# BAB 2

# TINJAUAN PUSTAKA DAN DASAR TEORI

## 2.1 Tinjauan Pustaka

Dalam sebuah penelitian diperlukan dukungan hasil-hasil penelitian sebelumnya yang berkaitan dengan penelitian saat ini. Berikut ini merupakan hasil dari penelitian yang pernah dilakukan :

**Tabel 2. 1 Tinjauan Pustaka**

| No | Peneliti | Objek | Metode | Hasil |
|---|---|---|---|---|
| 1 | Dyarsa Singgih Pamungkas, Noor Ageng Setiyanto, dan Erlin Dolphina (2015) | Sosial Media Twitter Terhadap Kata Kunci "Kurikulum 2013" | NBC | Sistem yang dapat mengklasifikasi sentimen secara otomatis dengan hasil pengujian 3000 data latih dan 100 tweet data ujicoba mencapai 91 %. |
| 2 | Nooraeni, dkk (2019) | Sistem zonasi sekolah | NBC | Prediksi sentimen masyarakat dengan akuransi 80,79%. |
| 3 | Thomas E Tarigan, Robby C Buwono, Sri Redjeki (2019) | Sosial Media Twitter Pada Perguruan Tinggi | NBC | Klasisifikasi data teks menjadi 3 kelas yaitu positif, negatif dan netral. Akuransi pengujian klasifikasi sebesar 75%. |
| 4 | Rizky Maulana, Sri | Sosial Media Twitter | *Support Vector Machine* | Hasil penelitian dengan 1.400 tweet pada dataset dan 200 |

| | | | | |
|---|---|---|---|---|
| | Redjeki (2016) | Terhadap Tokoh Publik | | data uji didapatkan akurasi sebesar 79,5%. |
| 5 | Irfangi (2019) | Sosial Media Twitter Terhadap Transportasi *Online* di Indonesia | NBC | Hasil uji akurasi pengujian dari 109 data dihasilkan sebesar 84%. |
| 6 | Septian Narsa (2019) | Berita dan *tweet* Divisi Humas Polri | NBC | Hasil analisis sentimen respon positif sebesar 55% untuk topik kegiatan kepolisian, 19,9% untuk topik komentar masyarakat dan 91,8% untuk layanan masyarakat. |

## 2.2 Dasar Teori

### 2.2.1 Perguruan Tinggi Yogyakarta

Yogyakarta menjadi salahsatu kota dengan Perguruan Tinggi yang banyak, mulai dari Universitas, Sekolah Tinggi, Politeknik hingga sekolah akademik. Berdasarkan pusatkampus.com terdapat 136 Perguruan Tinggi di Yogyakarta yang terdiri dari 11 Perguruan Tinggi Negeri dan sisanya adalah swasta.

Selain banyak, beberapa perguran tinggi di Yogyakarta juga menyandang status unggulan sebagai universitas dan institut serta berprestasi. Melalui peringkat Webometrics dengan sistem peringkat berdasarkan keberadaan web universitas, 10

peguruan tinggi di Yogyakarta dengan peringkat teratas dijadikan sebagai kata kunci untuk pengambilan data pada penelitian ini.

### 2.2.2 Analisis Sentimen

Analisis Sentimen masih bagian dari penelitian *opinion mining* merupakan salah satu bidang dari *Natural Languange Processing* (NLP) (Annisa, 2020) yaitu proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan pendapat tentang suatu masalah atau objek oleh seseorang, apakah mereka cenderung memiliki pandangan negatif atau pendapat positif (Thomas E Tarigan, 2019).

Tahapan yang umum dilakukan dalam analisis sentimen dengan *Machine Learning* (Septiar, 2019):

1. Data

Data yang dimaksud adalah data yang sudah dilabeli. Melakukan pelabelan terhadap data adalah usaha pertama yang besar untuk melakukan analisis sentimen.

2. *Preprocessing* Data

*Preprocessing* adalah suatu rangkaian langkah yang digunakan untuk menghasilkan dataset sesuai dengan kebutuhan. Pada kasus *Text Mining*, umumnya terdapat 5 rangkaian *preprocessing* yaitu *case folding, tokenizing, normalization, stopword removal,* dan *lemmalization*.

3. *Feature Extraction*

Komputer tidak dapat mengolah data selain data numerik, sehingga dibutuhkan langkah untuk mengekstrak data menjadi numerik dalam proses anasilis sentimen.

Secara umum terdapat 3 teknik ekstraksi fitur yaitu *Bag of Word (TF, IDF), Word Embedding (Glove, Word2vec, FastText)* dan *Character Embedding*.

4. *Modelling*

Pada tahap ini dilakukan pemodelan menggunakan metode yang telah ditentukan untuk dapat melakukan analisis sentimen.

5. *Evaluation*

Model yang terbentukan dari data pelatihan dievaluasi agar diketahui seberapa efektif model tersebut bekerja dalam menyelesaikan permasalahan. Evaluasai model biasanya menggunakan data uji untuk mengetahui seberapa prediktif model.

### 2.2.3 *Machine Learning*

*Machine Learning* adalah suatu kemampuan yang diberikan kepada komputer untuk belajar dari data tanpa diprogram secara eksplisit. Terdapat 2 tipe *machine learning* yang umum digunakan, diantaranya *Supervised Learning* dan *Unsupervised Learning*.

*Supervised learning* merupakan teknik pembelajaran yang diawasi dimana setiap data memiliki target kelas sedangkan *unsupervised learning* teknik pembelajaran yang tidak diawasi dimana data tidak memiliki target kelas.

### 2.2.4 *Naïve Bayes Classifier*

*Naïve Bayes Classifier* adalah model *Probabilistic Machine Learning* yang digunakan untuk tugas klasifikasi. Inti dari klasifikasi didasarkan pada teorema Bayes. Teorema Bayes adalah teorema yang digunakan dalam satatistika untuk menghitung peluang suatu hipotesis. Asumsi yang dibuat adalah fitur independen yaitu kehadiran satu fitur tidak mempengaruhi yang lain. (Eko Prasetyo, 2012)

Dalam klasifikasi menggunakan *Naïve Bayes* dibagi menjadi 2 proses, yaitu pelatihan dan proses pengujian. Proses pelatihan digunakan untuk menghasilkan model analisis sentimen yang nantinya akan digunakan sebagai referensi untuk mengklasifikasikan sentimen dengan data pengujian baru atau data mentah (Putra, 2019). Adapun persamaan dari Teorema Bayes dapat dilihat pada persamaan (1).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

B = Data dengan kelas yang belum diketahui

A = Hipotesis data B nerupakan kelas spesifik

P(A|B) = Probabilitas hipotesis A berdasarkan kondisi B(*posterior probability*)

P(A) = Probabilitas Hipotesis A(*prior probability*)

P(B|A) = Probabilitas B berdasarkan kondisi hipotesis A

P(B) = Probabilitas B

### 2.2.5 *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF bekerja dalam menentukan frekuensi relatif suatu kata kemudian dibandingan dengan proporsi kata tersebut pada seluruh dokumen (Robertson, 2004). Adapun rumus untuk menemukan bobot dari kata menggunakan TF-IDF adalah (Akhmad Pandhu Wijaya, 2016):

a. TF (*Term Frequency*)

*Term Frequency* adalah cara pembobotan *term* (kata) paling sederhana. Bobit kata t pada dokumen diberikan dengan :

$$w_{ij} = tf_{ij}.idf \quad (2)$$

b. IDF (*Inverse Document Frequency*)

Jika TF memperhatikan kemunculan kata dalam dokumen, IDF memperhatikan kemunculan kata pada kumpulan dokumen. Faktor IDF pada suatu kata t diberikan oleh:

$$idf = \log \frac{N}{df_i} \quad (3)$$

Dimana $w_{ij}$ adalah bobot kata *i* pada dokumen *j*, sementara N adalah jumlah dokumen dan *term frequency* adalah $tf_{ij}$ yaitu jumlah dari kemunculan kata *i* pada dokumen *j*, $df_i$ (*document frequency*) adalah jumlah dokumen *j* yang berisi kata *i*.

### 2.2.6 Akurasi

Metode evalusi dipergunakan untuk mengukur keakuratan hasil klasifikasi, digunakan perhitungan akurasi. Mengevaluasi banyaknya label prediksi yang sesuai dengan label actual. Semakin besar nilai akurasi, maka performansi *classifier* semakin bagus. (Akhmad Pandhu Wijaya, 2016)

$$Akurasi = \frac{Jumlah\ dokumen\ terklasifikasi\ dengan\ benar}{Jumlah\ dokumen\ keseluruhan}\ X\ 100$$