamaan ke dalam fungsi *lagrange*:

$$\min L_p(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1], \qquad (2.14)$$

Dimana:

L_p: fungsi *lagrange* (*primal problem*)

α_i: nilai dari koefisien *lagrange*,$\alpha_i \ge 0$ dengan $i = 1,2, \dots, n$.

Fungsi L_p diminimumkan terhadap $\boldsymbol{w}$, b dan dimaksimalkan pada variabel α, sehingga akan dicari turunan pertama dari fungsi L_p terhadap $\boldsymbol{w}$ dan b, sebagai berikut [16]:

1. Turunan pertama Lp terhadap w:

$$\frac{\partial}{\partial \boldsymbol{w}} L_p(\boldsymbol{w}, b, \alpha) = 0$$

Maka diperoleh:

$$\min L_p(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b)] + \sum_{i=1}^{n} \alpha_i\,,$$

$$\frac{\partial}{\partial \boldsymbol{w}} L_p(\boldsymbol{w}, b, \alpha) = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i\, y_i x_i$$

$$\Leftrightarrow \qquad 0 = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i\, y_i x_i$$

$$\Leftrightarrow \qquad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \, y_i x_i \tag{2.15}$$

2. Turunan pertama L_p terhadap b:

$$\frac{\partial}{\partial \boldsymbol{w}} L_p(\boldsymbol{w}, b, \alpha) = 0$$

Maka diperoleh:

$$\min L_p(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b)] + \sum_{i=1}^{n} \alpha_i \,,$$

$$\frac{\partial}{\partial \boldsymbol{w}} L_p(\boldsymbol{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i \, y_i x_i$$

$$\Leftrightarrow \qquad 0 = \sum_{i=1}^{n} \alpha_i \, y_i x_i \tag{2.16}$$

Persamaan 2.15 dan 2.16 merupakan optimalisasi fungsi *lagrangian.* Apabila *lagrange* L_p (*primal problem*) memiliki solusi optimal maka *lagrange* L_D (*dual problem*) juga akan memiliki solusi optimal yang sama [16]. Untuk mendapatkkan L_D, Formula *langrange Lp* (*primal problem*) diubah menjadi *LD* (*dual problem*) dengan mensubtitusi persamaan 2.15 dan 2.16 pada persamaan 2.14 sehingga diperoleh:

$$max \; L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j \, (x_i, x_j), \tag{2.17}$$

Dengan kendala:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \;\; \alpha_i \geq 0 \tag{2.18}$$

Nilai α_i diperoleh dari hasil perhitungan substitusi kendala pada persamaan (3.4). Nilai α_i akan digunakan untuk menemukan nilai $\boldsymbol{w}$. Setiap titik data selalu terjadi $\alpha_i = 0$. Titik-titik data dimana $\alpha_i = 0$ tidak akan muncul dalam perhitungan mencari nilai $\boldsymbol{w}$ sehingga tidak berperan dalam memprediksi data baru. Data lain dimana $\alpha_i > 0$ disebut support vector [20].

Prinsip kerja dari model SVM adalah menentukan *hyperplane* terbaik dari suatu *dataset* untuk dipisahkan berdasarkan kelasnya. Dalam beberapa kasus yang sering terjadi di dunia nyata, data yang dijumpai merupakan data yang tidak dipisahkan secara linear, yaitu saat tidak ada sebuah garis atau bidang yang dapat memisahkan antar kelas pada data [17]. Oleh karena itu, untuk mengatasi hal tersebut dibutuhkan *soft margin classifier* yaitu dengan mengklasifikasikan sebagian besar data benar dan memberikan kemungkinan pada model untuk terjadi kesalahan saat mengklasifikasi pada beberapa titik di sekitar bidang pemisah [18]. Dalam hal ini maka batas pemisah harus diubah agar lebih fleksibel dengan cara menambahkan variabel *slack* ξ dimana $\xi > 0$ pada setiap batas pemisah, sehingga batas pemisahnya berubah menjadi $x_i + b \geq 1 - \xi$ untuk kelas +1 dan $x_i + b \leq -1 + \xi$ untuk kelas -1 [19]. Dengan adanya penambahan variabel slack yang bertujuan untuk menunjukkan ketelitian pemisahan yang memungkinkan suatu titik berada pada kondisi misklasifikasi, pencarian *hyperplane* terbaik dirumuskan sebagai berikut:

$$\max margin = \min \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \qquad (2.19)$$

dimana parameter C merupakan variable yang bervariasi bergantung pada tujuan pengoptimalan [20]. Ketika nilai C semakin besar maka margin lebih ketat akan diperoleh, dalam hal ini dapat meminimalisisr kesalahan klasifikasi. Ketika nilai C semakin kecil artinya lebih banyak kesalahan klasifikasi yang dilakukan, dalam hal ini menekankan pada pemaksimalan margin yang menjadi tujuan utama dari metode SVM [21]. Perubahan juga terjadi pada fungsi *lagrange primal*, yaitu:

$$\min L_p(\boldsymbol{w}, b, \alpha) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n} \mu_i \xi_i \qquad (2.20)$$

pencarian *hyperplane* terbaik dilakukan dengan cara yang hampir sama tetapi dengan nilai α_i adalah $0 \geq \alpha_i \geq C$.

2.8 *Synthetic Minority Over Sampling Technique*

Synthetic Minority Over Sampling Technique atau *SMOTE* merupakan sebuah pendekatan untuk mengkonstruksi sebuah model klasifikasi dalam

menangani ketidakseimbangan data*set*. Sebuah data*set* dikatakan tidak seimbang jika kategori dari klasifikasi tidak terwakili secara merata. *Smote* mengusulkan pendekatan *over-sampling* dimana kelas minoritas melakukan *over- sample* dengan membuat data buatan (sintetis) dibanding melakukan *over-sample* dengan penggantian. Pendekatan *smote* terinspirasi oleh sebuah teknik yang menciptakan data pelatihan tambahan dengan melakukan operasi tertentu pada data nyata. *Smote* menghasilkan contoh sintetik yang lebih beroperasi di ruang fitur dibanding ruang data. *Over-sample* kelas minoritas dilakukan dengan mengambil masing-masing sampel kelas minoritas dan memperkenalkan contoh-contoh sintetis di sepanjang garis segmen yang bergabung dengan salah satu atau semua tetangga terdekat *k* kelas minoritas. Chawla, Bowyer, Hall dan Kagelmeyer [32] mengemukakan hasil bahwa pendekatan *smote* pada data yang tidak seimbang dapat menaikkan akurasi dari model klasifikasi untuk kelas minoritas, *smote* juga memberikan pendekatan baru dalam melakukan *over-sample*, kombinasi antara *smote* dan *under-sampling* bekerja lebih baik dibandingkan dengan *under-sampling* biasa. *Smote* telah diuji pada data*set* yang beragam, dengan berbagai tingkat ketidakseimbangan dan jumlah yang bervariasi pada *set* pelatihan, sehingga menyediakan *tesbed* yang beragam [32].

2.9 Optimasi parameter

Grid Search Cross Validation (*Grid Search* CV) merupakan salah satu proses untuk melakukan pemilihan hyperparameter terbaik dalam klasifikasi sehingga dapat memprediksi data yang tidak diketahui secara akurat. *Hyperparameter* yang diperoleh merupakan parameter terbaik yang akan dimasukkan pada model [21]. *Grid Search CV* melakukan kombinasi dari *hyperparameter* yang telah ditentukan dan menghitung rata-rata nilai *cross validation* dari setiap kombinasi yang akan dimasukan pada model [15].

Pada *Grid Search Cross Validation* dilakukan kombinasi nilai *hyperparameter* yang dimasukkan kemudian *hyperparamater* yang memiliki parameter terbaik akan digunakan untuk melatih model pada seluruh data. Pada *Support Vector Machine* (SVM) terdapat tiga parameter yang perlu dioptimalkan yaitu C, *gamma* dan *degree*. Pada Regresi Logistik terdapat teknik regulasi

dengan parameter yang perlu dioptimalkan yaitu antara nilai *lasso* dan *ridge*. Pada *Random Forest* parameter yang perlu dioptimalkan yaitu *N_estimator, max_depth, max_feature.*

2.10 *Evaluasi Model*

Evaluasi model dilakukan untuk mengetahui seberapa baik model dapat melakukan klasifikasi suatu kelas. Salah satu cara yang sering digunakan dalam melakukan evaluasi model adalah dengan menggunakan *confusion matrix. Confusion matrix* merupakan table yang menyatakan berapa banyak data uji yang benar dan salah diklasifikasikan seperti pada Tabel 2.1:

Tabel 2.1. Confusion Matrix.

Confusion Matrix	Prediksi	
	Negatif	Positif
Negatif	TN	FN
Positif	FP	TP

Dari tabel 2.1 diatas, evaluasi dan validasi hasil dari model dapat dihitung menggunakan rumus akurasi seperti pada persamaan berikut [22]:

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (2.21)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.22)$$

Dimana :

TP (*True Positive*): Jumlah data *positif* yang terklasifikasi benar.

TN (*True Negative*): Jumlah data *negative* yang terklasifikasi benar.

FP (*False Positive*): Jumlah data *positif* yang terklasifikasi salah.

FN (*False Negative*): Jumlah data *negative* yang terklasifikasi salah.

BAB III
METODOLOGI PENELITIAN

Pembahasan pada bab ini mengenai metode-metode yang digunakan dalam penelitian secara teori dan contoh penerapannya. Pada bab ini juga akan dijelaskan bagaimana alur penelitian.

3.1 Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder berupa data berita berbahasa Indonesia yang terdapat pada situs *turnbackhoax* yang memvalidasi kebenaran berita-berita yang beredar di masyarakat. Data yang diperoleh dengan menggunakan teknik *web scraping* ini berjumlah 4.551 data berita dari tahun 2015 sampai dengan tahun 2020.

Pada saat proses scraping, peneliti menggunakan *package Selenium* dan *browser chrome*. Berita yang didapat dari hasil scraping urut berdasarkan waktu. Data berisikan tiga kolom, diantaranya adalah tanggal yang merupakan tanggal pada saat berita dikeluarkan, judul yang berisikan judul dari berita, dan narasi yang berisikan isi dari berita pada web *turnbackhoax*. Data disimpan dalam bentuk *Comma Separated Values* (CSV). Berikut adalah beberapa data awal dari hasil scraping:

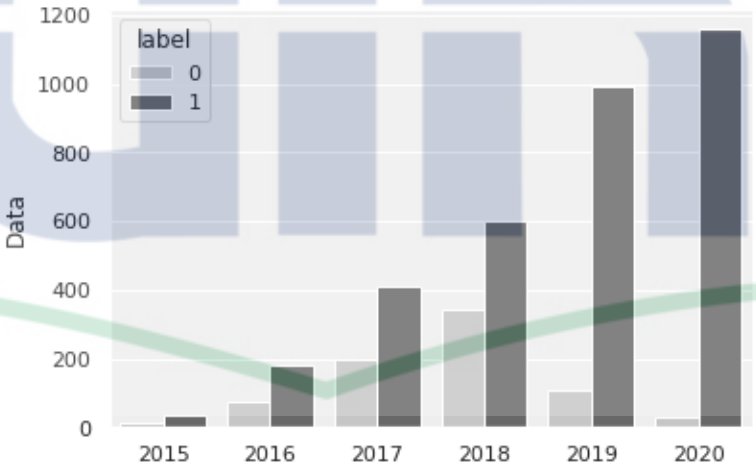
Tabel 3.1. Data Awal Berita Turnbackhoax

Tanggal	Judul	Narasi
03-Dec-19	Pergantian Kartu Keluarga Warga DKI Jakarta	Dinas Dukcapil Jakarta tidak mewajibkan warga Jakarta mengganti KK lamanya dengan tenggat waktu. Imbauan untuk segera memutakhirkan data memang terus didorong pemerintah, terutama input data status pernikahan, golongan darah, dan susunan anggota keluarga.

21-May-20	Koreksi Post Empat Ton Telur Bansos Pemprov Jabar Membusuk	Berkaitan juga dengan koreksi sebelumnya di: "[BERITA] Perkembangan Informasi Empat Ton Telur Bansos Pemprov Jabar Membusuk di Gudang Penyimpanan Bulog Garut" https://bit.ly/2WWv877 Tim Redaksi dan Pemeriksa Fakta MAFINDO memohon maaf atas ketidaknyamanan yang ditimbulkan. Selengkapnya di bagian PENJELASAN dan REFERENSI.

3.2 Pelabelan Berita

Setelah memperoleh data melalui *web scraping*, Langkah selanjutnya dilakukan analisis dengan memberikan label pada setiap narasi berita. Kelas-kelas yang digunakan yaitu *hoax* dan *non-hoax* dan label yang digunakan yaitu 0 untuk berita *non-hoax* dan 1 untuk berita *hoax*. Berita yang dilabeli sudah divalidasi kebenarannya melalui website turnbackhoax. Pelabelan dilakukan dengan tujuan untuk memberikan pebelajaran pada model.



Gambar 3.1. Pelabelan Berita

Hasil pelabelan yang dilakukan adalah seperti pada Gambar 3.1, dapat diketahui bahwa pada setiap tahun jumlah berita *hoax* lebih banyak dari berita *non-hoax*. Selain itu, jumlah berita *hoax* paling banyak berada pada tahun 2020

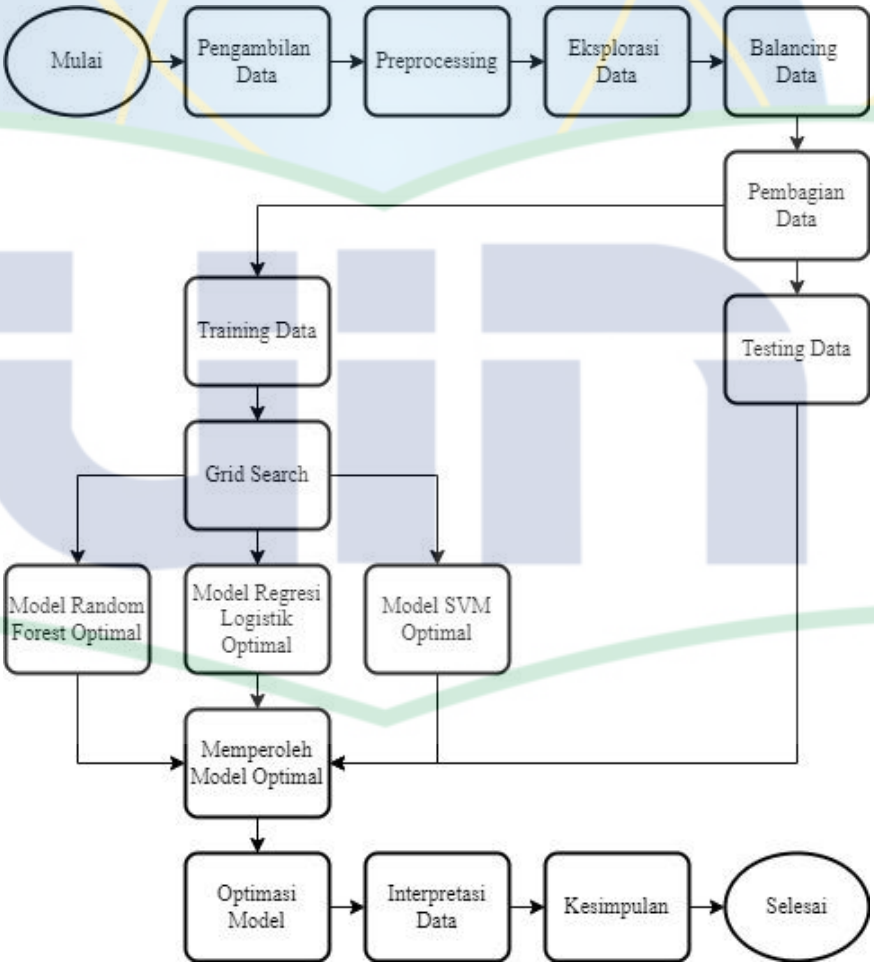
sebanyak 1.159 Berita. Tahap lanjut yang dilakukan setelah data diubah menjadi vektor adalah membagi data menjadi data *training* dan data *testing*. Data *training* atau data *latih* yang digunakan untuk membangun model dibentuk dengan ukuran 70% dan 30% sisanya digunakan sebagai data testing untuk menguji performa model yang sudah dilatih.

3.4 Alur Penelitian

Penulis melakukan proses pengolahan data menggunakan bahasa pemograman *Python 3* menggunakan *software Anaconda* dengan bantuan beberapa modul yaitu *sklearn, nltk, pandas, bs4 (beautifulsoup), imblearn, seaborn, matplotlib, numpy,* dan. Berdasarkan alur penelitian pada Gambar 3.2. Langkah awal yang dilakukan adalah mengambil data dengan cara scraping web *Turnbackhoax,* mengambil data isi narasi berita dari tahun 2015 sampai 2020, diperoleh sebanyak 4.551 data. Setelah data didapat dan disimpan ke CSV. Memberikan kelas untuk setiap keseluruhan dataset secara manual, memberi label 0 untuk berita *non-hoax* dan label 1 untuk *hoax.* Kemudian melakukan pembersihan data atau preprocessing, langkah preprocessing yang dilakukan di antaranya mengubah kata slang/singkatan, *lemmatization*, case-folding, dan menghapus stopword dan simbol. Data harus dibersihkan terlebih dahulu karena data yang didapat merupakan data mentah yang belum siap olah, dan mengandung *noise*, artinya masih ada karakter selain huruf yang harus dihapus, dan banyak kata-kata singkatan yang perlu diubah. Kata yang sering muncul dan tidak penting juga perlu dihapus supaya mudah untuk mengolah data dengan hanya melihat kata yang penting saja. Selanjutnya, model klasifikasi dan beberapa algoritma tidak dapat langsung memproses teks dokumen dalam bentuk aslinya. Oleh karena itu, dalam tahapan *preprocessing*, dokumen harus diubah supaya lebih direpresentasikan. Dokumen direpresentasikan sebagai fitur berbentuk vector disebut VSM (*Vector Space Model*).

Setelah data bersih dan dib, dilakukan proses eksplorasi data dengan melihat wordcloud dari hasil data yang telah dilakukan preprocessing. Jumlah data klasifikasi tidak seimbang, dimana lebih banyak data *hoax* dibanding data *non-*

hoax, maka dilakukan resampling data untuk mengatasi data tidak seimbang, dengan cara *Synthetic Minority Over Sampling Technique* atau *SMOTE*. Data dibagi menjadi *training* dan *testing*, sebanyak 70% untuk *training* dan 30% untuk *testing*. Kemudian, data dimasukkan ke model klasifikasi biner yaitu Random Forest, Regresi Logistik, dan SVM. Kemudian, melakukan pembentukan model dengan optimasi parameter dengan menggunakan *Grid Search CV* untuk mendapatkan parameter terbaik pada model sehingga menghasilkan model dengan hasil klasifikasi paling baik. Kemudian, melakukan evaluasi model dengan confusion matrix. Lalu interpretasi data dengan visualisai data menggunakan *wordcloud* dan wordlink dari setiap kelas yang telah diklasifikasi. Alur penelitian diberikan seperti pada Gambar 3.2 berikut:



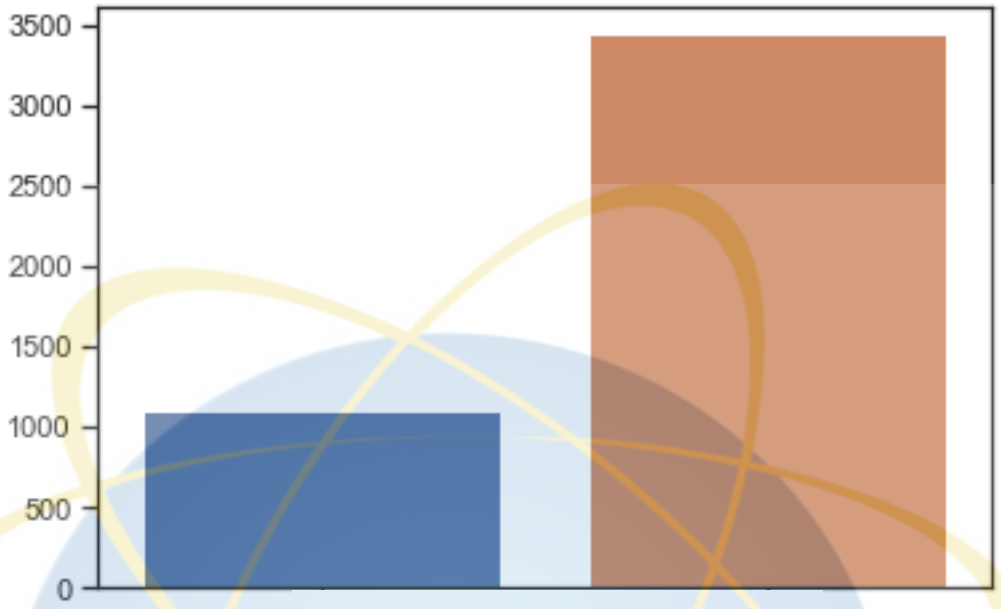
Gambar 3.2. Alur Penelitian

BAB IV
HASIL DAN PEMBAHASAN

Pada bab ini penulis akan menjelaskan hasil dari pengambilan data, data yang telah di*preprocessing* dan wordcloud dari setiap kelas berdasarkan labeling. Kemudian menjelaskan grafik training data serta menunjukan hasil uji coba beberapa model *machine learning* untuk menentukan model mana yang paling optimal. Lalu akan ditunjukan pula evaluasi model dengan *confusion matrix*.

4.1 Hasil Preprocessing dan Text Analytics

Data yang diperoleh melalui hasil *scraping* perlu dibersihkan terlebih dahulu sebelum diolah oleh mesin. Proses ini dinamakan *preprocessing*, lalu dilakukan pelabelan data. Gambar 4.1 memperlihatkan jumlah label klasifikasi data *Turnbackhoax*. Terlihat bahwa perbandingan antara jumlah label *hoax* dan *non-hoax* terdiri dari 3442 berita *hoax* dan 1106 berita *non-hoax*.



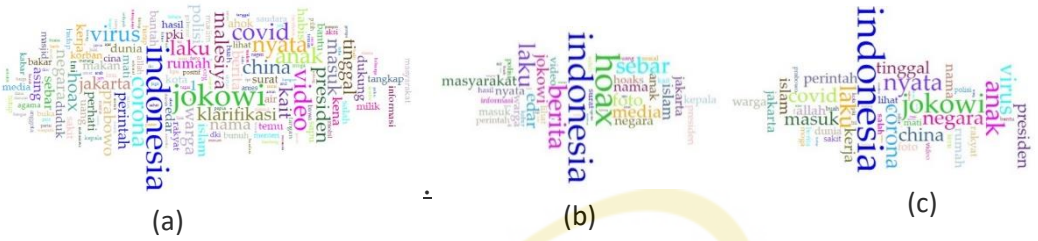
Gambar 4.1. Jumlah Label Tiap kelas

Adapun langkah *preprocessing* yang dilakukan antaranya adalah *case folding*, menghapus symbol, angka dan emotikon, *lemmatization*, penghapusan *stopword* serta *tokenizer*. Pada Table 4.1 merupakan hasil dari proses *preprocessing* dan pelabelan. Hasil berita yang sudah melewati tahap preprocessing dimasukan ke kolom "Berita preprocessing". Selanjutnya setiap kelas pada data berita divisualisasikan dengan wordcloud. Tujuannya adalah untuk melihat kata apa saja yang muncul pada setiap kelas.

Tabel 4.1. Hasil Preprocessing dan Pelabelan

Label	Berita	Berita Preprocessing
1	Profesor Dr Tasuku Honjo mengatakan bahwa virus korona itu tidak alami. Cina telah membuatnya Profesor Fisiologi Kedokteran Jepang, Profesor Dr Tasuku Honjo, menciptakan sensasi di	Profesor tasuku honjo virus korona alami cina profesor fisiologi dokter jepang profesor tasuku honjo cipta sensasi media virus korona alami alami pengaruh dunia sesuai sifat suhu beda negara alami dampak buruk negara milik suhu cina

	depan media hari ini dengan mengatakan bahwa virus korona itu tidak alami.Jika itu alami, itu tidak akan mempengaruhi seluruh dunia seperti ini. Karena, sesuai sifatnya, suhu berbeda di berbagai negara. Jika itu alami, itu akan berdampak buruk hanya pada negara-negara yang memiliki suhu yang sama dengan Cina.	
0	Charta Politika Tidak Pernah Melakukan Survei Nasional dengan Simulasi Jokowi - KH. Maruf Amin dan Prabowo Subianto - Sandiaga Uno Adapun data / gambar mengenai hasil Survei Nasional mengatasnamakan Charta Politika yang beredar di beberapa WhatsApp Group adalah HOAX dan kami akan memproses secara hukum	charta politika laku survei nasional simulasi jokowi maruf amin prabowo subianto sandiaga uno data gambar kena hasil survei nasional mengatasnamakan charta politika edar whatsap grup hoax proses hukum

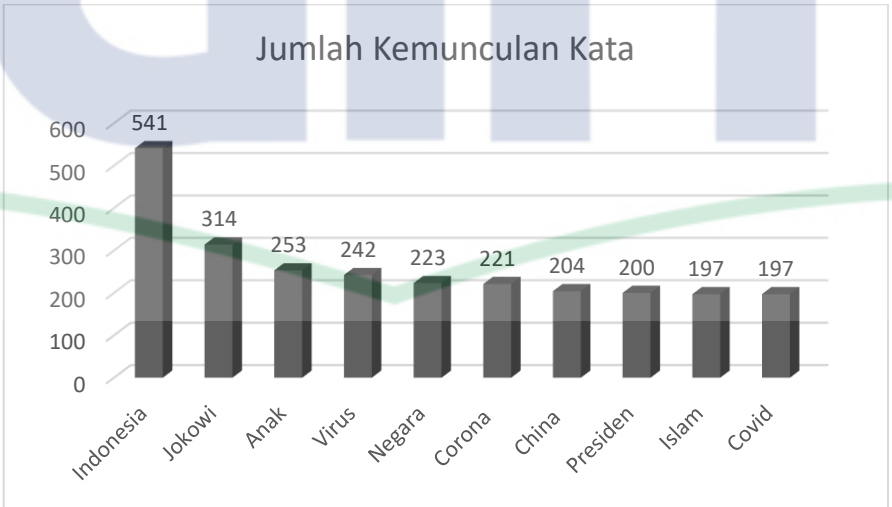


Gambar 4.2. Wordcloud Label berita (a) seluruh berita (b) non hoax dan (c) hoax

Berdasarkan Gambar 4.2 *wordcloud*, jika dilihat berdasarkan labelnya pada gambar (b) dan (c) terdapat kemiripan pada kata yang memiliki keberadaan yang dominan pada data dengan label hoax dan data dengan label non-hoax, seperti narasi berita yang memiliki label non hoax menunjukkan kata-kata yang banyak muncul diantaranya adalah indonesia, hoax, berita, edar, sebar dan lainnya. Sedangkan wordcloud dari narasi berita yang memiliki label hoax menunjukan kata-kata yang banyak muncul diantaranya adalah indonesia, corona, nyata, Jokowi, virus, dan lainnya.

4.2 Pembobotan Kata

Pada penelitian ini pembobotan yang digunakan adalah TF-IDF, nilai parameter pada fungsi Tfidf_vectorizer yang digunakan adalah sesuai dengan standar modul bentuk tfidf yang didapat adalah (4149, 12495), yang artinya bahwa terdapat 4.149 baris dokumen dan 12.495 kata unik.

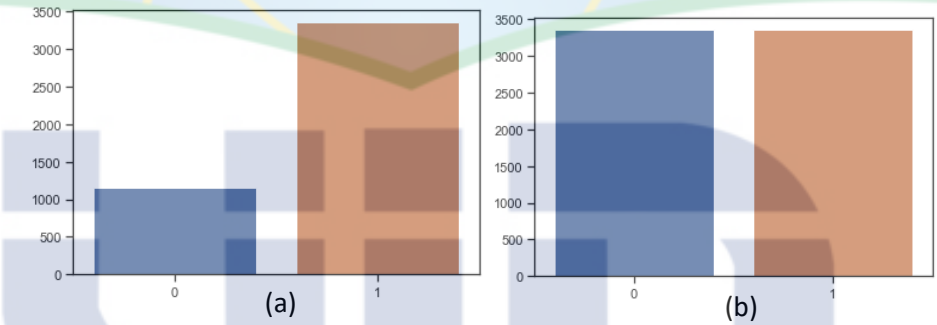


Gambar 4.3. Jumlah Kemunculan Kata

Pada Gambar 4.3 menampilkan sepuluh kata dengan jumlah kemunculan terbanyak pada data yang telah dilakukan *preprocessing*. Kata yang paling banyak muncul adalah kata Indonesia dengan jumlah kemunculan sebanyak 541. Hal ini menunjukkan bahwa kata tersebut tidak muncul di setiap berita sehingga tidak ada kata yang perlu untuk dihilangkan.

4.3 *Resampling* Data

Pada proses sebelumnya dapat diperhatikan bahwa hasil pelabelan yang didapat menunjukkan jumlah label yang tidak seimbang antara label satu dengan yang lainnya, hal ini dapat dilihat di Gambar 4.1. Dikarenakan label yang tidak seimbang tersebut maka perlu dilakukan suatu proses yaitu *Resampling* data yang dimana proses ini dapat membantu kinerja mesin agar lebih baik dalam proses memprediksi suatu model.



Gambar 4.4. Jumlah Label sebelum resampling(a) sesudah resampling(b)

Pada Gambar 4.4 menampilkan proporsi jumlah label sebelum dilakukan resampling dan setelah dilakukan resampling. Pada proses *balancing* data, teknik yang digunakan pada proses balancing data adalah metode SMOTE dimana proses pendekatannya dilakukan dengan pendekatan *over-sampling*.

4.4 Optimasi parameter pada *Machine Learning*

Pada pembentukan model penulis akan memodelkan dataset menggunakan algoritma *Support Vector Machine*, Regresi Logistik, dan *Random Forest* dimana masing-masing dari algoritma tersebut dilakukan optimasi parameter dengan

menggunakan *Grid Search Using Cross Validation (Grid Search CV)*. Tujuan dilakukannya *Grid Search CV* untuk mendapatkan parameter terbaik.

Pada pembentukan model *Random Forest Classifier* dilakukan *cross validation* menggunakan *Grid Search CV*. Parameter yang diujikan adalah nilai estimator 10, 20, 30, 40 dan max depth 4, 6, 8, 10, 20, 100. Hasil optimasi parameter terbaik adalah nilai estimator sebesar 30 dan max depth sebesar 100 dengan akurasi terbaik yang dihasilkan sebesar 0.829. Untuk pembentukan model Regresi Logistik, parameter yang diujikan adalah nilai konstanta c yakni nilai sebagai berikut, yaitu 0, 1, 2, 3 dan teknik regulasi L1 lasso dan L2 ridge. Hasil optimasi parameter terbaik adalah nilai c sebesar 1 dan regulasi L2 ridge dengan akurasi terbaik yang dihasilkan sebesar 0.826. Pada pembentukan model SVM, parameter yang diujikan adalah kernel linier dengan nilai konstanta c yakni 0.1, 1, 10, dan 1000. Hasil optimasi parameter terbaik yang dihasilkan adalah kernel linier dengan nilai konstanta c sebesar 1 dengan akurasi yang dihasilkan sebesar 0.83.

Table 4.2 Hasil Grid Search CV Tiap model

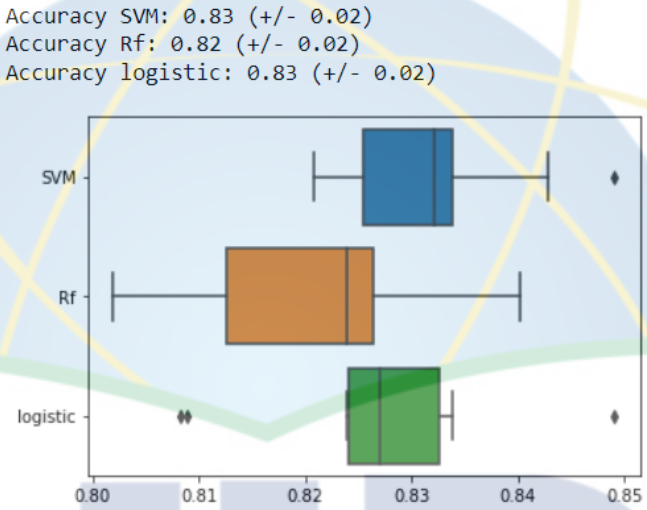
Model	Parameter yang diujikan	Akurasi
Random Forest	n_estimator = 10, 20, 30, 40 max depth = 4, 6, 8, 10, 20, 100	0.829
Regresi Logistik	C = 0.1, 1, 2, 3 Regulasi = l1, l2	0.826
SVM	C = 0.1, 1, 10, 100 dan 1000 Kernel = linier	0.831

Tabel 4.2 merupakan hasil optimasi parameter dengan *Grid Search CV* dari tiap model dengan parameter terbaiknya.

4.5 Hasil Klasifikasi dan Evaluasi Model

Pada tahap ini penulis mencoba untuk mengukur tingkat akurasi dari setiap model *cross validation* yang terbentuk. Hal ini dilakukan untuk menjadi acuan

penulis dalam mengetahui dan melihat model yang terbaik untuk dataset berita hoax dan sebagai alat bantu penulis melihat performa dari kinerja setiap model klasifikasi. Berdasarkan hasil optimasi parameter menggunakan *Grid Search CV*, diperoleh parameter-parameter terbaik untuk setiap model *machine* learning. Dengan menggunakan parameter-parameter terbaik tersebut kemudian dilakukan pengujian dengan menggunakan *data testing* dan diperoleh hasil akurasi seperti pada Gambar 4.5 berikut:



Gambar 4.5. Hasil Akurasi Data Traning Menggunakan Parameter Terbaik

Berdasarkan Gambar 4.5 hasil box plot akurasi data training menggunakan parameter terbaik dari setiap model, dapat dilihat bahwa model SVM dan Regresi Logistik menghasilkan akurasi yang sama yakni 0.83 dan Random Forest menghasilkan akurasi 0.82, dapat dilihat juga pada boxplot SVM dan Regresi Logistik memiliki IQR yang sama namun pada model Regresi Logistik memliki outlier lebih banyak dan nilai median lebih rendah dibanding model SVM. Akurasi pada box plot *cross validation* tidak cukup mewakili untuk menentukan model terbaik, perlu dilakukan evaluasi model dengan mempertimbangkan hasil *precision, recall*, dan *F1-Score* dari masing-masing model menggunakan confusion matrix dan diperoleh hasil perhitungan akurasi setiap model disajikan pada tabel 4.3.

Tabel 4.3. Evaluasi Model

Model	label	*precision*	*recall*	*F1-Score*	Akurasi
SVM dengan SMOTE	0	0.94	0.35	0.51	0.833
	1	0.82	0.99	0.90	
SVM Tanpa resampling	0	0.48	0.62	0.56	0.739
	1	0.86	0.78	0.89	
Regresi Logistik dengan SMOTE	0	0.80	0.40	0.53	0.826
	1	0.83	0.97	0.89	
Regresi Logistik Tanpa resampling	0	0.54	0.62	0.58	0.775
	1	0.87	0.83	0.85	
Random Forest dengan SMOTE	0	0.70	0.46	0.56	0.819
	1	0.84	0.94	0.89	
Random Forest Tanpa resampling	0	0.56	0.59	0.58	0.784
	1	0.86	0.85	0.86	

Berdasarkan Tabel evaluasi model, dapat diketahui bahwa pada masing masing model memiliki kenaikan akurasi yang cukup baik setelah dilakukan teknik resampling, dimana model SVM memiliki akurasi terbaik yakni sebesar 0.833. Jika dilihat dari nilai *precision, recall* dan *F1-score* model SVM pada kelas hoax lebih unggul dibandingkan dengan model lainnya, dalam klasifikasi berita hoax perlu mempertimbangkan nilai recall pada kelas hoax, dimana model SVM memiliki nilai recall sebesar 0.99, sehingga dapat disimpulkan bahwa model SVM merupakan model terbaik yang dapat digunakan untuk memodelkan klasifikasi berita hoax jika dibandingkan dengan model klasifikasi lainnya yang digunakan yaitu Regresi Logistik, dan Random Forest.

4.6 Visualisasi dan Interpretasi Model

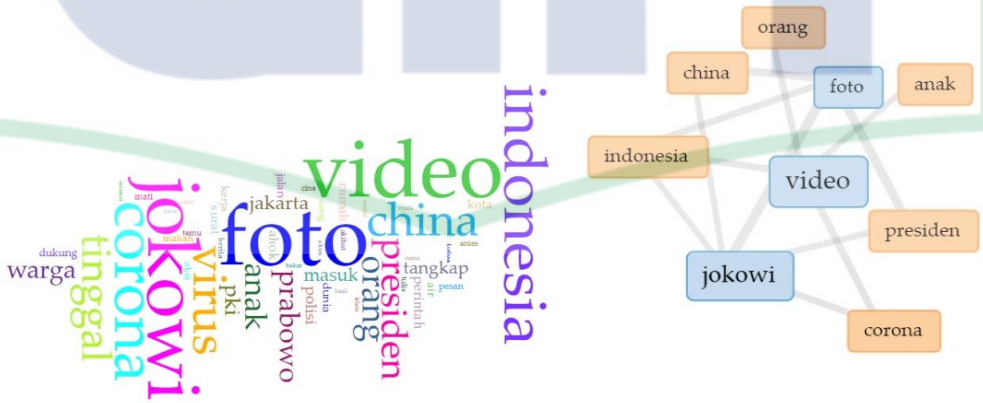
Dari hasil model terbaik yang telah memprediksi ulasan pada data testing yaitu model SVM, selanjutnya adalah visualisasi hasil prediksi untuk melihat opini apa saja yang ada dimasing-masing kelas. Visualisasi dilihat dari *wordclound* dan *wordlink* yang ada pada voyant tools.

Wordcloud berguna untuk melihat keterkaitan kata yang paling banyak muncul pada suatu kelas, sedangkan wordlink berguna untuk melihat keberhubungan antar kata didalam suatu kelas. Sehingga hasil dari visualisasi ini dapat berguna untuk membandingkan berita bedasarkan narasi disetiap kelasnya.

Pada Gambar 4.6 merupakan hasil visualisasi pada kelas *non-hoax*. Hasil visualisasi pada kelas *non-hoax* menunjukkan bahwa berita *non-hoax* dapat ditinjau dengan konten yang berisi klarifikasi dari berita *hoax* dengan menampilkan informasi relefan terkait kebenaran yang ada. Seperti pada narasi berikut "klarifikasi perintah daerah raja ampat kait isu gizi buruk kampung saporkren distrik waigeo selatan menanggapi isu tersebut, Sekertaris Daerah Raja Ampat mengatakan dua anak yang diklaim menderita gizi buruk tersebut bukan penderita gizi buruk melainkan menderita penyakit paru-paru". Pada visualisasi dikelas *non-hoax* ini dapat dijadikan acuan untuk membedakan berita yang benar serta tindakan-tindakan pencegahan berita *hoax*

(a) (b)

Gambar 4.6. Wordcloud dikelas non-hoax (a), Wordlink dikelas non-hoax (b)



(a) (b)

Gambar 4.7. Wordcloud dikelas hoax (a), Wordlink dikelas hoax (b)

Pada Gambar 4.7 merupakan hasil visualisasi pada kelas *hoax*. Dimana berita yang mengandung foto dan video pada kelas *hoax* menjadi kata yang mendominasi dimana isi berita menampilkan gambar dan video yang tidak sesuai dengan narasi berita seperti pada narasi berikut "video kendari memanas pada 2 agustus tentang tka china yang masuk indonesia kantor dan kendaraan polisi dibakar massa." Narasi pada kelas *hoax* ini menunjukkan video yang diklaim menyebutkan bahwa keadaan memanas itu disebabkan tka china masuk Indonesia. Berdasarkan hasil penelusuran, diketahui bahwa kejadian dalam video tersebut terjadi pada Februari 2011 dan terkait peristiwa kericuhan demonstrasi penolakan penambangan emas di Nusa Tenggara Barat. Dari hasil visualisasi di kelas *hoax* perlu dijadikan acuan untuk media yang menyajikan layanan untuk melaporkan berita yang diduga mengandung unsur *hoax*.

BAB V

PENUTUP

Bab ini akan membahas mengenai kesimpulan dari penerapan metode-metode yang digunakan dalam melakukan klasifikasi berita dan saran yang dapat dipertimbangkan untuk penelitian selanjutnya.

5.1 Kesimpulan

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa klasifikasi berita menggunakan machine learning, model SVM dan Regresi Logistik menghasilkan akurasi yang sama yakni 0.83 dan Random Forest menghasilkan akurasi 0.82 Karena Regresi Logistik dan SVM memiliki akurasi yang sama maka dilakukan evaluasi model dengan menggunakan confusion matrix pada model dan diperoleh akurasi terbaik pada model SVM sebesar 0.833 dan nilai recall sebesar 0.99 pada kelas hoax, Artinya model SVM mampu mengklasifikasi sebanyak 83% data uji dengan benar. Hal ini berarti, data berita memiliki performa yang cukup baik pada semua model *machine learning* dengan performa terbaik pada model SVM dengan akurasi 83%.

Opini yang didapatkan pada setiap kelas menghasilkan diantaranya pada kelas hoax yaitu banyaknya narasi berita yang mengandung foto dan video namun tidak sesuai dengan isi berita. Pada kelas non-hoax menghasilkan narasi yang berisi klarifikasi dari berita hoax dengan menampilkan informasi relefan terkait kebenaran tentang berita tersebut.

5.2 Saran

Selain metode yang digunakan pada penelitian ini masih banyak metode lain yang bisa dikembangkan dalam membentuk model deteksi berita *hoax*. Pengambilan data berita dapat dilakukan dari platform lainnya seperti sosial media twitter. Dalam penggunaan dataset, dapat dipertimbangkan untuk menggunakan variabel lain seperti gambar dan beberapa kategori berita.

DAFTAR PUSTAKA

[1] V. M. Rumata, K. Kominfo, and P. Subianto, "Berita palsu dalam segi konsep dan praktis," pp. 31–42, 2018.

[2] M. Athaillah, Y. Azhar, and Y. Munarko, "Perbandingan Metode Klasifikasi Berita Hoaks Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Repos.*, vol. 2, no. 5, pp. 675–682, Mar. 2020.

[3] 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, Oct. 2019.

[4] D. Arisandi, Z. Indra, and K. Kartini, "Mengidentifikasi Hoax pada Hasil Pencarian Berita Online Dengan Teknik Web Scraping dan Algoritma C4.5," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 6, no. 2, pp. 130–137, Jul. 2021.

[5] M. Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. October 2014, pp. 7–16, 2015.

[6] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. 2007.

[7] J. Pollock, E. Waller, and R. Politt, "Speech and language processing," *Day-to-Day Dyslexia Classr.*, pp. 16–28, 2010.

[8] R. E. Neapolitan and X. Jiang, *Artificial Intelligence With an Introduction to Machine Learning*. 2018.

[9] R. A. Haristu and P. H. P. Rosa, "Penerapan Metode Random Forest untuk Prediksi Win Ratio Pemain Player Unknown Battleground," *MEANS*, vol. 4, no. 2, pp. 120–128, 2019.

[10] R. A. Berk, *Statistical Learning from a Regression Perspective*. 2008.

[11] A. Amri, "Implementasi Algoritma Random Forest Untuk Mendeteksi Hate Speech Dan Abusive Language Pada Twitter Bahasa Indonesia," 2020.

[12] J. Ledolter, *Data Mining and Business Analytics with R*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[13] Ramli, D. Yuniarti, and R. Goejantoro, "Perbandingan Metode Klasifikasi Regresi Logistik Dengan Jaringan Saraf Tiruan (Studi Kasus: Pemilihan Jurusan Bahasa dan IPS pada SMAN 2 Samarinda Tahun Ajaran 2011/2012)," 2013.

[14] A. J. Scott, D. W. Hosmer, and S. Lemeshow, *Applied Logistic Regression*. 2013.

[15] S. Ailiyya, "Analisis sentimen berbasis aspek pada ulasan aplikasi tokopedia menggunakan support vector machine," 2020.

[16] F. A. Wijaya, "Studi empiris analisis sentimen kenaikan cukai rokok Indonesia di Tahun 2019 menggunakan data twitter dan ensemble learning," 2019.

[17] P. Wulan and R. Farizi, "Perbandingan klasifikasi tingkat keganasan breast cancer dengan menggunakan regresi logistik ordinal dan support vector machine (SVM)," *J. sains dan seni ITS*, vol. 1, pp. D130–D135, 2012.

[18] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. 2015.

[19] Pusphita Anna Octaviani, Y. Wilandari, and D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang," *J. Gaussian*, vol. 3, pp. 811–820, Jun. 2019.

[20] H. C. S. Ningrum, "Perbandingan metode Support Vector Machine (SVM) Linear, Radial Basis Function (RBF), dan Polinomial Kernel dalam klasifikasi bidang studi lanjut pilihan alumni UII," 2018.

[21] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA*, vol. 14, no. 4, p. 1502, 2016.

[22] Pedregosa, Fabian, Varoquaux, Dubourg, and Vincent, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, 2018.

[23] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*. 2014.

[24] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Appl. Stat.*, vol. 29, no. 2, p. 119, 1980.

[25] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[26] S. Salzberg, "Book review: C4. 5: by j. ross quinlan. inc., 1993. programs for machine learning morgan kaufmann publishers," *Mach. Learn.*, vol. 16, pp. 235–240, 1994.

[27] Breiman *et al.*, "Classification And Regression Trees," *CRC Press*, 1984.

LAMPIRAN

Memanggil Data yang Digunakan

```
df=pd.read_csv ('dataset.csv')
df.tail(12)
```

Membentuk VSM

```
##TFIDF
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
import gensim

vectorizer = TfidfVectorizer(max_df=1.0, min_df=5, ngram_range=(1, 3))

listdata=data['narasi'].values.astype('object')
listdata = [d for d in listdata]

listdata
v = TfidfVectorizer(decode_error='replace', encoding='utf-8')
tfidf = v.fit_transform(data['narasi'].values.astype('U'))
y = data.iloc[:, 2].values
print(tfidf.shape, len(y))
```

Membagi Data Train dan Test

```
from sklearn.model_selection import train_test_split

X_train_tf, X_test_tf, y_train_tf, y_test_tf = train_test_split(tfidf, y, test_size=.2)
print(X_train_tf.shape, X_test_tf.shape)
```

Optimasi Parameter

```
from sklearn.model_selection import GridSearchCV

# defining parameter range
param_grid = {'C': [1, 5, 10]}

grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
grid.fit(X_train_tf, y_train_tf)
print(grid.best_params_)
print(grid.best_score_)

rfc=RandomForestClassifier(random_state=42)
param_grid = {
    'n_estimators': [200, 500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [4,5,6,7,8],
    'criterion' :['gini', 'entropy']
}
CV_rfc = GridSearchCV(estimator=rfc, param_grid=param_grid, cv= 5)
CV_rfc.fit(X_train_tf, y_train_tf)
CV_rfc.best_params_
print("tuned hpyerparameters :(best parameters) ",CV_rfc.best_params_)
print("accuracy :",CV_rfc.best_score_)

from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
grid={"C":np.logspace(-3,3,7), "penalty":["l1","l2"]}# l1 lasso l2 ridge
logreg=LogisticRegression()
logreg_cv=GridSearchCV(logreg,grid,cv=10)
logreg_cv.fit(X_train_tf,y_train_tf)
```

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

Plotting akurasi dengan parameter terbaik

```
clf = LogisticRegression(solver='liblinear')
rf = RandomForestClassifier()
svm_ = svm.SVC()
Models = [('Regresi Logistik', clf), ('Random Forest', rf), ('SVM', svm_)]
Scores = {}
for model_name, model in Models:
    Scores[model_name] = cross_val_score(model, X_train, y_train, cv=10, scorin
g='accuracy')

dt = pd.DataFrame.from_dict(Scores)
plt.figure(figsize=(12,8))
ax = sns.boxplot(data=dt)
for m, s in Scores.items():
    print(m,list(s)[:1])
```

Evaluasi Model

```
svm_.fit(X_train, y_train)
y_pred = svm_.predict(X_test)
import sklearn
sklearn.metrics.accuracy_score(y_test, y_pred)
from sklearn.metrics import accuracy_score, confusion_matrix, classification_r
eport
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
print('Akurasi : ', accuracy_score(y_test, y_pred))
```

```
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
import sklearn
sklearn.metrics.accuracy_score(y_test, y_pred)
from sklearn.metrics import accuracy_score, confusion_matrix, classification_r
eport
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
print('Akurasi : ', accuracy_score(y_test, y_pred))
```

Confusion Matrix

```
fig, ax = plt.subplots(figsize=(10, 10))
plot_confusion_matrix(svm_, X_test, y_test, cmap=plt.cm.Blues, ax=ax)
plt.savefig('confusion matrix.png')
plt.show()
fig, ax = plt.subplots(figsize=(10, 10))
plot_confusion_matrix(rf, X_test, y_test, cmap=plt.cm.Blues, ax=ax)
plt.savefig('confusion matrix.png')
plt.show()
```