BAB II
LANDASAN TEORI

2.1 Dasar Teori

2.1.1 Twitter

Twitter adalah jejaring sosial dan *microblogging* yang dapat membuat pengguna memberikan informasi tentang dirinya sendiri, bisnis, dan informasi lainnya. Update yang dilakukan pengguna disebut dengan *tweet* yang berbasis teks dan dibatasi sejumlah 140 karakter. Twitter diluncurkan pada tahun 2006 oleh Jack Dorsey, Biz Stone, Evan Williams. Twitter memiliki julukan yaitu *"SMS of The Internet"* karena menjadi aplikasi internet yang dapat mengirim pesan pendek ke aplikasi-aplikasi lain [8].

2.1.2 Text Mining

Text mining adalah proses pengumpulan teks dengan menggunakan komputer untuk mendapatkan sesuatu informasi baru yang sebelumnya tidak diketahui yang kemudian informasi baru tersebut akan diekstrak secara otomatis dari berbagai sumber yang berbeda [9]. Proses penambangan data yang dilakukan pada *text mining* banyak mengadopsi penambangan data pada *data mining* tetapi ada perbedaan pola dalam proses kerjanya yaitu *text mining* menggunakan sekumpulan bahasa alami yang belum memiliki struktur sedangkan *data mining* menggunakan data yang sudah terstruktur [10].

2.1.3 Text Preprocessing

Tahap *text preprocessing* merupakan tahapan pertama yang harus dilakukan dalam *text mining*. Pada tahapan *text preprocessing* akan dilakukan proses mempersiapkan dataset yang akan digunakan pada proses berikutnya [9]. Tahapan yang dilakukan pada *text preprocessing* yaitu :

a. *Cleaning* yaitu menghapus kata-kata yang tidak diperlukan seperti *hashtag*, *url*.

b. *Case Folding* yaitu mengubah teks menjadi menjadi huruf kecil atau besar.
c. *Tokenizing* yaitu memisahkan atau memecah kalimat menjadi sebuah potongan kata.

2.1.4 Klasifikasi

Klasifikasi merupakan contoh masalah yang dapat diselesaikan dengan cara *text mining* [3]. Klasifikasi merupakan proses pencarian sebuah model yang dapat membedakan kelas dari data, tujuannya adalah untuk menggunakan model dalam melakukan prediksi kelas dari objek yang belum diketahui kelasnya [11]. Hasil yang didapat dari proses klasifikasi biasanya berbentuk menjadi sebuah *decision tree* atau bisa disebut dengan pohon keputusan [12]. Algoritma yang bisa digunakan untuk melakukan proses klasifikasi adalah Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, C4.5, ID3. Artificial Neural Network [4].

2.1.5 Naïve Bayes

Naïve Bayes adalah metode klasifikasi yang menerapkan teorema Bayes. Naïve Bayes berfungsi untuk menghitung probabilitas tiap kelas pada data dengan menganggap tiap kelas yang ada pada data tersebut tidak saling ketergantungan (independen). Pada metode Naïve Bayes setiap atribut memiliki bobot yang sama penting dan saling bebas sehingga semua atribut dapat memberikan kontribusi dalam proses pengambilan sebuah keputusan [13]. Berikut persamaan 2.1 yang merupakan algoritma Naïve Bayes :

$$P(H|X) = \frac{P(\boldsymbol{X}|\boldsymbol{H}).P(\boldsymbol{H})}{P(\boldsymbol{X})} \quad (2.1)$$

Keterangan :

X : Data dengan class yang tidak diketahui

H : Hipotesis data merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi posteriori

P(H) : Probabilitas hipotesis H

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

2.1.6 Natural Language Processing

Natural Language Processing (NLP) merupakan salah satu ilmu kecerdasan buatan yang digunakan dalam mengolah bahasa alami. Natural Language Processing (NLP) merupakan kemampuan sebuah sistem untuk memproses bahasa dalam bentuk lisan maupun tulisan yang digunakan oleh manusia dalam kehidupannya dengan menerapkan proses komputasi agar sistem dapat memahami perintah-perintah dalam standar bahasa manusia [14]. Proses yang dilakukan Natural Language Processing (NLP) dalam mengolah bahasa alami yaitu dengan melakukan konversi teks menjadi kueri atau dokumen yang kemudian dilakukan analisis untuk menemukan morfologi kata-kata dan melakukan proses pengolahan leksikal dan sintaksis untuk mendapatkan bentuk karakteristik dari setiap kata yang diolah, pengenalan *part-of-speech,* menentukan frasa, dan menguraikan kalimat yang diolah [15].

2.1.7 Named Entity Recognition

Named Entity Recognition (NER) adalah kemampuan yang digunakan dalam melakukan ekstraksi informasi dengan menerapkan pemodelan Natural Language Processing (NLP) dalam proses ekstraksinya. Named Entity Recognition digunakan untuk mengidentifikasi atau mengenali entitas dari informasi yang telah di ekstraksi. Setelah proses identifikasi entitas selesai maka selanjutnya entitas akan diklasifikasian agar sesuai dengan tiap entitas yang didapat. Entitas yang dikenali pada NER biasanya meliputi orang, lokasi, organisasi, tanggal, waktu, durasi, uang, persen, numerik, dan kardinal pada sebuah informasi yang diekstraksi. Pola untuk mengenali entitas pada NER dapat didefinisikan secara manual dalam penentuan entitas atau dilakukan secara otomatis menggunakan pembelajaran mesin sehingga ada dua pendekatan yang bisa digunakan dalam melakukan pengenalan sebuah

entitas dari informasi yang ada yaitu berbasis aturan dan berbasis pembelajaran [16].

2.1.8 N-Gram

N-gram adalah potongan dari n-karakter yang berasal dari sebuah kata tertentu atau sebuah potongan dari n-kata yang berasal dari sebuah kalimat tertentu [6]. Keuntungan dari penggunaan n-gram adalah ketika terjadi kesalahan pada sebagian string hanya mengakibatkan perbedaan pada sebagian dari N-gram [17]. Berikut contoh dari penerapan n-gram dapat dilihat pada Tabel 2.1.

Contoh kalimat : Kasus demam berdarah di Lampung

Tabel 2.1 Contoh penerapan N-Gram

Nama	Hasil N-gram
Unigram	Kasus, demam, berdarah, di, Lampung
Bigram	Kasus demam, demam berdarah, berdarah di, di lampung
Trigram	Kasus demam berdarah, demam berdarah di, berdarah di Lampung

2.2 Tinjauan Pustaka

Adapun hasil tinjauan pustaka yang diambil dari penelitian terkait yang digunakan untuk menjadi acuan penulis mengajukan penelitian dapat dilihat pada Tabel 2.2.

Tabel 2.2 Tinjauan pustaka

No	Penulis	Objek	Metode	Hasil
1	Lazqar Markus Oktavianto	Data Kasus Penyakit Yang Didapatkan Dari Postingan Twitter Tentang Penyakit	Naïve Bayes & NER	NB = 91,5 % NER = 96 %
2	Harold Situmorang	Data Kasus Demam Berdarah Yang Didapatkan Dari Dinkes Medan	SVM	75 %

3	Aries Setiawan, Adi Prihandono	Data Kerentanan Malaria Pada Kabupaten Purworejo	Naïve Bayes	93,75 %
4	Sandi Fajar Rodiyansyah, Edi Winarko	Data Kemacetan Yang Didapatkan Dari Postingan Twitter Tentang Kemacetan Bandung	Naïve Bayes	91,60 %
5	Ziza Amira Syafini, Muhammad Nasrun, Casi Setianingsih	Data Kemacetan Yang Didapatkan Dari Akun Twitter @TMCPoldaMetro dan @lewatmana	KNN	71,7 %
6	Nelly Nur Hamidah, Ahmad Fathan Hidayatullah	Data Bencana Alam Dari Akun Twitter BMKG & BNPB	Naïve Bayes	96,5 % - 99,5 %
7	Kennedy Espina, Regina Justina	Data Demam Berdarah & Tifus Yang Didapatkan Dari Postingan Twitter Masyarakat Filipina	SVM	90,09%
8	Romy Ranovan, Afrizal Doewes, Ristu Saptono	Data Penyakit Daerah Tropis Yang Didapatkan Dari Postingan Twitter Tentang Penyakit Tropis Di Indonesia	Multinomial Naïve Bayes	87,26% & 93,68%
9	Gigih Rezki Septianto, Firman Fakhri Mukti, Muhammad Nasrun, Alfian Akbar Gozali	Data Kemacetan Yang Didapatkan Dari Akun Twitter @TMCPoldaMetro dan @lewatmana	Naïve Bayes	61,66%
10	Son Doan, QuocHung-Ngo, Ai Kawazoe, Nigel Collier	Data Penyakit Menular Yang Didapatkan Dari Berita Wabah Penyakit Di Internet	Naïve Bayes	88,10%
11	Sweta Swain, K.R. Seeja	Data Demam Berdarah Dan Chikungunya Yang Didapatkan Dari Postingan Twitter Terkait Demam Berdarah dan Chikungunya	Naïve Bayes	92%
12	Yuda Munarko	Data Informasi Lalu Lintas Yang Didapatkan	NER	P = 99,43 %

		Dari Akun Twitter RTMC_Jatim, SbyTrafficServ, Radio Suara Surabaya		R = 98.89 % F1 = 99,16 %
13	John Lingad, Sarvnaz Karimi, Jie Yin	Data Bencana Alam Yang Didapatkan Dari Postingan Twitter Terkait Bencana Alam	NER	0.872 & 0.902
14	Yuda Munarko	Data Tweet Yang Diambil Dari 100 Akun	NER & POS Tagger	P = 0.9289 R = 0.8385 F1 = 0.8814

Penelitian Lazqar Markus Oktavianto (2018), menerapkan sistem *biosurveilance* untuk melakukan klasifikasi dan pemetaan lokasi yang sedang terjangkit jenis penyakit yang disebabkan oleh virus dan bakteri. Data yang digunakan adalah *tweets* dengan kata kunci jenis penyakit yang memiliki jumlah sebanyak 1000 data *tweets*. Metode yang digunakan dalam penelitiannya yaitu Naïve Bayes dan Named Entity Recognition. Tujuan dari penelitiannya adalah memberikan kemudahan pada pengguna untuk mengetahui daerah yang terjangkit penyakit sesuai data *tweet* yang digunakan [16].

Penelitian Harold Situmorang (2015), melakukan klasifikasi wilayah di Medan yang positif sebagai daerah epidemi demam berdarah berdasarkan data kasus demam berdarah di kota Medan dari tahun 2010 sampai tahun 2013 yang didapatkan dari dinas kesehatan kota Medan menggunakan metode Naïve Bayes dengan tujuan dapat menunjukkan wilayah – wilayah di kota Medan yang menjadi pusat epidemi penyakit demam berdarah [18].

Penelitian Aries Setiawan dan Adi Prihandono (2019), melakukan klasifikasi tingkat kerentanan penyakit malaria pada suatu wilayah. Data yang digunakan pada penelitiannya adalah data variabel kerentanan malaria pada kabupaten Purworejo tahun 2011 yang kemudian akan dilakukan pengklasifikasian menggunakan metode Naïve Bayes. Hasil klasifikasi kerentanan malaria pada suatu wilayah tersebut akan digunakan untuk upaya menekan tingkat penularan penyakit malaria di suatu wilayah yang rentan malaria [19].

Penelitian Sandi Fajar Rodiyansyah, Edi Winarko (2012), melakukan klasifikasi kemacetan lalu lintas di kota Bandung menggunakan metode Naïve Bayes dengan sumber data dari postingan pada Twitter. Data Twitter yang digunakan adalah postingan pengguna dengan rentang waktu sejak 24 mei 2011 sampai 11 september 2011 dengan jumlah data sebanyak 15401. Hasil klasifikasi akan memberikan informasi jalan di kota Bandung dalam kondisi macet atau tidak macet yang kemudian dilanjutkan dengan visualisasi pada *Google Map* [20].

Penelitian Ziza Amira, Muhammad Nasrun, Casi Setianingsih (2018), melakukan klasifikasi kemacetan lalu lintas di kota Jakarta menggunakan metode K-Nearest Neighbor dengan sumber data yang ada pada Twitter. Data Twitter yang digunakan adalah postingan pengguna dengan rentang waktu sejak 01 september 2016 sampai 31 september 2017 dengan jumlah data sebanyak 1481. Hasil klasifikasi akan memberikan informasi jalan di kota Jakarta dalam kondisi lancar, padat, atau macet [21].

Penelitian Nelly Nur Hamidah, Ahmad Fathan Hidayatullah (2019), membuat sebuah sistem informasi pemetaan bencana yang terjadi di Indonesia dengan menggunakan data yang ada pada Twitter. Data Twitter yang telah dikumpulkan kemudian dilakukan pengklasifikasian dengan metode Naïve Bayes. Data Twitter yang digunakan adalah postingan yang ada pada akun BMKG dan BNPB. Hasil klasifikasi akan memberikan sebuah informasi mengenai bencana di seluruh Indonesia dan menampilkan bencana yang sering terjadi di tiap daerah di Indonesia [22].

Penelitian Kennedy Asfina dan Regina Justina (2017), melakukan klasifikasi penyakit demam berdarah dan tifus di filipina menggunakan metode SVM dengan sumber data dari Twitter. Data Twitter yang digunakan adalah postingan pengguna dengan rentang waktu sejak 10 agustus 2016 sampai 10 september 2016 dengan jumlah data sebanyak 1.931.561 dalam bahasa filipina dan 8.150.017 dalam bahasa inggris. Hasil klasifikasi berupa gejala-gejala yang menandakan terkena penyakit demam berdarah atau tifus. Hasil klasifikasi juga kemudian divisualisasikan ke dalam peta berdasarkan lokasi detail dari postingan yang dibuat oleh pengguna Twitter [23].

Penelitian Romy Ranovan, Afrizal Doewes, Ristu Saptono (2018), melakukan klasifikasi pemetaan penyakit tropis di Indonesia dengan menggunakan metode Multinomial Naïve Bayes dengan sumber data berasal dari Twitter. Data twitter yang digunakan adalah postingan pengguna dengan rentang waktu 23 mei 2016 sampai 22 november 2016 dengan jumlah data sebanyak 33.613 *tweet*. Pada metode ini peneliti melakukan dua klasifikasi yaitu untuk mengidentifikasi bahasa dan juga klasifikasi berita tentang wabah penyakit tropis. Hasil klasifikasi kemudian divisualisasikan ke dalam peta dan menunjukkan daerah yang terkena wabah penyakit tropis di Indonesia [24].

Penelitian Gigih Rezki Septianto, Firman Fakhri Mukti, Muhammad Nasrun, Alfian Akbar Gozali (2015), melakukan klasifikasi kemacetan lalu lintas di kota Jakarta menggunakan metode Naïve Bayes dengan sumber data yaitu dari postingan yang ada pada Twitter. Hasil klasifikasi akan memberikan informasi lokasi dan arah kemacetan lalu lintas di kota Jakarta dalam kondisi tersendat, padat merayap, padat, atau macet yang kemudian divisualisasikan ke dalam peta [25].

Penelitian Son Doan, QuocHung-Ngo, Ai Kawazoe, Nigel Collier (2008), melakukan pembuatan sistem berbasis web yang dapat mendeteksi dan memetakan wabah penyakit menular di dunia yang muncul di berita yang ada di internet. Agar sistem dapat melakukan pendeteksian dan pemetaan wabah penyakit menular dengan melakukan klasifikasi menggunakan metode Naïve Bayes. Hasil dari klasifikasi kemudian divisualisasikan ke dalam peta dan diberikan plot di tiap lokasi yang terjangkit wabah penyakit menular sesuai dengan berita [26].

Penelitian Sweta Swain, K.R. Seeja (2017), melakukan klasifikasi wabah penyakit demam berdarah dan juga chikungunya di New Delhi dengan metode Naïve Bayes. Data yang digunakan merupakan *tweets* yang terkait dengan demam berdarah, chikungunya, dan delhi lalu kemudian melakukan dua klasifikasi yaitu klasifikasi penyebaran wabah demam berdarah dan penyebaran klasifikasi wabah chikungunya. Hasil dari klasifikasi adalah informasi bulan yang paling tinggi terjadinya wabah penyakit demam berdarah dan chikungunya di New Delhi dalam bentuk grafik [5].

Penelitian Yuda Munarko (2015), melakukan ekstraksi informasi data twitter yang berasal dari tiga akun yaitu RTMC_Jatim, SbyTrafficServ, dan Radio Suara Surabaya. Jumlah data yang didapatkan sebanyak 3000 sampai 3500 dari tiap akun yang diambil sejak bulan april hingga bulan juni 2013. Hasil dari ekstraksi informasi dengan menggunakan Named Entity Recognition akan digunakan untuk mengidentifikasi nama lokasi dari tweet yang memberikan informasi tentang lalu lintas [27].

Penelitian John Lingad, Sarvnaz Karimi, Jie Yin (2013), melakukan eksperimental dalam melakukan ekstraksi informasi data dengan menggunakan Named Entity Recognition yang selanjutnya digunakan untuk menentukan lokasi yang ada pada sebuah *tweet*. Data yang digunakan merupakan postingan terkait bencana alam yang diambil dari Twitter. Jumlah data yang didapatkan sebanyak 3203 *tweets* diambil sejak akhir tahun 2010 hingga akhir tahun 2012. Hasil dari eksperimen yang dilakukan menunjukkan bahwa Named Entity Recognition dapat mengenali lokasi secara efektif dari sebuah *tweet* [28].

Penelitian Yuda Munarko (2016), melakukan sebuah uji coba dan analisis dari penelitiannya yaitu melihat peranan dari POS Tagger dalam proses ekstraksi informasi menggunakan NER. Data yang digunakan dalam penelitian adalah data twitter yang mewakili *tweet* dengan bahasa baku dan tidak baku. Data Twitter tersebut diambil dari 100 akun dengan 100 tweet dari tiap akun. Hasil dari uji coba yang dilakukan mendapatkan hasil bahwa penggunaan POS Tagger dalam proses NER berpotensi meningkatkan ketepatan hasil dari proses NER [29].

Dari beberapa penelitian yang sudah dijabarkan sebelumnya, maka untuk menyelesaikan penelitian yang dikerjakan penulis yaitu klasifikasi dan deteksi lokasi kasus demam berdarah di Indonesia berdasarkan *tweet* maka akan menggunakan metode *Naïve Bayes* dalam proses klasifikasi dikarenakan metode ini cukup mudah untuk diimplementasikan dan juga efisien dalam prosesnya sedangkan untuk proses pencarian lokasi penulis akan menggunakan metode *Named Entity Recognition*.