BAB II
TINJAUAN PUSTAKA

Pada tahap ini peneliti akan menjabarkan penelitian terdahulu sebagai penunjang pengerjaan tugas akhir yang diperoleh dari skripsi dan jurnal yang berkaitan dengan penelitian.

2.1 Penelitian Terdahulu

Di bawah ini penelitian sebelumnya yang dijadikan acuan dalam tugas akhir.

Tabel 1. Literatur Review

No	Judul	Metode	Dataset	Hasil akurasi
1.	Analisis Sentiment Menggunakan Naïve Bayes Classifier Dan Inset Lexicon Pada Twitter (Studi Kasus Mie Gacoan)	Naïve Bayes & Inset Lexicon	756 tweet	Sebelum dan sesudah mendapat sertifikat halal sebesar 62% dan 65%
2.	Analisis Sentiment Data Tweet Terhadap Penanganan Covid-19 Di Indonesia Menggunakan Metode Naïve	Naïve bayes & Lexicon based	1020 tweet	75%

	Bayes Dan Pemelihan Kata Bersentimen Menggunakan Lexicon Based			
3.	Analisis Sentiment Terhadap Perkuliahan Daring Di Indonesia Dari Twitter Dataset Menggunakan Inset Lexicon	Naïve bayes & Inset Lexicon	5811 tweet	79.2%
4.	Analisis Sentiment Opini Pengguna Twitter Pada Aplikasi Bibit Menggunakan Multinomial Naïve Bayes	Multinomial naïve bayes & Lexicon Based	2764 tweet	88%
5.	Analisis Sentimen Dan Information Extraction	Lexicon	73.120 tweet	Positif 77,58% dan negatif 17,97%

	Pembelajaran Daring Menggunakan Pendekatan Lexicon			

Penelitian mengenai analisis sentimen di media sosial, khususnya Twitter, telah banyak dilakukan sebelumnya dengan berbagai metode dan pendekatan seperti tabel diatas . Pada jurnal 1 yang terbit pada tahun 2023 memiliki kelebihan dalam menggabungkan 2 metode yaitu machine learning naïve bayes dan pendekatan berbasis kamus lexicon based approach dan memiliki kekurangan dalam proses emoticon, bahasa asing, dan bahasa daerah [6]. Pada jurnal 2 yang terbit pada tahun 2022 memiliki kelebihan yang sama seperti pada jurnal satu dan memiliki kekurangan dalam menangani kata bernegasi [7]. Pada jurnal 3 yang terbit pada tahun 2021 memiliki kelebihan yang sama seperti pada jurnal yang disebutkan sebelumnya dan memiliki kekurangan dalam tahapan preprocessed data dalam menangani singkatan dari bahasa yang tidak formal dan kurangnya dalam menyaring tweet non opini [8]. Pada jurnal 4 yang terbit pada tahun 2023 memiliki kelebihan dalam penggunaan yang sama juga tapi lebih dijelaskan menggunakan multinomial naïve bayes dan memiliki kekurangan hanya menggunakan satu pendekatan lexicon based [9]. Sementara itu, pada jurnal 5 yang terbit pada tahun 2022 memiliki kelebihan menggunakan inset lexicon dan memiliki kekurangan tidak mengkombinasikannya dengan metode machine learning[10]. Penelitian-penelitian tersebut menunjukkan variasi metode dan akurasi dalam analisis sentimen, yang dapat memberikan wawasan berharga untuk penelitian lebih lanjut di bidang ini.

2.2. Generasi Z Dalam Dunia Kerja

Generasi Z, juga dikenal sebagai "Digital Natives", adalah kelompok orang yang lahir antara tahun 1996 hingga 2010. Mereka adalah generasi yang sejak lahir akrab dengan teknologi digital [2]. Generasi Z memiliki tujuan karir

yang unik yaitu, membangun berbagai karir sehingga mereka dapat melakukan banyak pekerjaan sekaligus. Generasi Z senang menerima umpan balik dari atasan tentang bagaimana mereka melakukan pekerjaan mereka. Generasi Z akan cenderung bertahan lama dalam pekerjaan jika terjalin hubungan personal yang kuat. Dalam pekerjaan sehari-hari Generasi Z sangat mengandalkan teknologi, mulai dari menyusun dokumen hingga berkomunikasi dengan klien melalui email dan pesan instan. Selain itu, mereka merasa nyaman dengan teknologi baru dan sering menggunakannya dalam pekerjaan. Kenaikan cepat dalam karier adalah prioritas bagi Generasi Z, baik dari segi pribadi maupun professional [3].

2.3 Analisis Sentimen

Bidang studi yang melakukan analisis opini, sentimen, evaluasi, penilaian, sikap, serta emosi individu pada topik atau kegiatan tertentu. Analisis sentimen ialah pendekatan untuk ditentukannya sentimen berdasarkan nilai polaritas teks dalam kalimat, sehingga kategori label diklasifikasikan sebagai sentimen positif, negatif, atau netral [6]**.**

2.3.1 Inset Lexicon

Dalam analisis sentimen, pendekatan lexicon juga dikenal sebagai pendekatan kamus dengan cara memeriksa dokumen yang sudah clean [11]. Metode ini menggunakan kamus yang dilengkapi bobot tiap kata sebagai sumber bahasa. Hasil analisis mencakup kategorisasi sentimen menjadi positif, negatif, serta netral. Kualitas produk dipengaruhi oleh kamus kata atau corpus yang digunakan [12]. Dengan menggunakan InSet (Indonesian Sentiment) Lexicon, berisi 3.609 kata positif serta 6.609 kata negatif, dengan bobot polarity score -5 dan +5 untuk masing-masing kata metode ini cukup sederhana untuk digunakan yaitu, bobot akan dihitung oleh sistem dan hasilnya akan diklasifikasikan ke dalam sentimen. Untuk menentukan nilai polaritas, digunakan persamaan sebagai berikut [10].

$$Sentiments_{score} = \sum_{i=1}^{n} Sentiments_{score} + W_{positive} + W_{negativ}$$

(2.1)

Dimana sentimen score ialah nilai polaritas opini Twitter, nilai yang didapat dari hasil penjumlah bobot positif serta negatif dalam kamus InSet lexicon. Polaritas sentimen ditentukan menggunakan persamaan berikut ini [13].

$$Sentiments_{score} \begin{cases} positif, & jika\ Sentiments_{score} > 0 \\ netral, & jika\ Sentiments_{score} = 0 \\ negatif, & jika\ Sentiments_{score} < 0 \end{cases}$$

(2.2)

Pemilihan lexicon inset ini didasarkan pada pertimbangan yang kuat sebagai berikut.

1. Keandalan dan Konsistensi: Kamus InSet menyediakan daftar kata yang telah dikategorikan dan diberi bobot berdasarkan polaritas sentimen yang diakui secara luas. Dengan menggunakan kamus ini, proses pelabelan sentimen menjadi lebih sistematis dan konsisten. Hal ini mengurangi subjektivitas yang mungkin muncul jika pelabelan dilakukan secara manual oleh manusia, karena setiap kata diberi bobot secara objektif sesuai dengan kamus.
2. Efisiensi Waktu dan Sumber Daya: Metode lexicon-based memungkinkan proses pelabelan data yang lebih cepat dibandingkan dengan pelabelan manual. Mengingat dataset yang digunakan dalam penelitian ini terdiri dari 1000 tweet, penggunaan kamus InSet membantu mengotomatisasi proses pelabelan, sehingga menghemat waktu dan sumber daya yang diperlukan untuk analisis.

3. Relevansi dengan Bahasa Indonesia: Kamus InSet dirancang khusus untuk bahasa Indonesia, sehingga sangat relevan dan sesuai untuk digunakan dalam analisis sentimen tweet berbahasa Indonesia. Ini memberikan keunggulan dibandingkan kamus sentimen yang mungkin dikembangkan untuk bahasa lain. Nuansa dan konotasi kata dalam bahasa Indonesia dapat ditangkap dengan lebih baik, sehingga analisis sentimen menjadi lebih akurat dan kontekstual. Metode ini tidak hanya memastikan bahwa proses pelabelan dilakukan secara objektif dan konsisten, tetapi juga membantu menghemat waktu dan sumber daya, sambil tetap menjaga akurasi dan relevansi hasil analisis sentimen.

Namun, metode lexicon memiliki beberapa kekurangan. Salah satunya adalah bahwa hasil klasifikasi sentimen bergantung pada kualitas nilai polaritas serta jumlah kata yang digunakan. Selain itu, metode ini tidak dapat memperoleh arti dari kata-kata yang memiliki arti yang dekat, yang berdampak pada penilaian polaritas [10].

2.4 Twitter

Twitter ialah media sosial favorit yang banyak untuk berinteraksi serta mendapatkan informasi. Twitter menawarkan jejaring sosial microblog atau dengan kata lain memungkinkan penggunan mengirim dan membaca pesan terbatas hanya 140 katakter [14]. Pengguna dapat menulis terkait topik serta membahas isu yang terjadi [15]. *User* twitter mengemukakan opini melewati tweet. Tiap tweet yang di posting *user* beraneka ragam sesuai kehendak mereka. Tweet berupa pendapat, saran, atau kritik terkait topik tertentu. Twitter merupakan media yang cukup baik dalam mendapat data pendapat masyarakat terhadap suatu topik [8].

2.5 Ekstrasi Fitur TFIDF

Term Weighting ialah proses dihitungnya bobot tiap kata sehingga diketahui kesamaan suatu kata pada dokumen [16]. Salah satu metode ekstrasi

fitur terbanyak dipergunakan ialah *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF ialah statistik angka yang menerangkan bahwasanya seberapa penting sebuah kata dalam dokumen. Metode ini dapat digunakan diberbagai bidang subjek termasuk peningkasan teks serta klasifikasi teks. Metode ini sering dipergunakan sebagai faktor pembobot dalam text mining dengan cara memberikan bobot hubungan suatu kata terhadap dokumen. Pembobotan kata guna diberikannya nilai pada suatu kata untuk dijadikan input pada proses klasifikasi [6]. TF-IDF terdiri atas 2 statistik yakni TF serta IDF. TF didefinisikan sebagai berapa kali istilah muncul dalam dokumen.

$$tf_{x,d} = count\ (x, d)$$

(2.3)

Keterangan :

- x : kata
- d : dokumen
- $tf_{x,d}$: frekuensi banyaknya kata *x* yang muncul dalam dokumen *d*

IDF adalah bobot statistik yang digunakan mengukur pentingnya istilah dalam sekumpulan dokumen teks. Fitur IDF digabungkan untuk meminimalisir bobot istilah yang sering muncul dalam kumpulan dokumen serta ditingkatkannya bobot istilah yang jarang muncul.

$$idf_x = log\ \frac{N}{d_x}$$

(2.4)

Keterangan :

- N : Total dokumen

- d_x : Dokumen yang mengandung kata x
- idf : Frekuensi banyaknya kata *x* yang muncul didokumen *d*

Kemudian TF-IDF dihitung untuk tiap kata mempergunakan rumus.

$$W_{x,d} = tf_{x,d}\ x\ idf_{x,d}$$

(2.5)

Keterangan :

- $tf_{x,d}$: Frekuensi banyaknya *x* kata yang muncul dalam dokumen *d*
- $idf_{x,d}$: Frekuensi banyaknya *x* kata yang muncul di dokumen *d*
- $W_{x,d}$: nilai TF-IDF (bobot dokumen ke-*d* terhadap kata ke-*x*)

2.6. Naïve Bayes

Dalam text mining, Naive Bayes adalah metode klasifikasi yang dipergunakan untuk analisis sentimen. Ini ialah salah satu metode klasik terkenal yang telah banyak digunakan untuk kategorisasi teks yang memiliki struktur sederhana dan tingkat efektifitasnya yang tingggi sehingga digunakan dalam data mining. Gambaran umum klasifkasi naïve bayes ditunjukan dibawah ini [17].

1.Menghitung jumlah kelas/label

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

(2.6)

Keterangan :

- $P(c|x)$: *Posterior* ialah peluang kelas *c* ketika terdapat kemunculan kata *x*
- $P(w|c)$: *Likehood* ialah peluang sebuah kata *x* terhadap kelas *c*
- $P(c)$: Prior ialah peluang munculnya kelas *c*
- $P(w)$: Evidence ialah peluang munculnya kata *x*

2. Menghitung jumlah kasus per kelas

Perhitungan likehood mempergunakan model multinomial naïve bayes yang menghitung nilai frekuensi munculnya tiap kata yaitu dokumen yang memuat kata yang diolah mempergunakan distribusi multinomial. Model multinomial memiliki kelemahan dalam perhitungan model yaitu, dimana terdapat kata yang tidak pernah muncul menjadi penyebab perhitungan dengan nilai nol. Penggunaan laplace smoothing memecah permasalahan dengan ditambahkannya nilai 1 pada kata sehingga dianggap pernah muncul sekali serta ditambahkannya kata unik. Untuk rumusnya di bawah ini.

$$P\ (wi|c) = \frac{count(wi, c) + 1}{count(c) + |v|}$$

(2.7)

Keterangan :

- $P\ (wi|c)$: Likehood, peluang kata ke-I kelas c.
- $count(wi, c)$: Jumlah kata ke-i yang ada pada suatu kelas *c*.
- $count(c)$: Jumlah semua kemunculan kata kelas *c*.
- $|v|$: Jumlah kemunculan term unik seluruh kelas.

3.Menghitung prior

$$P(c) = \frac{Nc}{N}$$

(2.8)

Keterangan :

- $P(c)$: Peluang kemunculan kelas c.
- Nc : Jumlah dokumen latih pada kelas c.
- N : Jumlah semua dokumen latih yang digunakan.

4.Menghitung Evidence

$$P(w) = \frac{|x|}{|S|}$$

(2.9)

Keterangan :

- $P(x)$: Evidence ialah jumlah kata x yang muncul .
- $|S|$: Jumlah seluruh kemunculan kata pada semua dokumen.

2.7 Evaluasi Performa

2.7.1 Confusion Matrix

Sebuah tabel yang berisi jumlah data uji benar serta salah dilakukan klasifikasi. Confusion matrix menjadi pengukuran yang sangat popular digunakan saat memecahkan masalah klasifikasi. Metode ini dapat diterapkan untuk klaisifikasi biner serta untuk masalah klasifikasi multikelas. Metode ini mendefinisikan berbagai metric kinerja seperti accuracy, precision, recall dan f-1 score [6].

Tabel 2 .Confusion Matrix

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:

- *True Positive* (TP) : Jumlah dokumen yang benar serta diklasifikasikan sebagai kelas 1.

- *True Negative* (TN) : Jumlah dokumen yang benar serta diklasifikasikan sebagai kelas 0.
- *False Positive* (FP) : Jumlah dokumen yang salah serta diklasifikasikan sebagai kelas 1.
- *False Negative* (FN) : Jumlah dokumen yang salah serta diklasifikasikan sebagai kelas 0.

Pada Penelitian ini , penulis menggunakan multiclass confusion matrix 3x3 dikarenakan terdapat 3 kelas sentimen yakni positif, negatif serta netral. Sehingga tabel confusion matrix yang ada pada penelitian ini yakni.

Tabel 3. Confusion Matrix Multiclass

		PREDIKSI		
		Positif (A)	**Negatif (B)**	**Netral (C)**
AKTUAL	**Positif (A)**	TPos (AA)	FPosNeg (AB)	FPosNet (AC)
	Negatif (B)	FNegPos (BA)	TNeg (BB)	FNegNet (BC)
	Netral (C)	FNetPos (CA)	FNetNeg (CB)	TNet (CC)

Keterangan :

- TPos : Jumlah prediksi positif dari data aktual positif.
- FPosNeg : Jumlah prediksi negatif dari data aktual yang positif.
- FPosNet : Jumlah prediksi netral dari data aktual positif.
- FNegPos : Jumlah prediksi positif dari data aktual negatif.
- TNeg : Jumlah prediksi negative dari data aktual negatif.
- FNegNet : Jumlah prediksi netral dari data aktual negatif.
- FNetPos : Jumlah prediksi positif dari data aktual netral.

- FNetNeg : Jumlah Prediksi negatif dari data actual netral.
- TNet : Jumlah prediksi netral dari data aktual netral.

2.7.2 Classification Report

Dari hasil tabel confusion matrix sebelumnya maka dapat di identiifkasi berbagai matrix evaluasi kinerja dibawah ini [13].

1. Accuracy

   Accuray merupakan visualisasi keakuratan model dalam mengelompokan dengan benar.

   Rumus accuracy ialah $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

   (2.10)

   Sedangkan Pada multiclass confusion matrix 3x3 dapat dilihat pada rumus dibawah ini.

$$accuracy = \frac{TPos + TNeg + TNet}{TPos + FPosNeg + FPosNet + FNegPos + TNeg + FNegNet + FNetPos + FNetNeg + TNet}$$

   (2.11)

2. Precision

   Precission merupakan visualisasi dari persentase keakuratan model untuk memprediksi nilai positif. Rumus dari precision dapat dilihat dibawah ini.

$$precision = \frac{TP}{TP + FP}$$

   (2.12)

3. Recall

   Recall merupakan visualisasi kesesuaian metode dalam pencarian ulang informasi yang bernilai positif. Rumus dari reccal dapat dilihat dibawah ini.

$$recall = \frac{TP}{TP + FN}$$

   (2.13)

4. F1-Score

F1-Score membandingkan antara rata-rata nilai presisi serta recall dari hasil pengujian.Rumus F1-Score dapatdlihat dibawah ini.

$$f1 - score = 2 * \frac{precision * recall}{precision + recall}$$

(2.14)

2.8 K-Fold Cross Validation

Cross validation ialah teknik validasi model yang dipergunakan mengevaluasi seberapa baik hasil analisis statistik yang dapat digeneralisasi ke kumpulan data independen. Teknik ini dipakai membuat prediksi model serta untuk memperkirakan keakuratan model prediksi saat dijalankan dalam eksperimen. Teknik cross validation adalah k-fold cross validation yang membagi data menjadi k-set data berukuran sama yang digunakan untuk menghilangkan bias dalam data. Untuk melakukan pengukuran kinerja klasifikasi, yakni membandingkan semua data uji yang diklasifikasikan dengan benar dengan jumlah data uji. Dengan rumus sebagai berikut [6].

$$akurasi = \frac{\sum klasifikasi\ benar}{data\ uji\ x\ 100\%}$$

(2.15)

Selain itu *standard deviation* atau simpang baku juga akan dihitung, yang merupakan ukuran disebarkannya data dan ditunjukkannya jarak rata-rata median ke titik nilai. Makin besar simpang baku yang didapat, maka penyebaran median juga semakin besar, begitu juga sebaliknya. Tujuan perhitungan simpang baku ialah untuk dilihatnya jarak antar rata-rata akurasi dengan akurasi tiap percobaan. Dengan rumus sebagai berikut.

$$\sigma = \sqrt{\frac{\sum(x_{i-\mu})^2}{N}}$$

(2.16)

Keterangan :

- N : Banyak percobaan
- μ : Mean
- X : Percobaan ke-i
- i : Indek x tiap percobaan