BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Untuk membedakan penelitian yang dilakukan dengan penelitian sebelumnya maka penulis melakukan studi pustaka yang telah dilakukan peneliti - peneliti terdahulu, adapun tinjauan pustaka terdahulu.

Tinjauan pertama penelitian yang dilakukan oleh Budi Kurniawan dkk (2017). Peneliti melakukan proses klasifikasi berita dengan data yang berasal dari twitter yang kemudian berita diklasifikasikan menjadi 8 yaitu ekonomi, entertainment, olahraga, teknologi, kesehatan, makanan, otomotif, travel. Pada proses klasifikasi dilakukan mulai preprocessing kemudian proses case folding, tokenizing, filtering dan stemming. Setelah Proses preprocessing selesai dilakukan perhitungan kemunculan kata dan probabilitas, kata yang muncul akan 1 sedangkan yang tidak akan diberi nilai 0. Setelah itu baru dilakukan klasifikasi. Jumlah data latih mempengaruhi performa, semakin banyak data latih semakin baik hasilnya.

Tinjauan yang kedua penelitian yang ditulis oleh Dio Ariadi dan Kartika Fithriasari (2015). Data yang diklasifikasikan adalah berita bahasa Indonesia dengan sumber data adalah artikel berita dari koran online kompas.com yang akan dibagi menjadi 12 kategori antara lain berita nasional,

internasional, olahraga, sains, edukasi, ekonomi, tekno, entertaintment, otomotif, health, properti, dan travel. Artikel diambil dari Januari hingga Desember 2014, yang kemudian data tersebut dibagi menjadi data *traning* dan data *testing* dengan porsi 70:30. Metode yang digunakan dalam proses klasifikasi adalah *Naive Bayesian Classification* dan *Suport Vector Machine.*

Tinjauan ketiga penelitian yang ditulis oleh Rusdi Efendi dkk (2012). Objek yang diteliti adalah dokumen berbahasa Indonesia dengan 30 dokumen latih dan 30 dokumen uji yang kemudian akan diklasifikasikan menjadi 6 katagori yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan.

Tinjauan yang keempat ditulis oleh Yuna Sophia Dewi Febriant pada tahun 2017. Objek yang diteliti adalah *tweets* yang ada di akun twitter STMIK Akakom yang kemudian diklasifikasikan dalam tweets positif atau negatif degan metode *Naive Bayes*.

Tinjauan yang kelima ditulis oleh Naufal Riza Fatahillah pada tahun 2017 . Dalam proses pencariannya didasarkan hastag tertentu yang selanjutnya diklasifikasikan negatif atau positif metode yang digunakan adalah *Naive Bayes*.

Pada penelitian ini, sistem akan melakukan klasifikasi terhadap berita yang diupload pada akun twitter Divis Humas Polri ke dalam tiga katagori yaitu berita kegiatan polisi, komentar masyarakat dan layanan masyarakat selama empat tahun ke belakang. Setelah berita diklasifikasi kemudian dicari sentimen dari

setiap topik. Metode yang digunakan adalah Naive Bayes Classifier.

Perbandingan dari penelitian - penelitian di atas dengan yang akan dibuat dapat dilihat pada tabel 2.1

Tabel 2.1 Tabel Perbandingan Penelitian

Penulis	Objek	Metode	Hasil
Budi Kurniawan, Mochammad Ali Fauzi, Agus Wahyu Widodo (2017)	Berita Twitter	Metode Improved Naïve Bayes	Aplikasi ini mampu mengkasifikasikan delapan kategori yaitu: ekonomi, entertainment, olahraga, teknologi, kesehatan, makanan, otomotif, travel. Data yang diambil berupa judul dari berita.
Dio Ariadi dan Kartika Fithriasari (2015)	Berita bahasa Indonesia	Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer	Berita dikategorikan sebagai berikut berita nasional, internasional, olahraga, sains, edukasi, ekonomi, tekno, entertaintment, otomotif, health, properti, dan travel. Tiap kategori akan diambil sebanyak 100 artikel sehingga data artikel keseluruhan berjumlah 1200. Hasil yang didapatkan pada saat data testing pada masing-masing pengukuran performa akurasi, precision, recall, dan F-Measure sebesar 82,2%; 83,9%; 82,2%; dan 82,4%.

Tabel 2.1 Lanjut

Nama	Objek	Metode	Hasil
Rusdi Efendi, Reza Firsandaya Malik, Jeni Mila Sari U (2012)	Dokumen Berbahasa Indonesia	Naive Bayes Classifier	Perangkat lunak pada penelitian ini telah dapat mengklasifikasikan dokumen berita dengan akurasi perangkat lunak ini mencapai 86,67% dengan menggunakan 60 dokumen yang terdiri dari 30 dokumen latih dan 30 dokumen uji untuk enam kategori berita yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan.
Yuna Sophia Dewi Febriant (2017)	Tweets pada akun Twitter STMIK Akakom	Naive Bayes Classifier	Aplikasi mampu menganalisa dan mengklasifikasikan sentimen yang ada pada twitter STMIK Akakom.
Naufal Riza Fatahillah (2017)	Tweet pada twitter	Naive Bayes Classifier	Mendeteksi tweets negatif atau positif berdasarkan hastag.
Septian Narsa Putra (2018)	Berita dan twet Divisi Humas Polri	Naive Bayes Classifier	Mengklasifikasikan berita berdasarkan topik kemudian dicari sentimen dari setiap topik.

2.2. Dasar Teori

2.2.1. Twitter

Twitter merupakan salah satu media sosial dengan layanan *microbloging* yang terkenal dan memungkinkan para penggunanya untuk menulis sesuatu atau yang biasa disebut tweet. Twitter digunakan untuk mengutarakan opini publik

maupun berita resmi dari suatu instansi atau dari akun kantor berita. Twitter dibangun oleh Jack Dorsey pada tahun 2006 dengan alamat http://www.twitter.com, jika seseorang ingin menggunakan twiiter seseorang harus terlebih dahulu memiliki akun, untuk registrasinya dapat dilakukan pada alamat tersebut.

Pengguna dapat mennulis pesan berdasarkan topik dengan tanda #(tagar). Sedangkan untuk menyebut atau membalas pesan dari pengguna lain bisa menggunakan tanda @(diikuti nama akun yang akan dibalas).

2.2.2. Twiiter API

API (*Application Programing Interface)* merupakan sekumpulan perintah, fungsi dan protokol yang dapat digunakan dalam membangun perangkat lunak untuk sistem operasi tertentu, juga merupakan suatu dokumentasi yang terdiri dari antar muka, fungsi, kelas, struktur untuk membangun *software*. Dengan API seorang *programmer* mengintegrasikan satu perangkat lunak dengan perangkat lunak lainnya.

Terdapat berbagai jenis sistem API yang dapat digunakan, termasuk sistem operasi, library, dan web

2.2.3. Text mining

Text Mining adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah besar sumber data tak terstruktur. Penambangan teks

memiliki tujuan dan menggunakan proses yang sama dengan penambangan data, namun memiliki masukan yang berbeda. Masukan untuk penambangan teks adalah data yang tidak (atau kurang) terstruktur, sedangkan masukan untuk penambangan data adalah data yang terstruktur (Ronen Feldman, 2006). *Text mining* dapat diterapkan untuk pelacakan topik, katagorisasi dan analisis sentimen.

Dalam proses penambangan teks terdapat proses pengubahan data menjadi datta yang terstruktur seesuai dengan kebutuhan, proses ini disebut *Text Prepocessing*, dengan tahapannya adalah Case Folding, Tokenizing, Filtering, Stemming, Tagging dan Analyzing.

Case Folding adalah proses mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” yang diterima dan karakter selain huruf dihilangkan. Tahapan *Tokenizing* adalah proses pemotongan kalimat (*string*) berdasarkan kata peyusunnya. Kemudian tahap *filtering*, merupakan tahapan mengambil kata - kata penting dari hasil token dan membuang sebagian kata tertentu. Tahapan *Steamming* merupakan tahapan pemetaan dan penuraian suatu kata menjadi kata dasarnya. Tahapan *Tagging* merupakan tahap memberikan label dari setiap kategori. Yang terakhir adalah tahap *analyzing* yaitu tahap untuk mencari seberaoa jauh keterhubungan antar kata - kata setiap dokumen.

2.2.4. Naive Bayes Classifier

Teorema Bayes adalah teorema yang digunakan dalam statistika untuk menghitung peluang suatu hipotesis. Bayes merupakan teknik prediksi berbasis *probalistic* sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi independensi yang kuat (Eko Prastyo, 2012). Naive Bayes merupakan salah satu contoh metode *supervised document classification* yang berarti membutuhkan data latih dalam melakukan klasifikasi.

Yang perlu dilakukan adalah menghitung nilai probabilitas masing - masing katanya dengan persamaan 1.

$$P(W_k|V_j)=\frac{n_k+1}{n+|kosakata|} \quad \text{.....(1)}$$

$P(W_k|V_j)$ = Peluang kategori j ketika terdapat kemunculan kata i

n_k+1 = Frekuensi kemunculan kata pada sebuah katagori

n = Jumlah seluruh kata pada dokumen dalam suatu kategori

$|kosakata|$ = Jumlah total kata (*distinc*) pada semua data latihan

Setelah itu, hitung probabilitas katagori dengan menggunakan persamaan 2.

$$P(V_j)=\frac{|docsj|}{|contoh|} \quad \text{.....(2)}$$

$P(V_j)$ = Probabilitas dokumen kategori

$|docsj|$ = Jumlah seluruh dokumen pada sebuah kategori

$|contoh|$ = Jumlah keseluruhan data yang dilatih

Pada proses klasifikasi, data yang diinputkan belum diketahui katagorinya. Pada proses ini, dilakukan tahap mencari kata - kata pada data yang diinputkan sesuai dengan pengetahuan data latih. Untuk mengklasifikasikan data yang diinputkan menggunakan persamaan 3.

$$V_{MAP} = \underset{vj \in V}{argmax}\, P(V_j)\, \Pi_i\, P(a_i|V_j) \quad \text{.....(3)}$$

Pada pengklasifikasian menggunakan Naïve Bayes dibagi kedalam 2 proses, yaitu proses training dan testing. Proses training digunakan untuk menghasilkan model analisis yang nantinya akan digunakan sebagai acuan untuk mengklasifikasikan data mentah yang baru.