ISSN : 2715-6982

Pengembangan Sistem Identifikasi Fakta Dan Tidak Fakta Berita di Media Informasi Berbahasa Indonesia

Muh Maulana Stiawan[1], Rinto Hidayat [2]

Universitas Tanri Abeng
Jl. Swadarma Raya Blok Haji Rohimin No.58, Ulujami, Pesanggrahan, Jakarta Selatan 12250
maulana.setiawan@student.tau.ac.id[1], rinto.hidayat@student.tau.ac.id[2]

Abstrak--Media informasi dapat kita temukan dalam bentuk media elektronik atau media cetak. Media sangat memudahkan masyarakat untuk mendapatkan informasi dan memiliki sebuah kerentanan dalam memahaminya jika tidak bijaksana. Oleh sebab itu menjadi salah satu point penting dalam menilai tingkat kebeneran sebuah berita yang beredar di media elektronik seperti website berita dan media social. Untuk mengatasi hal tersebut peneliti bertujuan mambantu pengguna dalam mendeteksi nilai propabilitas dari sebuah berita. Berita akan diklasifikasikan menjadi dua yaitu: fakta dan tidak fakta. Dataset diambil dari beberapa website berita yang beredar di internet, lalu di tag fakta dan tidak fakta. Dataset diproses dengan text preprocessing lalu dibuat model machine learning dengan algoritma Support Vektor Machine untuk menghitung nilai probabilitas sebuah berita tersebut.

***Kata Kunci*—Berita, Fakta, Support Vektor Machine, Machine Learning**

I. PENDAHULUAN

1.1. Latar Belakang

Saat ini terdapat banyak berita yang beredar di tengah masyarakat yang belum tau letak kebenarannya, sehingga menimbulkan isu yang tidak sesuai dengan keadaan sebenarnya atau sering disebut berita hoaks.

Dampak dari penyerbaran berita hoax secara langsung mampu mempengaruhi pemikiran pembaca terhadap isu-isu berita yang beredar, yang memelintir kebenaran yang sebenarnya, yang mengakibatkan pergerakan massa untuk mempercayai suatu kepentingan yang berpihak.

Perkembangan teknologi saat ini seharusnya bisa berperan membantu mengatasi hal tersebut, salah satunya dengan menggunakan text mining dan machine learning.

Dalam Penelitian ini penulis mengusulkan pendekatan dengan algoritma support vector machine (SVM) untuk mengklasifikasi berita yang berdar di media elektronik. Penelitian ini menggunakan dataset sebanyak 50 berita yang di kasih tag berita valid dan hoaks. Berdasarkan hasil uji coba system ini menghasilkan akurasi 90%, presisi 90% dan recall 90%.

II. LANDASAN

TEORI 2.1. Hoaks

Menurut Muhammad Alwi Dahlan berpendapat ”hoaks merupakan berita yan dimanipulasi secara sengaja dengan tujuan untuk memberikan pengakuan atau pemahaman yang salah”

Menurut Badan Pengembangan dan Pembinaan Bahasa,“Dalam bahasa Indonesia istilah *hoax* diserap menjadi hoaks, padanan kata untuk *hoax* tersebut tercantum dalam Kamus Besar Bahasa Indonesia yang mana hoaks diartikan sebagai berita bohong.”

Menurut Novald Pakar Information Technology (IT) dari Universitas Kristen Duta Wacana Yogyakarta itu juga menilai “ bahwa literasi media dapat menangkal persebaran berita hoaks”

2.2. Web Crawler

Web crawler adalah proses penelusuran atau menrayapi suatu laman atau halaman informasi dari suatu lamam. Tidak hanya merayapi, web crawler juga mengambil halaman informasi dari laman tersebut. fungsi utama dari web crawler adalah untuk melakukan penelurusan atau merayapi informasi dari suatu laman (Muzad & Rahutomo, 2016)

Web crawler dapat digunakan untuk beragam tujuan. Tujuan utamanya adalah mengumpukan data baik berupa informasi tekstual (*textual*) maupun intekstual (*hypertextual*) dari halaman web.

2.4. Text Preprocessing

Text preprocessing adalah untuk membangun satu set fitur yang relevan dari text pada dokumen. Semua set fitur dipilih untuk koleksi dokumen yang disebut repsentional model. Setiap dokumen diwakili oleh nilai numeric vector. Tujuan dari text preprocessing adalah untuk mengekstrak vector fitur berkualitas tinggi untuk setiap dokumen, sehingga values yang diambil adalah values yang bernilai tinggi atau penting (Utami dan Sari 2018).

2.5. Case Folding

Case folding bagian awal dari Text Preprocessing Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu, peran Case Folding dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (huruf kecil atau lower case). Sebagai contoh, user yang ingin mendapatkan informasi **"**BERITA**"** dan mengetik **"**BERITA**",** **"**BeRIta**",** atau **"**Berita**",** tetap diberikan hasil retrieval yang sama yakni **"**berita**".** Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

2.6. Stopword

Stopword adalah tahap mengambil kata-kata penting dari kalimat. Bisa menggunakan algoritma stoplist (membuang kata kurang penting) atau wordlist (menyimpan kata penting). Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah "yang", "dan", "di", "dari" dan seterusnya.

2.7. Pembobotan kata

2.7.1 Term Frekuensi

TF (Term Frequency) adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Pada Term Frequency (TF), terdapat beberapa jenis formula yang dapat digunakan :

- TF biner (binary TF), hanya memperhatikan apakah suatu kata atau term ada atau tidak dalam dokumen, jika ada diberi nilai satu (1), jika tidak diberi nilai nol (0).
- TF murni (raw TF), nilai TF diberikan berdasarkan jumlah kemunculan suatu term di dokumen. Contohnya, jika muncul lima (5) kali maka kata tersebut akan bernilai lima (5).
- TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit term dalam query, namun mempunyai frekuensi yang tinggi.

$$=\{1+ \quad (\quad), \qquad (1)$$

Dimana nilai ft, d adalah frekuensi term (t) pada document (d)**.** Jadi jika suatu kata atau term terdapat dalam suatu dokumen sebanyak 5 kali maka diperoleh bobot = 1 + log (5) = 1.699. Tetapi jika term tidak terdapat dalam dokumen tersebut, bobotnya adalah nol (0).

2.7.2 Inverse Document Frequency (IDF)

IDF (Inverse Document Frequency) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. IDF menunjukkan hubungan ketersediaan sebuah term dalam seluruh dokumen. Semakin sedikit jumlah dokumen yang mengandung term yang dimaksud, maka nilai IDF semakin besar.

Sedangkan untuk Inverse Document Frequency (IDF) dihitung dengan menggunakan formula sebagai berikut:

$$= \quad (\quad) \qquad (2)$$

Dimana D adalah jumlah semua dokumen dalam koleksi sedangkan dfj adalah jumlah dokumen yang mengandung term (tj).

Jenis formula TF yang biasa digunakan untuk perhitungan adalah TF murni (raw TF)**.** Dengan demikian rumus umum untuk Term Weighting TF-IDF adalah penggabungan dari formula

perhitungan raw TF dengan formula IDF dengan cara mengalikan nilai TF dengan nilai IDF:

(3)

Dimana Wij adalah bobot term (tj) terhadap dokumen (di). Sedangkan tfij adalah jumlah kemunculan term (tj) dalam dokumen (di). D adalah adalah jumlah semua dokumen yang ada dalam database dan dfj adalah jumlah dokumen yang mengandung term (tj) (minimal ada satu kata yaitu term (tj)**).**

Berapapun besarnya nilai tfij**,** apabila D = dfj**,** maka akan didapatkan hasil 0 (nol), dikarenakan hasil dari log 1, untuk perhitungan IDF**.** Untuk

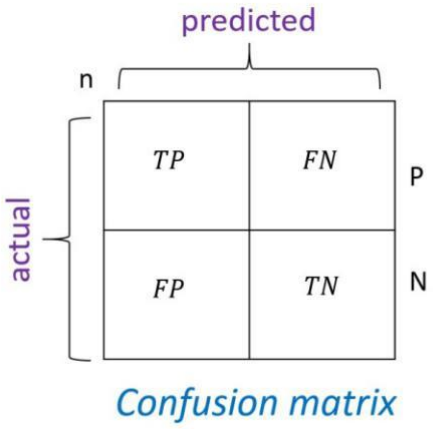
itu dapat ditambahkan nilai 1 pada sisi IDF, sehingga perhitungan bobotnya menjadi sebagai berikut :

(4)

2.8. Confusion Matrik

Ketika sebuah sistem telah berhasil dirancang sebagaimana mestinya dan sudah di implementasikan yang kemudian menghasilkan nilai seperti yang diinginkan, maka tahapan selanjutnya adalah pengukuran performa.

Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Adapun Perhitungan presisi dan recall dalam Confusion Matrix dinyatakan dalam persamaan berikut:



Gambar 1. Confusion Matrix

- TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
- TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

2.9. Precision

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban oleh pengguna dengan jawaban yang di berikan oleh system. Adapun rumus sebagai berikut:

(5)

2.10. Recall

Tingkat keberhasilan system dalam menemukan kembali sebuah informasi.Adapun rumus sebagai berikut :

(6)

2.11. Klasifikasi SVM

Secara umum, Support Vector Machines dianggap sebagai pendekatan klasifikasi, tetapi dapat digunakan dalam kedua jenis klasifikasi dan masalah regresi. Ini dapat dengan mudah menangani beberapa variabel kontinu dan kategorikal. SVM membuat hyperplane dalam ruang multi dimensi untuk memisahkan kelas yang berbeda. SVM menghasilkan hyperplane optimal secara berulang, yang digunakan untuk meminimalkan kesalahan. Ide inti dari SVM adalah untuk menemukan hyperplane marginal (MMH) maksimum yang membagi dataset menjadi kelas - kelas terbaik.

2.12. Bahasa Pemrograman Python

Bahasa Pemrograman Python termasuk ke dalam kategori *highlevel language* atau bahasa pemrograman yang mendekati bahasa manusia. Berbeda dengan *lowlevel language*, *highlevel language* tidak dapat dijalankan secara langsung

oleh mesin sehingga perlu diproses terlebih dahulu agar dapat dijalankan. Karena Bahasa Python merupakan jenis *highlevel language* maka terdapat beberapa keuntungan di dalamnya yaitu dalam penulisan program yang tidak akan memakan banyak waktu, mudah untuk dibaca dan lebih mudah untuk dibenarkan. Kemudian Bahasa Python ini juga dapat dijalankan diberbagai macam komputer.

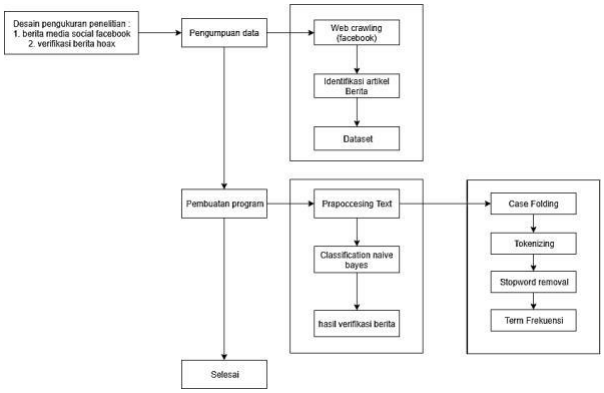
2.13. Postman

Postman adalah aplikasi yang digunakan mengenarate data set berita yang telah di buat dan sebagai GUI API Caller

3. METODE PENELITIAN

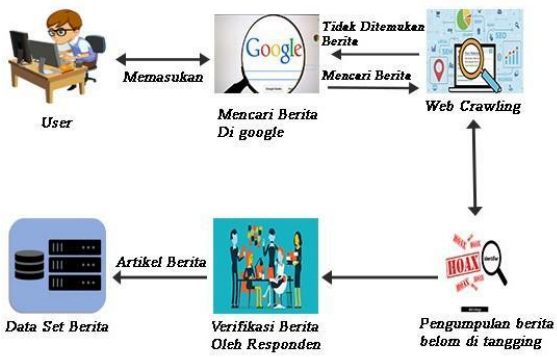
3.1. Kerangka Pemikiran

Kerangka pemikiran dalam penelitian ini menjelaskan tentang perancangan sistem identifikasi fakta dan tidak fakta berita di Media Informasi berbahasa indonesia. Penelitian ini bertujuan untuk mambantu pembaca dalam menyaring sebuah berita yang beredar di media elektronik agar tidak di provokasi oleh pihak tertentu.



Gambar 2. Kerangka Pemikiran

3.2. Teknik Pengumpulan Data



Gambar 3. Alur Pengumpulan Data

Peneliti mengumpulkan dataset secara manual melului internet, dimana website google.com sebagai mesin pencari. Peneliti memasukan kata kunci berita di website google.com. Google.com akan mencari berita yang sesuai dengan kata kunci yang dimasukan lalu akan mengarahkan ke portal - portal berita yang tersebar diinternet, berita - berita tersebut dikumpulan, setelah berita terkumpul dilalukan tag fakta dan tidak fakta pada berita tersebut berdasarkan hasil voting dari 3 responden.

Tabel 1. Penentuan Tag Berita

Berita	R1	R2	R3	Tag
Radi Hartono Meninggal	0	0	0	0
operasi militer di distrik gome	0	1	0	0
Keterangan	0 = Tidak Fakta 1 = Fakta R = Responden			

Data berita yang dikumpulkan secara manual dari portal berita diinternet berjumlah 50 berita, 26 berita ditag tidak fakta dan 25 berita ditag fakta.

4. HASIL DAN PEMBAHASAN

Sistem klasifikasi pada berita ini terdiri dari beberapa proses yakni dari text preprocessing, pembobotan kata, split data training dan data testing lalu klasifikasi menggunakan metode Support Vektor Machine.

4.1. Pengujian Klasifikasi SVM

Pengujian model klasifikasi Support vector machine dilakukan 3 rasio pembagian data yaitu:

1. Pembagian data traning dan data testing sebesar 70:30
2. Pembagian data training dan data testing sebesar 60:40.
3. Pembagian data training dan data testing sebesar 50:50.

Hasil pengujian ditunjukan pada Tabel 1.

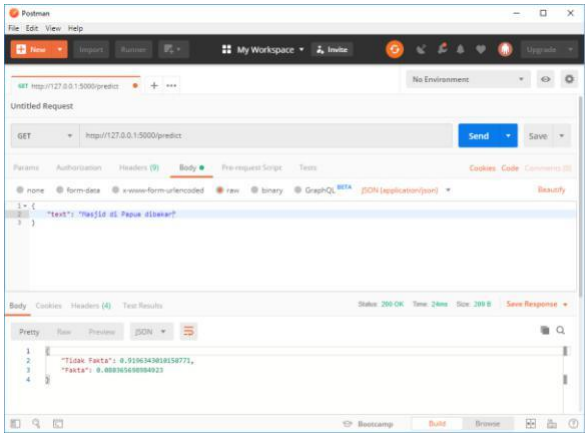
Tabel 2. Hasil Pengujian

Algoritma SVM			
		Presisi	Recall
Radio 30:70	Training	0.87	0.87
	Testing	0.92	0.92

Radio 40:50	Training	0.90	0.90
	Testing	0.94	0.94
Radio 50:50	Training	0.88	0.88
	Testing	0.92	0.93

4.2. Pengujian Sistem

Setelah model dibuat dan hasil pengujian dengan nilai Precision dan Recall diatas 87%. Peneliti melakukan pengujian kembali menggunakan aplikasi postman.



Gambar 3. Pengujian menggukan postman

Dari gambar diatas pengujian dilakukan dengan menginput text berita "Masjid di Papua di bakar", system menyatakan nilai probabilitas Tidak Fakta(Hoaks) : 0.9196343010150771, dan nilai Fakta : 0.080365698984923.

5. PENUTUP

5.1. Kesimpulan

Berdasarkan penelitian dan pengujian yang telah dilakukan, terdapat beberapa kesimpulan sebagai berikut :

- Algoritma Support Vektor Machine dapat digunakan untuk menklasifikasi berita berupa berita Fakta dan Tidak Fakta
- Algoritma Support Vektor Machine mempunyai kelebihan pada dataset yang memiliki output 2 label.
- Pengujian yang dilakukan dengan 3 rasio pembagian data, ketiga mencapai nilai minimal 87%.

5.2. Saran Pengembangan

Beberapa Saran yang diberikan dari hasil penelitian dan pengujian yang telah dilakukan untuk pengembangan sistem sebagai berikut:

- Perngembangan dataset dengan perbanyak berita dengan berbagai isu yang ada.
- Pengembangan dataset dengan berita yang realtime.
- Labelin dataset secara manual dapat dikaji kembali agar system lebih akurat.

DAFTAR PUSTAKA

[1] Sadewo Angger Dimas Bayu, Widasari Edita Rosana,MuttaqinAdharul.2016, *"Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia,"* Jurnal Penelitian Komunikasi dan Opini Publik,. 23(1):1-15.

[2] Utami Putri Dinda dan Sari Risam. 2018 "*Filtering Hoax Menggunakan Naive Bayes Classifier*,". Jurnal Multimatics,. 4(1).

[3] Prasetyo Andre Rino, Indriati, Adikara Putra Pandu. 2018 *"Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor,"* J-PTIIK, 2(12):7466-7473.

[4] Muadz Aad Miqdad Muadz dan Rahutomo Faisal. 2016 *"Korpus Berita Daring Bahasa Indonesia dengan Depth first Focused Crawling,"* belum terbit.

[5] Informatikkalogi. 2017, *"Pembobotan kata atau Term Weighting TF-IDF,"* di https://informatikalogi.com/term-weighting-tf-idf/ (akses 14 September 2019).

[6] Informatikkalogi. 2017, *"Text Preproccesing,"* di https://informatikalogi.com/text-preprocessing/ (akses 14 September 2019).

[7] Aldwairi Monther dan Alwahedi Ali. 2018 *"Detecting Fake News in Social Media Networks."* The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks. Elsevier, 215-222.

[8] Afriza Aulia dan Adisantoso Julio. 2018 *"Metode Klasifikasi Rocchio untuk Analisis Hoax,"* ISSN, 5(1):1-10.

ISSN : 2715-6982

[9] Rasywir, Errissya, Ayu Purwarianti. 2015, *"Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbassis Pembelajaran Mesin,"* Jurnal Cybermatika,. 3 (2).

[10] Garuda Cyber Indonesia. 2018, *"Belajar Mengenal Metode Klasifikasi SVM,"* di https://garudacyber.co.id/artikel/654-belajar-mengenal-metode-klasifikasi-svm (akses 14 November 2019).