BAB II

TINJAUAN PUSTAKA

2.1 Vaksin Covid-19

Vaksin adalah zat biologis-imun dirancang untuk menghasilkan perlindungan khusus terhadap penyakit tertentu. Proses pemberian vaksin disebut vaksinasi. Dengan kata lain, vaksinasi adalah proses melindungi individu yang rentan dari penyakit dengan pemberian agen yang hidup atau yang dimodifikasi (misalnya, vaksin polio oral), penangguhan organisme yang dimatikan (seperti pada pertusis), atau toksin yang tidak aktif (seperti di tetanus). Tujuan vaksinasi yaitu untuk melindungi individu yang berisiko terkena penyakit seperti anak-anak, orang tua, individu dengan gangguan kekebalan, orang yang hidup dengan penyakit kronis, dan orang yang tinggal di daerah endemis penyakit merupakan yang paling berisiko. Vaksinasi adalah strategi umum untuk mengontrol, menghilangkan, memberantas, atau menahan penyakit (seperti strategi imunisasi massal) [5].

Coronavirus merupakan virus RNA dengan ukuran partikel 120-160 nm. Virus ini terutama menginfeksi hewan, termasuk kelelawar dan unta. Sebelum merebaknya COVID-19, terdapat 6 jenis virus corona yang dapat menginfeksi manusia yaitu α-coronavirus 229E, α-coronavirus NL63, β-coronavirus OC43, β-coronavirus HKU1, dan penyakit saluran pernapasan akut berat (SARS-CoV).) dan Virus Corona Sindrom Pernafasan Timur Tengah (MERS-CoV). Virus corona penyebab COVID-19 termasuk dalam genus Beta Coronavirus [6].

2.2 Text *Mining*

Menurut jurnal [7] yang dikutip dari buku (Ronen Feldman, 2007) menjelaskan *Text Mining* merupakan sebuah proses yang dianalisa dalam informasi yang berupa teks dimana sumber *data* didapatkan dari sebuah dokumen. Konsep yang ada pada *text Mining* digunakan sebagai klasifikasi dokumen tekstual di mana dokumen-dokumen yang ada akan diklasifikasikan menurut dokumen yang akan di proses. Dengan adanya konsep ini membuat artikel yang diteliti akan diperjelas kategori jenisnya melalui kata-kata yang akan muncul dari artikel yang ada. Kata-kata yang terdapat dari artikel tersebut dapat dicocokan dan akan dianalisa dengan basis kata kunci yang telah ditentukan, sehingga dengan adanya proses ini dapat mengelompokkan dokumen tersebut dengan waktu yang efisien [7].

Menurut [8] *Text Mining* dideskripsikan sebagai sebuah *technology* yang digunakan untuk menganalisis *data* teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan *data Mining*.

Tujuan dari *text Mining* adalah mendapatkan infromasi yang berguna dari sekumpulan banyak dokumen. Adapun tugas lain dari *text Mining* yaitu melakukan kategorisasi dan pengelompokan *data*.

Perbedaan dari *text Mining* dengan *data Mining* dapat dilihat dari sumber *data* yang akan dipakai saat analisis. Dalam *text Mining* pola-pola yang diekstrak dari *data* tekstual yang tidak terstruktur bukan berasal dari suatu *database.* [8]

2.3 Analisis Sentimen

Analisis sentimen adalah bagian dari cabang ilmu penambangan teks, program bahasa alami, dan kecerdasan buatan. Proses yang dilakukan oleh *Sentiment analysis* ini dilakukan untuk mengolah dan mengekstrak *data* teks secara cepat dan otomatis agar dapat menjadi *data* yang lebih mudah untuk diolah menurut jurnal [9]. Selain itu analisis sentimen digunakan utnuk menganalisia pendapat, sikap, evaluasi, dan penilaian terhadap suatu peristiwa, topik, organisasi, maupun perseorangan.

2.4 Media Sosial

Media sosial yaitu sebuah perkembangan teknologi *web internet* based yang memiliki tujuan untuk memudahkan masyarakat untuk melakukan berbagai macam aktivitas seperti berkomunikasi, berpartisipasi, saling berbagi informasi dan membentuk sebuah jaringan secara *online*. Seperti postingan di video *youtube* yang dapat dilihat secara *live* oleh ratusan juta orang secara cuma cuma [10].

Sosial Media bentuknya bermacam-macam, contohnya seperti *microblogging (twitter), facebook,* dan *youtube. Twitter* merupakan situs web yang menjadi bagian dari layanan dari *microblog*. *Twitter* merupakan sosial media yang sangat mudah

digunakan, karena dengan waktu yang sedikit dapat melakukan penyebaran informasi secara luas [11].

2.5 *Data Preprocessing*

Preprocessing teks adalah tahap awal dari penambangan teks. Pada tahap ini dilakukan proses penyusunan dokumen dan *data* agar dokumen / *data* siap untuk diolah dan proses klasifikasi dapat diolah dengan baik [12]. Ada juga tahapan dalam *Preprocessing* teks, yaitu:

1. *Combining Data*: Pengabungan *data* menjadi 1 document dengan format csv.

2. *Cleansing*: digunakan untuk menghapus karakter yang dianggap kurang penting seperti *username* (@), hashtag (#), URL, tanda baca (punctuation), angka, dan emoticon.

3. *Transform Cases*: Perubahaan semua huruf menjadi lowercase.

4. *Tokenizing*: digunakan untuk memisahkan kata menggunakan whitespace.

5. *Filter Stopwords* : digunakan untuk kata-kata yang sering digunakan dalam kehidupan sehari dan tidak memiliki penting dicari apabila digunakan dalam proses pencarian.

6. *Filter Token* : Proses penghilangan dari token-token yang ada dengan huruf kurang dari 2.

7. *Generate Bigram :* Proses mempersatukan 2 kata menjadi 1 token untuk menaikan konteks akurasi dan kalimat yang ada.

Tujuan dari pemrosesan teks adalah supaya *data* yang didapatkan lebih terstruktur agar lebih mudah untuk dilakukan pengolahan *data*[13].

2.6 *Naïve Bayes Classifier (NBC)*

Naive Bayes adalah salah satu algoritma pembelajaran induktif yang paling efisien, juga efektif untuk melakukan proses me-*Mining data*. Performa dari algoritma *Naive Bayes* cukup tinggi, hal ini dibuktikan pada berbagai penelitian empiris [14].

Dalam algoritma *Naive Bayes Classifier* setiap dokumen direpresentasikan dengan pasangan atribut "x_1 , x_2 , $x_{3\ldots\ldots}$ x_n" di mana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Sedangkan V_{map} adalah himpunan kategori opini [15]. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan, di mana persamaanya adalah sebagai berikut

$$V_{map} = \frac{\arg max}{V\ jev} \left(\frac{P(X1,X2,X3,\ldots\ldots Xn,|VJ)P(Vj)}{P(X1,X2,X3,\ldots\ldots Xn)}\right)$$

$$V_{map} = (X1, X2, X3, \ldots\ldots Xn, |VJ)P(Vj)) \qquad (2)$$

$$V_{map} = \frac{\arg max}{V\ jev} \prod_{i=1}^{n} P(Xi|Vj)\mathrm{P(Vj))} \qquad (3)$$

Rumus 2. 1. Naïve Bayes Classifier

Di mana:

V_j = Kategori Opini

$P(X_i|V_j)$ = Probabilitas X_i pada kategori V_j

$P(V_j)$ = Probabilitas dari V_j

Menurut jurnal [16] Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah *data* pelatihan (Training *Data*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* berfungsi lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

2.7 *Support Vector Machine*

SVM adalah serangkaian metode pembelajaran yang diamati terkait yang digunakan untuk klasifikasi dan regresi. SVM termasuk dalam keluarga klasifikasi linier umum. Properti Khusus SVM, SVM meminimalkan kesalahan klasifikasi empiris dan mengoptimalkan margin geometris. Jadi SVM disebut margin classifier maksimum. SVM didasarkan pada minimalisasi risiko struktural (SRM). Vektor input kartu SVM pada dimensi yang lebih tinggi di mana hiperplan pemisah maksimum dibangun. Dua hyperplan paralel dibuat di setiap sisi hyperplan yang digunakan untuk memisahkan *data*. Hyperplan pemisahan adalah hyperplan yang mengoptimalkan jarak antara dua hyperplan paralel. Hipotesis dibuat bahwa semakin besar margin atau jarak antara hiperplanes paralel ini, yang terbaik adalah generalisasi kesalahan klasifikasi [17].

2.8 *K - Nearest Neighbor*

Algoritma *K-Nearest Neighbor* merupakan sebuah algoritma yang sering digunakan untuk mengklasifikasikan obyek berdasarkan atribut dan *data* training. Dengan adanya titik query, akan ditemukan sejumlah k obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi yang akan digunakan yaitu menggunakan voting terbanyak di antara klasifikasi dari k obyek. Algoritma *K-Nearest Neighbor* (K-NN) menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru [17].

2.9 *Random Forest*

Random Forest adalah algoritma yang digunakan untuk klasifikasi dan regresi. Metode ini adalah satu set (koleksi) metode pembelajaran menggunakan pohon keputusan sebagai classifier-classifier yang dibangun dan dikombinasikan. Ada tiga aspek penting dalam metode hutan acak, yaitu: sampling bootstrap untuk membangun pohon prediktif; Setiap pohon keputusan memprediksi dengan prediktor acak; Kemudian hutan acak melakukan prediksi dengan menggabungkan hasil setiap pohon keputusan melalui sebagian besar suara untuk klasifikasi atau rata-rata regresi [19].

Langkah pertama adalah melakukan hasil entri *data* dari transformasi *data* yang terdiri dari atribut eksplanatori dan atribut tujuan. Setelah *data* dibagi menjadi dua jenis (pelatihan *data* dan tes *data*) menggunakan metode validasi silang 10 kali. Selain itu, penentuan *data* tentang pelatihan dan tes *data* dilakukan dengan menggunakan semua *data*. Selain itu, hasil hasil akan dilakukan antara dua jenis metode untuk menentukan pelatihan *data* dan *data* uji [19].

2.10 *PyCharm*

PyCharm adalah aplikasi yang disarankan untuk pemrograman Python, yang dibuat oleh jetbrains. JetBrains adalah perusahaan yang memproduksi berbagai ide (lingkungan pengembangan terintegrasi) untuk berbagai bahasa pemrograman, seperti IntelliJ Java, PHPStorm untuk PHP, Rubymine untuk Ruby, WebStorm untuk JavaScript, *PyCharm* for Python, dll. Selain digunakan untuk belajar, *PyCharm* ini juga berlari oleh guru dan guru yang akan diajarkan di agensi dan sekolah [13].

2.11 Rapidminer

RapidMiner adalah perangkat lunak sumber terbuka, di mana ia bekerja untuk menganalisis hal-hal seperti *data* dan penambangan teks. RapidMiners menggunakan beberapa teknik, sebagai deskriptif dan prediktif untuk memberikan output informatif kepada pengguna sehingga hasil yang dihasilkan dapat berfungsi sebagaimana mestinya. RapidMiner pertama kali diklasifikasikan sebagai perangkat lunak penambangan *data* dalam survei KDNuggets, portal penambangan *data* pada 2010-2011 [13].

2.12 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

No	Penulis	Nama Jurnal	Judul	Permasalahan	Metode	Hasil
1	Prasetyo, Andre Rino Indrianti Adikara, Putra Pandu	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer	Klasifikasi *Hoax* Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor	Banyaknya berita mengenai kesehatan yang disebar melalui grup di beberapa jejaring sosial merupakan *hoax*	Modified K-Nearest Neighbor	Hasil dari pengujiannya yakni akurasi tertinggi sebesar 75% terdapat pada pengujian k-values yang bernilai 4, dengan nilai *precision* 0,83, nilai *Recall* 0,75 dan nilai f-measure 0,79.
2	Trisna Astono Putri, Tansa Warra, HendryxS Yanti Sitepu, Irma Sihombing, Marita	JIPN (Journal Of Informatics Pelita Nusantara) On Machine Learning	Analysis And Detection Of *Hoax* Contents in Indonesian News Based	Penyebaran hoaks di masyarakat dapat menimbulkan dampak negatif, seperti kerusakan, kerugian, baik materil maupun psikis, ketidakpercayaan	Support Vector Machine, Naïve Bayes, *Random Forest*, dan Decision Tree *Algorithms*.	Algoritma yang mendapatkan hasil terbaik yaitu *Random Forest Algorithm* dengan akurasi 76.47%

				masyarakat dan sebagainya.		

3	Faisal Rahutomo, Inggrid Yanuar Risca Pratiwi, Diana Mayangsari Ramadhani	Jurnal Penelitian Komunikasi dan Opini Publik Vol. 23 No. 1, Juni 2019: 1-15	Eksperimen Naïve Bayes pada deteksi berita *hoax* berbahasa Indonesia	Artikel berita yang tersebar di situs- situs website dapat diarahkan oleh penulis berita ke arah tipuan atau sesuatu yang tidak benar. Informasi palsu dan menyesatkan ini berbahaya karena dapat menyesatkan persepsi manusia dengan menyampaikan informasi yang tidak benar	Naïve Bayes	Metode naïve bayes ini memiliki kelebihan dapat bekerja baik pada *data*set yang jumlahnya sedikit, mudah dibuat, cepat dalam proses penghitungan.
4	Christevan Destitus, Wella, Suryasari	ULTIMA InfoSys, Vol. XI, No. 2	Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di	cyberbullying yang terjadi pada media sosial Twitter berupa *Tweet* – *Tweet* yang mengandung kata – kata yang	Support Vector Machine , Information Gain	hasil identifikasi *Tweet* cyberbullying dengan kedua metode memiliki hasil yang cukup maksimal.

			Twitter Indonesia	berisikan konten negatif.		

2.13 Analisa Jurnal Terdahulu

Pada jurnal [18], permasalahan yang ada yaitu banyaknya berita mengenai kesehatan yang disebar melalui grup di beberapa jejaring sosial merupakan *hoax*. Hasil dari pengujiannya yakni akurasi tertinggi sebesar 75% terdapat pada pengujian k-values yang bernilai 4, dengan nilai *precision* 0,83, nilai *Recall* 0,75 dan nilai f-measure 0,79.

Pada jurnal [12], permasalahannya penyebaran hoaks di masyarakat dapat menimbulkan dampak negatif, seperti kerusakan, kerugian, baik materil maupun psikis, ketidakpercayaan masyarakat dan sebagainya. Hasilnya akurasi Algoritma yang mendapatkan hasil terbaik yaitu *Random Forest Algorithm* dengan akurasi 76.47%.

Pada jurnal [19], permasalahan yang ada yaitu adanya *hoax* yang menyesatkan membuat orang bingung dalam memutuskan informasi yang *hoax* dan *non hoax*. Dengan penggunaan metode *Naïve Bayes* maka hasilnya yaitu algoritma ini bekerja dengan sangat baik.

Pada jurnal [20], permasalahannya yaitu cyberbullying yang terjadi pada media sosial menggunakan banyak kata negatif. Menggunakan algoritma SVM dan Information Gain, hasil yang dapat dianalisa yaitu hasil identifikasi menggunakan kedua algoritma tersebut berjalan dengan baik.

Dengan melakukan perbandingan dengan jurnal [12], terdapat perbedaan dengan penelitian ini yaitu perbandingan yang dilakukan pada jurnal tersebut yaitu perbandingan antar berita, untuk penelitian ini membandingkan berita dengan *Tweet* yang didapat dari *twitter*.