

## BAB II

### TINJAUAN PUSTAKA

#### 2.1 Tinjauan Penelitian Terdahulu

Tinjauan penelitian yang pertama adalah karya tulis dari Veronika Effendy tahun 2015 dengan judul Analisis Sentimen Berbahasa Indonesia Dengan Pendekatan Lexicon Based (Studi kasus: Solusi Pengolaan Sampah). Hasil dari penelitian ini, Klasifikasi dapat berjalan dengan cukup bagus dengan akurasi mencapai diatas 70% dan TPR diatas 85% bisa dikatakan bahwa pendekatan sentimen analis bisa diterapkan dalam pemilahan sentimen solusi dan bukan solusi, khususnya dalam kasus pengelolaan sampah. Dan juga akurasi ini sangat tergantung dengan kamus data, sehingga akurasi bisa diperbaiki dengan pengayaan kamus.[2]

Tinjaun penelitian terdahulu yang ke dua adalah karya tulis Ira Zulfa dan Edi Winarko tahun 2017 dengan judul Sentimen Analisis Tweet Berbahasa Indonesia Dengan *Deep Belief Network*. Hasil dari penelitian ini, dari pengujian pada sistem yang memperlihatkan dengan metode Deep Belief Network memberikan hasil pengujian klasifikasi lebih baik dengan akurasi sebesar 93.31%, presisi 93%, recall 93%, f1-score 93%, dan support 2378, sedangkan hasil pengujian pada sistem menggunakan metode Naive Bayes memberikan hasil pengujian klasifikasi dengan akurasi sebesar 79.10%, presisi 79%, recall 79%, f1-score 79%, support 2378 dan Support Vector Machine sebesar 92.18%, presisi 92%, recall 92%, f1-score 92% , support 2378.[3]

Tinjauan penelitian terdahulu yang ke tiga adalah karya tulis Rosa Delima tahun 2014 dengan judul Implementasi Metode K-Nearest Neighbor Dengan Decision rule untuk Klasifikasi Sub topik Berita. Hasil dari penelitian ini, Penggunaan K-Nearest Neighbor sebagai klasifikasi menunjukkan persentasi yang baik, dengan nilai  $k = 3$ , menunjukkan hasil persentase 88,29%. Dari  $k$  yang sama, digunakan Decision Rule yang ada dan persentase hasil akhir dari keakuratan K-Nearest Neighbor dengan Decision Rule adalah 89,36%.[4]

Tinjauan penelitian terdahulu yang ke empat adalah karya tulis Moh.Aziz Nugroho dan Heru Agus Santoso tahun 2016 dengan judul Klasifikasi Dokumen Komentar Pada Situs Youtube Menggunakan Algoritma K-Nearest Neighbor (K-NN). Hasil dari penelitian ini, Dengan Pendekatan Lexicon Based Hasil pengujian klasifikasi dokumen komentar pada situs youtube menggunakan metode K-NN ini mencapai nilai akurasi tertinggi yaitu dengan rata 0,806 dengan banyak dokumen training 240 menggunakan proses stemming. Sedangkan akurasi terendahnya adalah 0,669 dengan dokumen training 210 yang mempunyai porsi 120 dokumen berlabel positif dan 90 berlabel negatif. Untuk pengujian tanpa stemming memiliki rata-rata mencapai akurasi hingga 0,794.[5]

Tinjauan penelitian terdahulu yang ke lima adalah karya tulis Antonius Rahmat tahun 2016 dengan judul Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan metode Naive Bayes. Hasil dari penelitian ini, Algoritma Naive Bayes mampu memberikan klasifikasi

sentimen positif atau negatif dengan tingkat akurasi rata-rata mencapai tertinggi 82%. [6]

**Table 1.** Penelitian Terdahulu

No	Peneliti	Judul Penelitian	Perumusan Masalah	Hasil
1	Veronika Effendy (2015)	<p>Analisis Sentimen Berbahasa Indonesia Dengan Pendekatan Lexicon Based (Studi kasus: Solusi Pengolahan Sampah).</p>	<p>Seberapa besar akurasi dengan menggunakan Pendekatan Lexicon Based (Studi kasus: Solusi Pengolahan Sampah).</p>	<p>Klasifikasi dapat berjalan dengan cukup bagus dengan akurasi mencapai diatas 70% dan TPR diatas 85% bisa dikatakan bahwa pendekatan sentimen analisis bisa diterapkan dalam pemilahan sentimen solusi dan bukan solusi, khususnya dalam kasus pengelolaan sampah.</p>
2	Ira Zulfa	SentimenAnalisi	Membandingkan	Hasil dari

	dan Edi Winarko (2017)	s Tweet Berbahasa Indonesia Dengan <i>Deep Belief Network</i> .	akurasi antara Dengan <i>Deep Belief Network</i> dengan Naive Bayes	penelitian ini, dari pengujian pada sistem yang memperlihatkan dengan metode Deep Belief Network memberikan hasil pengujian klasifikasi lebih baik dengan akurasi sebesar 93.31%, presisi 93%, recall 93%, f1-score 93%, dan support 2378, sedangkan hasil pengujian pada sistem menggunakan metode Naive Bayes memberikan hasil
--	------------------------------	------------------------------------------------------------------------------	-------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

				<p>pengujian klasifikasi dengan akurasi sebesar 79.10%, presisi 79%, recall 79%, f1-score 79%, support 2378 dan Support Vector Machine sebesar 92.18%, presisi 92%, recall 92%, f1-score 92% , support 2378.</p>
3	Rosa Delima (2014)	<p>Implementasi Metode K-Nearest Neighbor Dengan Decision rule untuk Klasifikasi Subtopik Berita</p>	<p>Seberapa tingkat keakuratan K-Nearest Neighbor dengan Decision Rule untuk klasifikasi subtopic berita</p>	<p>Penggunaan K-Nearest Neighbor sebagai klasifikasi menunjukkan persentasi yang baik, dengan nilai k = 3, menunjukkan hasil</p>

				<p>persentase 88,29%. Dari k yang sama, digunakan Decision Rule yang ada dan persentase hasil akhir dari keakuratan K-Nearest Neighbor dengan Decision Rule adalah 89,36%</p>
4	<p>Moh. Azi Nugroho dan Heru Agus Santoso (2016)</p>	<p>Klasifikasi Dokumen Komentar Pada Situs Youtube Menggunakan Algoritma K-Nearest Neighbor (K-NN)</p>	<p>Seberapa baik penggunaan Algoritma K-Nearest Neighbor (K-NN) untuk Klasifikasi Dokumen Komentar Pada Situs Youtube</p>	<p>Dengan Pendekatan Lexicon Based Hasil pengujian klasifikasi dokumen komentar pada situs youtube menggunakan</p>

				<p>metode k-nn ini memiliki nilai accuracy tertinggi yaitu dengan rata 0,806 dengan jumlah dokumen training 240 menggunakan proses stemming. Sedangkan accuracy terendahnya adalah 0,669 dengan dokumen training 210 yang mempunyai porsi 120 dokumen berlabel positif dan 90 berlabel negatif. Untuk pengujian tanpa stemming</p>
--	--	--	--	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

				memiliki rata-rata accuracy hingga 0,794
5	Antonius Rachmat (2016).	Klasifikasi Sentimen Komentar Politik Facebook Page Menggunakan Naive Bayes	permasalahan yang dibahas pada tulisan ini adalah belum banyaknya sistem yang dapat mengklasifikasikan pro kontra dari komentar-komentar yang terdapat pada Facebook Page	Algoritma Naive Bayes mampu mengklasifikasikan sentimen dengan tingkat akurasi rata-rata tertinggi 82%.

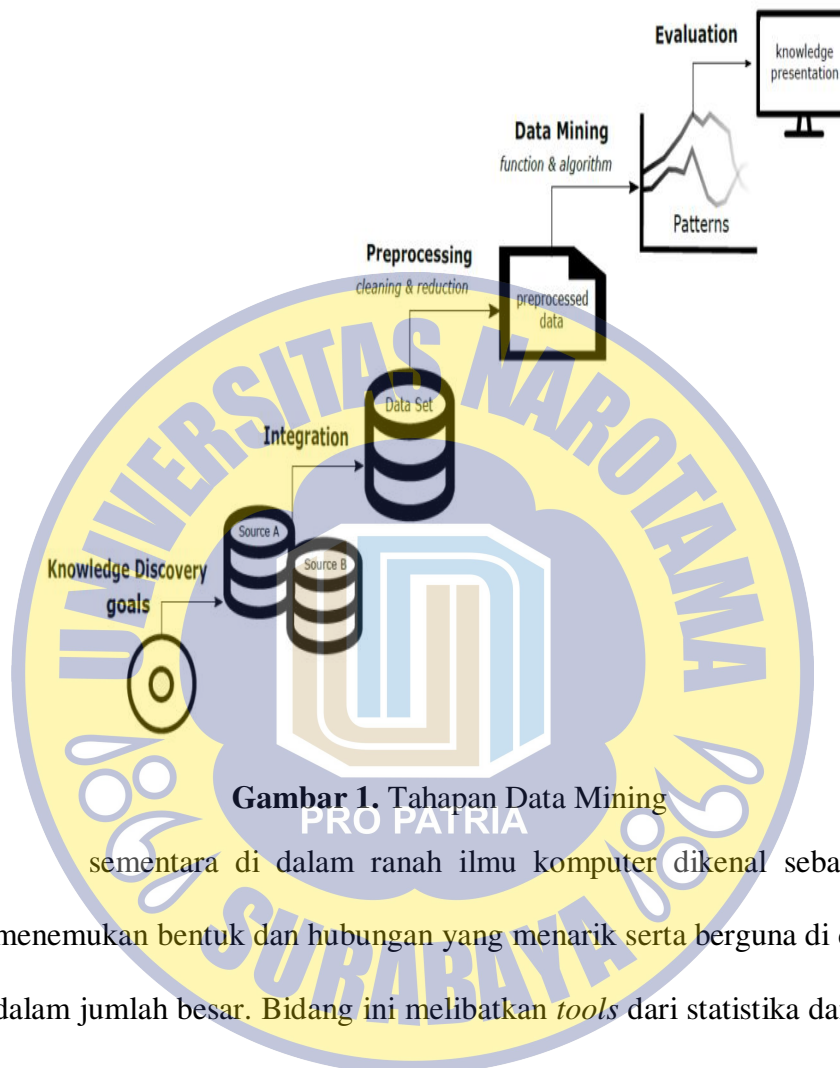
## 2.2 Teori Dasar yang Digunakan

### 2.2.1 Data Mining

*Data mining* adalah suatu proses untuk mengolah data mentah menjadi informasi yang bermanfaat. Saat ini *data mining* digunakan hampir dipakai pada semua aspek yang melibatkan data-data digital.[7] Hal ini yang kemudian menimbulkan kekhawatiran terkait keamanan dan etika di dalam pelaksanaannya, karena istilah *data mining* sendiri dapat di artikan



penggalian data, dan sering juga dikaitkan dengan penggalian data terkait perilaku manusia. Berikut gambaran tahapan pada data *mining*.



**Gambar 1.** Tahapan Data Mining

sementara di dalam ranah ilmu komputer dikenal sebagai proses menemukan bentuk dan hubungan yang menarik serta berguna di dalam data dalam jumlah besar. Bidang ini melibatkan *tools* dari statistika dan *Artificial Intelligence* (seperti *Neural Network* dan *Machine Learning*) serta manajemen sistem basis data (*database*) untuk menganalisis data digital berskala besar (disebut dengan istilah *dataset*), dan mungkin saja juga memakai perangkat pendukung keputusan yang lainnya.[8]

### 2.2.1 Twitter

Twitter adalah salah satu layanan media sosial yang cukup terkenal dan memungkinkan para penggunanya untuk menulis status yang sering

dinamakan kicauan atau *tweet*. Media sosial Twitter digunakan untuk mengungkapkan berbagai pendapat atau opini akan suatu produk, layanan atau hal lainnya. Twitter diciptakan oleh Jack Dorsey di tahun 2006 dan pertama meluncur di dunia maya saat Juli 2006 dengan alamat <http://www.Twitter.com> yang masih digunakan hingga saat ini. Pengguna dapat menulis pesan berdasarkan topik dengan menggunakan tanda #(hashtag). Sedangkan untuk menyebutkan atau membalas pesan dari pengguna lain bisa menggunakan tanda @.

### 2.2.3 *Python standard library*

*Python* memiliki referensi *library* untuk menjelaskan *standard library* yang didistribusikan bersama *Python*. *Standard library* menjelaskan beberapa komponen opsional yang biasanya disertakan dalam distribusi *Python*. Fasilitas ini memberikan kemudahan saat melakukan penulisan kode (*coding*), terdiri dari *built-in functions*, modul matematika, dan modul *input* atau *output* (I / O). Bahkan tersedia modul HTTP yang memudahkan untuk mendapatkan data dari internet.

### 2.2.4 *Text Mining*

*Text mining* merupakan bagian dari data mining dimana proses yang dilakukan utamanya adalah mengerjakan ekstraksi pemahaman dan informasi dari bentuk-bentuk yang terdapat dalam sekelompok dokumen teks menggunakan alat analisis tertentu.[9] *Text mining* bertujuan untuk mencari kata-kata yang dapat menggantikan dari dokumen kata-kata yang lain.

*Text Mining* adalah salah satu proses yang bisa digunakan untuk melakukan klasifikasi dokumen yang berusaha untuk menemukan pola yang menarik dari banyak jumlah data yang terkumpul. *Text Mining* merupakan analisis teks dimana data didapat dari dokumen yang bertujuan mencari kata-kata yang bisa mewakili dari dokumen, sehingga dapat dilakukan analisa keterhubungan, keterkaitan dan kelas antar dokumen. [10][7]

#### **2.2.5 Analisis Sentimen**

Analisis sentimen atau opinion mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk memperoleh informasi sentimen yang ada dalam suatu kalimat opini. Analisis sentimen dilaksanakan untuk melihat pendapat atau kecenderungan pandangan terhadap suatu masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini positif atau negatif (B. Liu. 2010).

#### **2.2.6 Naïve Bayes**

*Naïve Bayes* merupakan teorema bayes yang memiliki kemampuan hampir serupa dengan *decision tree* dan *neural network*. Teorema Bayes adalah teorema yang digunakan dalam untuk menghitung peluang suatu hipotesis. Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi independensi (ketidak tergantungan) yang kuat atau naïf.[10] Dengan kata lain, dalam Naïve Bayes, model yang digunakan adalah “model fitur independen”, maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur

lain dalam data yang sama. *Naïve Bayes* adalah suatu metode pengklasifikasian data dengan model statistik yang dapat digunakan untuk menghitung probabilitas keanggotaan suatu kelas.

Metode Bayesian classification digunakan menganalisis dalam membantu tercapainya pengambilan keputusan terbaik suatu permasalahan dari sejumlah alternatif. Kaitan antara *Naïve Bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi.[10]

Metode *Naïve Bayes* populer di dalam penerapan *machine learning* dikarenakan kesederhanaannya dalam memungkinkan setiap atribut berkontribusi terhadap keputusan akhir secara merata dan independen tanpa bergantung pada atribut lainnya. Kesederhanaan ini merupakan bentuk dari efisiensi komputasi, yang membuat teknik *Naïve Bayes* menarik dan sesuai di banyak area.[11][10] Namun, hal yang sama yang membuat *Naïve Bayes* populer, juga alasan mengapa beberapa peneliti menganggap pendekatan ini lemah. Asumsi independen bersyarat sifatnya kuat, dan membuat sistem berbasis *Naïve Bayes* tidak mampu menggunakan dua atau lebih keterangan bersamaan, namun jika digunakan di area yang sesuai, *Naïve Bayes* menawarkan pelatihan cepat, analisis data dan pengambilan keputusan yang cepat, serta interpretasi langsung terhadap hasil tes. Secara umum berikut

ini adalah persamaan *Naive Bayes* yang didasari oleh rumusan probabilitas bersyarat :

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c) \quad (2)$$

**Keterangan :**

$P(c|x)$  = Probabilitas yang dicari (*posterior probability*)

$P(c)$  = Probabilitas *class* berdasarkan hipotesis (*prior probability*)

$P(x|c)$  = Probabilitas prediktor berdasarkan *class* yang diberikan (*likelihood*)

$P(x)$  = Probabilitas prediktor

