BAB II
TINJAUAN PUSTAKA

A. Peneliti Terdahulu

Penelitian yang dilakukan oleh (Irawanto et al., 2023) dengan judul "*Sentiment Analysis And Classification Of Forest Fire In Indonesia*" memiliki tujuan untuk menganalisis sentimen bagaimana masyarakat merespon isu kebakaran hutan di Indonesia melalui data yang diambil dari twitter dengan membandingkan metode *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine* untuk mengetahui perbedaan akurasi ketiganya ketika menggunakan data yang sama. Hasil penelitian menggunakan metode *Random Forest* mendapatkan akurasi sebesar 59%, menggunakan metode *Naïve Bayes* mendapatkan akurasi sebesar 69%, dan menggunakan metode *Support Vector Machine* mendapatkan akurasi sebesar 66%. Penelitian ini memiliki beberapa keterbatasan yaitu data yang didapatkan masih kurang maksimal karena pemilihan kata kunci yang digunakan pada saat proses *crawling* data dari *twitter* yang hanya mendapatkan 650 *tweet* dan tersisa 285 *tweet* setelah dilakukan *preprocessing*, pada tahap normalisasi kosakata yang digunakan masih kurang lengkap sehingga masih ada beberapa kata *slang* yang belum diganti dengan kata baku, dan pada proses pelabelan menggunakan leksikon VADER masih mungkin terjadi kesalahan yang dapat mempengaruhi hasil analisis.

Penelitian yang dilakukan oleh (Syah & Witanti, 2022) memiliki tujuan untuk menganalisis sentimen opini publik terhadap kebijakan pemerintahan mengenai vaksinasi *covid-19* di Indonesia, dengan data yang digunakan diambil dari *twitter* yang terdiri dari 4.708 *indeks tweet* yang telah di proses. Hasil analisis menunjukkan bahwa 83,6% respon positif dan 16,4% respon negatif. Metode yang diterapkan dalam penelitian ini adalah *Support Vector Machine* yang menghasilkan akurasi sebesar 89%, *F1 score* sebesar 93%, *precision score* sebesar 88%, dan *recall score* sebesar 99%. Keterbatasan dalam penelitian ini kurangnya dukungan dari *library* yang dapat menghitung polaritas dengan data *text* dalam bahasa Indonesia,

sehingga menyebabkan penulis mengalami kendala dalam proses *text processing* dan pada tahap *preprocessing* tidak ada normalisasi, sehingga kata-kata yang tidak baku tidak diubah menjadi kata yang lebih baku.

Penelitian yang dilakukan oleh (Tineges et al., 2020) yang memiliki tujuan untuk membangun model klasifikasi sentimen menggunakan metode *Support Vector Machine* (SVM) untuk menganalisis opini pengguna terhadap layanan indihome, yang merupakan salah satu penyedia layanan internet di Indonesia. Penelitian ini menggunakan data dari *twitter* yang dikumpulkan sejak 16 maret 2020 hingga 22 maret 2020, dengan fokus pada *tweet* berbahasa Indonesia. Hasil penelitian menunjukkan bahwa model SVM yang diterapkan berhasil mengklasifikasikan sentimen pengguna dengan akurasi mencapai 87%, *precision* sebesar 86% dan *recall* sebesar 95%. Penelitian ini mengidentifikasikan opini menjadi opini positif dan negatif. Keterbatasan dalam penelitian ini yaitu terbatasnya jumlah data karena hanya mengambil data dalam rentang waktu yang singkat dan hanya mencakup *tweet* yang menyebut akun tersebut, sehingga tidak sepenuhnya mewakili keseluruhan opini pengguna di *platform twitter*.

Penelitian yang dilakukan oleh (Arsi & Waluyo, 2021) memiliki tujuan untuk menganalisis sentimen publik terkait wacana pemindahan ibu kota Indonesia menggunakan algoritma *Support Vector machine*. Data yang digunakan dalam penelitian ini menggunakan data yang bersumber dari *twitter* yang terdiri dari 1.236 *tweet*, yang menghasilkan 404 *tweet* dengan sentimen positif dan 832 *tweet* dengan sentimen negatif. Hasil penelitian menunjukkan bahwa penerapan SVM berhasil mencapai akurasi sebesar 96,68%, dengan *precision* 95,82%, *recall* 94,04%, dan AUC 0,979 yang menunjukkan efektivitas metode ini dalam mengklasifikasikan sentimen. Keterbatasan dalam penelitian ini mencakup perlunya kajian lebih mendalam mengenai penanganan negasi dalam analisis sentimen dalam bahasa Indonesia, sehingga sentimen yang mengandung kata negasi belum dapat diidentifikasi polaritasnya dengan optimal. Selain itu, tidak adanya

normalisasi pada tahap teks *processing,* sehingga data *tweet* yang memiliki kata tidak baku tidak diubah.

Penelitian yang dilakukan oleh (Manwombreidy Kafiar & Supatman, 2024) bertujuan untuk untuk menganalisis sentimen menggunakan algoritma *Support Vector Machine* mengenai isu pembabatan hutan adat di Papua melalui kata kunci tagar #AllEyesOnPapua di media sosial X. Data yang digunakan dalam penelitian ini dikumpulkan menggunakan teknik crawling data. Hasil penelitian ini menggunakan metode *Support Vector Machine* mandapatkan akurasi sebesar 67%, dimana pada sentimen positif didapatkan nilai *precision* sebesar 73%, *recall* 64%, dan *f1-score* 68%, pada sentimen negatif didapatkan nilai *precision* sebesar 56%, *recall* 54%, dan *f1-score* 55%, dan pada sentimen netral didapatkan nilai *precision* sebesar 61%, *recall* 79% dan *f1-score* 73%. Keterbatasan dalam penelitian ini adalah waktu untuk pengambilan data yang terlalu singkat yaitu hanya dalam rentang 25 Mei sampai 20 Juni 2024, sehingga jumlah data yang berhasil dikumpulkan menjadi terbatas.

Penelitian yang dilakukan oleh (Hasibuan et al., 2024) bertujuan untuk menganalisis sentimen masyarakat mengenai kebijakan ekspor pasir laut di Indonesia, terutama setelah kebijakan yang mengizinkan *ekspor* tersebut kembali diterapkan pada tahun 2023, yang sebelumnya dilarang sejak tahun 2003 denngan menggunakan algoritma Sup*port Vector Machine*. Pengumpulan data diambil melalui *twitter* dengan menggunakan *tweet* yang relevan dengan kata kunci ‘ekspor pasir laut’ dan menghasilkan data sebanyak 856 tweet dalam rentang waktu 14 juni hingga 8 september 2023. Namun, hanya 661 data yang digunakan karena 195 data tidak menunjukkan kecenderungan positif atau negatif. Hasil analisis menunjukkan bahwa sentimen negatif sebesar 50,5%, sentimen positif sebesar 25,25% dan sentimen netral sebesar 24,5%. Hasil pengujian menggunakan algoritma SVM menunjukkan hasil yang baik mencapai akurasi sebesar 80,94%, *precision* 80,91%, *recall* 80,94% dan *f-measure* 80,75%. Keterbatasan dalam penelitian ini meliputi jumlah data yang

terbatas, sehingga tidak sepenuhnya mewakili opini publik secara keseluruhan. Selain itu, penelitian ini juga hanya menggunakan satu sosial media untuk pengambilan datanya, sehingga tidak dapat memberikan wawasan tambahan mengenai opini masyarakat.

Penelitian yang dilakukan oleh (Shofiya & Abidi, 2021) memiliki tujuan untuk menganalisis data *twitter* dengan menggunakan pendekatan *Support Vector Machine* (SVM) dan untuk memahami sentimen warga Kanada tentang wacana seputar jarak sosial terkait COVID-19. Data yang digunakan dalam penelitian terdiri dari 629 *tweet* yang dikumpulkan dalam rentang 20 maret - 20 april 2020. Hasil dari penelitian ini menunjukkan bahwa 40% orang Kanada menunjukkan sentimen netral terhadap pembatasan sosial, 35% yang menunjukkan sentimen negatif, dan hanya seperempat orang kanada yang bersikap positif terhadap pembatasan sosial. Akurasi yang didapatkan menggunakan algoritma SVM sebesar 87%. Penelitian ini memiliki beberapa keterbatasan, meliputi ukuran dataset mempengaruhi kinerja algoritma SVM penambahan data pelatihan dapat meningkatkan akurasi, hilangnya 20% data karena proses hidrasi *ID tweet* ke CSV, dan kesalahan teknis yang menyebabkan hilangnya *tweet* pada periode tertentu, 40% *tweet* bersifat netral dapat menjadi salah satu alasan berkurangnya akurasi, serta performa *SentiStrength* tidak dievaluasi.

B. Landasan Teori

1. Analisis sentimen

   Analisis sentimen merupakan proses memahami, mengekstrak, dan mentransformasikan data teks secara otomatis untuk memperoleh informasi dengan tujuan melihat kecenderungan opini terhadap suatu objek, apakah cenderung beropini positif, negatif, atau netral (Parlika et al., 2020). Analisis sentimen juga dikenal untuk menganalisis dan mengukur sikap serta opini seseorang terhadap berbagai aspek, seperti topik, produk, peristiwa dan lain-lain dengan cara mengidentifikasi pola emosi dan pandangan yang ekspresikan dalam teks (Undap et al., 2021). Analisis sentimen termasuk dalam cabang ilmu yang melibatkan *text*

mining, pemrosesan bahasa alami, dan kecerdasan buatan (Bei & Sudin, 2021).

2. *X*

   X merupakan media sosial gratis dan terpopuler menyediakan layanan jaringan yang memungkinkan pengguna untuk berbagi pemikiran atau opini melalui pesan singkat yang sering disebut dengan *tweet* (Hendrastuty et al., 2021). *X* adalah *platform* media sosial *microblogging* yang memberi pengguna ruang untuk menulis dan membagikan aktivitas atau opini mereka. *X* menyediakan ruang dengan batas maksimal 280 karakter. Pengguna *X* dapat berinteraksi, berbagi informasi, mendukung pandangan pengguna lain, serta membahas isu terhangat (*trending topic*) dengan membuat *tweet* dan menggunakan hashtag tertentu (Undap et al., 2021).

3. *Crawling*

   Crawling merupakan teknik yang secara otomatis digunakan untuk mengumpulkan informasi dari web. Proses ini dilakukan berdasarkan kata kunci yang ditentukan oleh pengguna. Alat yang digunakan untuk melakukan *crawling* disebut *crawler*, yaitu program yang dirancang menggunakan algoritma tertentu untuk memindai situs web sesuai dengan alamat atau kata kunci yang ditentukan pengguna (Putra et al., 2020). *Crawling* juga dapat dilakukan dengan memanfaatkan indeks informasi pada halaman web melalui *Uniform Resource Locator* (URL) dan menggunakan *Application Programming Interface* (API) untuk memperoleh dataset dalam jumlah besar (Kosasih et al., 2022).

4. Python

   Python merupakan bahasa pemrograman tingkat tinggi yang berorientasi pada objek dan dikembangkan oleh Guido van Rossum (Junaidi et al., 2023). Python adalah bahasa pemrograman yang popular untuk analisis data, karena *python* dapat dipelajari dan digunakan oleh semua kalangan usia. Selain itu, bahasa pemrograman *python* memiliki beragam *library* yang masing-masing dapat digunakan oleh siapa saja

di berbagai sistem operasi atau dengan kata lain bersifat *open source* (Ua et al., 2023).

5. Deforestasi

Deforestasi adalah istilah yang menggambarkan penghilangan kawasan hutan. Deforestasi terjadi saat area hutan ditebang dan dialihfungsikan untuk kegiatan lain. Istilah lain yang sering digunakan untuk menyebut deforestasi adalah penggundulan hutan. Proses ini biasanya melibatkan perubahan penggunaan lahan untuk tujuan seperti pertanian, peternakan, atau pemukiman (Ika Febryanti et al., 2023). Perubahan fungsi hutan menjadi lahan non-hutan ini memicu pemanasan global karena seringnya terjadi kebakaran hutan. Deforestasi juga berkaitan dengan aktivitas penebangan atau pembalakan liar yang mengancam semua makhluk hidup, terutama karena kebakaran hutan yang berkontribusi pada pemanasan global (Nakita & Najicha, 2022).

6. *VADER (Valence Aware Dictionary and Sentiment Reasoner)*

VADER (*Valence Aware Dictionary and Sentiment Reasoner*) merupakan pendekatan berbasis *lexicon* yang dikembangkan oleh CJ Hutto dan Eric Gilbert dari Georgia Institute of Technology untuk mengklasifikasikan informasi dari teks menjadi kategori sentimen positif, negatif, atau netral. VADER melakukan klasifikasi dengan memberikan nilai pada setiap kata dalam teks, Penilaian ini didasarkan pada angka yang ditemukan oleh Hutto, C.J and Gilbert melalui penelitiannya yang melibatkan manusia sebagai penilai (Maulana et al., 2023). *Lexicon* VADER dibangun menggunakan data dalam bahasa Inggris, sehingga mengharuskan proses *translate* data ke dalam bahasa Inggris sebelum digunakan (Juli et al., 2024).

Kelebihan VADER ini adalah adanya kamus yang telah dilengkapi dengan nilai untuk setiap kata, sehingga pengklasifikasian dapat dilakukan secara otomatis oleh penggunanya. Metode VADER didasarkan pada fakta bahwa nilai diberikan pada setiap kata dalam teks berdasarkan penilaian manusia, bahkan makna tersirat dari tanda baca

dalam teks juga dapat diidentifikasi oleh VADER (Fauziah et al., 2024). Penentuan label sentimen ditentukan berdasarkan nilai *compound score* yaitu, jika skor bernilai ≥ 0,05 masuk dalam kategori sentimen positif, jika skor > -0,05 dan < 0,05 masuk dalam kategori sentimen netral, dan jika skor bernilai ≤ -0,05 masuk dalam kategori sentimen negatif (Aminullah et al., 2024). Perhitungan compound score dapat dilihat pada persamaan (2.1) (Rahmada & Qoiriah, 2024).

$$compound\ score = \frac{x}{\sqrt{x^2+\alpha}} \quad (2.1)$$

Keterangan:

x : Jumlah dari sentimen skor unsur kata dari kalimat

α : Parameter normalisasi yaitu 15 (default)

7. TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode algoritma yang menghitung bobot setiap kata dalam teks berdasarkan frekuensi dan distribusi kata tersebut di seluruh dokumen. TF atau (*Term Frequency*), menunjukkan seberapa sering sebuah kata muncul dalam satu dokumen, sedangkan IDF (*Inverse Document Frequency*) adalah nilai *invers* dari dokumen yang mengandung kata tersebut. TF dan IDF akan dikalikan sehingga menghasilkan nilai bobot dari kata tersebut (Fikri et al., 2020). Persamaan TF-IDF dapat dilihat pada persamaan (2.2), (2.3), (2.4).

$$TF - IDF_{(d,t)} = TF_{(d,t)} * IDF_{(t)} \quad (2.2)$$

$$TF_{(d,t)} = \frac{jumlah\ kata\ t\ pada\ dokumen\ d}{total\ kata\ pada\ dokumen\ d} \quad (2.3)$$

$$IDF_{(t)} = log\ \frac{total\ dokumen}{jumlah\ dokumen\ yang\ mengandung\ kata\ t} \quad (2.4)$$

Keterangan:

t = kata

d = dokumen

8. *Support Vector Machine*

Support vector machine (SVM) adalah algoritma dalam pembelajaran mesin dan didasarkan pada prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperplane* optimal yang dapat memisahkan dua kelas dalam ruang input (Gifari et al., 2022). SVM sebagai salah satu algoritma *machine learning* dapat memberikan hasil yang baik dalam pengklasifikasian (Bei & Sudin, 2021). Keunggulan lain dari SVM yaitu kemampuannya untuk menangani data dimensi tinggi dan dengan adanya ruang kernel hanya data tertentu yang dipilih untuk mengklasifikasikan model (Permata Aulia et al., 2021). Rumus persamaan SVM (Adams et al., 2021) dapat dilihat pada persamaan (2.5) dan (2.6).

$$d(x) = w.x + b \tag{2.5}$$

Atau

$$d(x) = \sum_{i=1}^{l} d_j\, y_i K(X, X_i) + b \tag{2.6}$$

Keterangan:

w : Vektor bobot

x : Titik data

$d_j y_i$: Bobot pada nilai di setiap titik data

$K(X, X_i)$: Fungsi kernel

b : Nilai bias

Kernel linear merupakan kernel yang digunakan dalam penelitian ini, dimana persamaan *linear* dapat dilihat pada persamaan (2.7).

$$K(x, y) = x.y \tag{2.7}$$

Keterangan:

K(x, y) : Nilai *kernel* dari data x dan y

x : Nilai fitur data 1

y : Nilai fitur data 2

9. *Confusion Matrix*

Metode yang banyak digunakan untuk mengevaluasi hasil penelitian adalah *confusion matrix*. *Confusion matrix* berfungsi untuk menilai

kinerja model klasifikasi dengan menunjukkan jumlah prediksi yang benar dan salah (Ridwan, 2020). Cara kerja metode ini adalah dengan membandingkan matriks dari prediksi dengan kelas asli serta prediksi nilai klasifikasi (Hasibuan et al., 2024). *Confusion matrix* dalam penelitian ini menggunakan empat kondisi, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Murtopo et al., 2024). Berikut tabel *confusion matrix* ditunjukan pada Tabel 2.1.

Tabel 2. 1 Confusion Matrix

Actual Data	*Predicted (Positive)*	*Predicted (Negative)*
Actual (Positive)	*TP (True Positive)*	*FN (False Negative)*
Actual (Negative)	*FP (False Positive)*	*TN (True Negative)*

Keterangan:

- TP (*True Positive*): hasil prediksi dan aktual keduanya menunjukkan kelas positif.
- FP (*False Positive*): prediksi menunjukkan bahwa kelas positif, namun aktualnya menunjukkan kelas negatif.
- FN (*False Negative*): prediksi menunjukkan bahwa kelas negatif, namun aktualnya menunjukkan kelas positif.
- TN (*True Negative*): hasil prediksi dan aktual keduanya menunjukkan kelas negatif.

Dalam *confusion matrix*, metrik yang digunakan untuk mengevaluasi kinerja model klasifikasi meliputi akurasi, *precision*, *recall*, dan *F1 score*, dirumuskan dalam persamaan berikut:

a. Akurasi

Akurasi menunjukkan persentase prediksi yang benar dari total prediksi yang dilakukan oleh model. Perhitungan akurasi dapat dilihat pada persamaan (2.8).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.8)$$

b. *Precision*

Precision mengukur ketepatan prediksi positif dari model, yaitu seberapa banyak prediksi positif yang dibuat oleh model sesuai dengan kenyataannya sebagai positif. Perhitungan *precision* dapat dilihat pada persamaan (2.9).

$$Precision = \frac{TP}{TP+FP} \tag{2.9}$$

c. *Recall*

Recall merupakan metrik yang menunjukkan seberapa besar proporsi dari seluruh sampel positif yang berhasil dikenali secara benar oleh model sebagai positif. Perhitungan *recall* dapat dilihat pada persamaan (2.10).

$$Recall = \frac{TP}{TP+FN} \tag{2.10}$$

d. *F1-score*

F1-Score memberikan penilaian yang seimbang dari kinerja model dengan menghitung rata-rata harmonis antara *precision* dan *recall*. Perhitungan *F1-score* dapat dilihat pada persamaan (2.11).

$$F1\text{-}score = 2x\frac{recall\ x\ precision}{recall + precision} \tag{2.11}$$