BAB II
TINJAUAN PUSTAKA

2.1 Peneliti Terdahulu

Penelitian terdahulu berperan penting dalam memperkuat landasan teori serta menjadi referensi dalam pemilihan metode dan pendekatan pada penelitian ini. Temuan dari studi sebelumnya juga memberikan perbandingan yang relevan terhadap hasil yang diperoleh dalam penelitian ini.

Tabel 2. 1 Peneliti Terdahulu

No	Penulis (Tahun)	Kontribusi
1	H. M. Lee et. al (2023) [12]	Topik: Perbandingan antara model deep learning IndoBERTweet dan algoritma machine learning Support Vector Machine (SVM) dalam analisis sentimen pada data Tweet mengenai pembangunan sirkuit balap di Indonesia Metode: IndoBERTweet, Support Vector Machine Dataset: Data tersebut dikumpulkan dari Indonesia selama periode Januari hingga September 2022 dengan menggunakan kata kunci "vaksin booster". Hasil: Hasil penelitian menunjukkan bahwa IndoBERTweet lebih unggul dibandingkan SVM dalam analisis sentimen, dengan akurasi 86% serta skor presisi, recall, dan F1 di atas 88%, sementara SVM hanya mencapai akurasi sekitar 82%, sehingga IndoBERTweet terbukti lebih efektif untuk menganalisis sentimen tweet terkait pembangunan sirkuit balap di Indonesia.

2	N. Hadi et. al (2025) [19]	Topik: Analisis Sentimen Masyarakat terhadap Pembangunan Ibu Kota Negara (IKN) melalui Data Media Sosial X menggunakan Metode SVM, Logistic Regression, dan Naïve Bayes Metode: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes Dataset: Dataset yang digunakan dalam penelitian ini berasal dari data media sosial X (Twitter) dengan kata kunci "IKN". Hasil: Hasil penelitian menunjukkan bahwa SVM dengan fungsi RBF memberikan akurasi pengujian tertinggi sebesar 80% dalam analisis sentimen terkait pembangunan IKN, mengungguli Logistic Regression dan Naïve Bayes yang masing-masing mencapai 79%. Secara keseluruhan, sentimen masyarakat cenderung seimbang, dengan dominasi sentimen netral.
3	M. Kurmasih et. al [18]	Topik: Analisis sentimen terhadap opini masyarakat mengenai patriarki di media sosial menggunakan metode Naïve Bayes Metode: Naïve Bayes Dataset: Data opini yang dikumpulkan melalui proses crawling di media sosial Twitter dengan menggunakan kata kunci terkait patriarki. Hasil: Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memiliki akurasi 92,50% dalam analisis sentimen terkait patriarki, namun hanya mampu mengenali opini negatif (recall 100%) dan gagal mengidentifikasi opini positif (presisi 0%).

4	A. D. Maulana et. al (2023) [20]	Topik: Analisis sentimen terhadap tweet-tweet terkait tragedi Kanjuruhan menggunakan metode IndoBERTweet, dengan perbandingan terhadap algoritma Naive Bayes Dataset: Dataset yang digunakan dalam penelitian ini berupa data komentar tweet dari media sosial Twitter yang berkaitan dengan tragedi Kanjuruhan. Metode: IndoBERTweet, Naïve Bayes Hasil: IndoBERTweet unggul dalam analisis sentimen tragedi Kanjuruhan dengan akurasi 88% dan F1-score 84%, jauh lebih baik dibandingkan Naive Bayes yang hanya mencapai akurasi 62% dan F1-score 59%.
5.	N. Fitriyah et. Al (2020) [21]	Topik: Analisis sentimen terhadap gojek pada media sosial twitter dengan Support Vector Machine (SVM). Dataset: Dataset yang digunakan dalam penelitian ini dari media social twitter, diambil dari kata kunci “gojek”. Metode: Support Vector Machine (SVM) Hasil:

Pada penelitian-penelitian sebelumnya juga telah mengeksplorasi berbagai pendekatan pembelajaran mesin, dari yang metode klasik hingga metode *transfer learning,* yaitu salah satunya yang dilakukan Nur dan Dedy (2025) melakuka analisis sentiment masyarakat terhadap pembangunan Ibu Kota Negara (IKN) dengan menggunakan data dari media social X, dengan menggunakan tiga algoritma yaitu, Support Vector Machine (SVM), Logistic Regression, Naïve Bayes. Proses ini dilakukan untuk menguji dan membandingkan keefektifan masing-masing algoritma dari metode klasik,

dalam mengklasifikasikan positif, negative, dan netral. Dibandingkan dengan algoritma lainnya, Support Vector Machine adalah yang paling efektif dan akurat untuk menganalisis sentimen masyarakat terkait pembangunan IKN. Hasil pengujian menunjukkan bahwa SVM memiliki akurasi tertinggi sebesar 80%, sedangkan Logistic Regression dan Naive Bayes masing-masing memiliki akurasi sekitar 79% [19]. Dibalik hasil performa yang baik dari penelitian terdapat kekurangan juga pada pelabelan data dilakukan otomatis menggunakan Inset Lexicon Indonesia, yang mungkin kurang akurat dalam menangkap konteks, ironi, atau bahasa daerah dan slang dalam tweet. Hal ini dapat mempengaruhi keakuratan klasifikasi sentimen akhir

Selain itu, pada penelitian yang dilakukan Hanvito dan Yulian (2023) membandingkan performa antara model IndoBERTweet dan Support Vector Machine (SVM) dalam melakukan analisis sentiment terhadap tweet yang berkaitan dengan pembangunan sirkuit balap di Indonesia. Hasil penelitian menunjukkan bahwa IndoBERTweet memiliki performa yang lebih baik dibandingkan SVM. Secara umum, IndoBERTweet mencapai akurasi sebesar 86% dengan precision 88.2%, recall 88.6%, dan F1-score 88.4%. Sedangkan SVM memperoleh akurasi sebesar 82%, precision 87.3%, recall 84.3%, dan F1-score 85.8% [19]. Selain itu, IndoBERTweet mampu mencapai akurasi tertinggi hingga 94% dalam beberapa iterasi, sedangkan SVM mencapai 93%. Walaupun hasil pengujian mendapatkan hasil yang baik, tetapi dari segi data masih terbatas, yang mungkin mempengaruhi generalisasi model terhadap data yang lebih luas di kehidupan nyata.

Dari penjelasan penelitian-penelitian sebelumnya, belum ada yang melakukan analisis sentimen terkait topik fatherless dengan menggunakan metode IndoBERTweet dan Support Vector Machine (SVM). Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi baru dalam ranah analisis sentimen berbahasa indonesia, khususnya dalam mengangkat isu sosial yang relevan namun masih jarang dibahas secara mendalam. Selain itu, penggunaan kombinasi metode IndoBERTweet dan SVM akan menunjukkan efektivitas

metode ini dalam mengkategorikan sentimen pada teks berbahasa Indonesia yang informal, seperti unggahan di media sosial.

2.2 Analisis Sentimen

Analisis sentimen adalah teknik berbasis pemrosesan bahasa alami (NLP) yang digunakan untuk menemukan dan mengukur emosi positif, negatif, dan netral dalam teks. Sentimen ini menunjukkan pendapat atau perasaan penulis tentang topik tertentu. Sektor seperti perusahaan, pemerintah, dan lembaga pendidikan dapat memperoleh informasi tentang cara meningkatkan layanan, kebijakan, atau penelitian dengan melihat data teks seperti ulasan produk atau komentar media social [10].

2.3 Media Sosial X

X atau yang sebelumnya disebut Twitter, ialah *platform* media sosial yang memungkinkan penggunanya untuk menulis, membaca, dan berpartisipasi dalam diskusi tentang hal-hal baru dan tren. Pengguna sering menggunakan X sebagai platform untuk mengungkapkan perasaan mereka tentang berbagai hal, seperti pujian, kritik, dan kritik [22]. Pujian, kritik maupun celaan itu dapat digunakan sebagai data untuk melihat tren atau topik. Media sosial X dapat digunakan untuk memahami perspektif masyarakat tentang suatu topik karena opini yang ada di platform tersebut dianggap sangat akurat [23].

2.4 Term Frequency – Inverse Document Frequency (TF-IDF)

Dalam representasi data, setiap kata dalam setiap *tweet* akan diberi nilai boolean yang menunjukkan ada atau tidaknya kata tersebut dalam *tweet* tersebut. Tiap *tweet* diasumsikan memiliki sejumlah kata yang dapat mewakili tweet tersebut. Untuk menemukan kata representasi, harus menghitung berat kata-kata yang ada [24].

Nilai *tf-idf* dihitung dengan membandingkan frekuensi kata dalam teks (term fequency) dengan frekuensi kata di seluruh dokumen (inverse document frequency) [25]. Nilai tf-idf dapat diperoleh dengan menggunakan persamaan (1):

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

Dimana t adalah kata atau *term*, d adalah kalimat, dan D adalah dokumen atau kumpulan kalimat. Dalam analisis sentimen, kata "baik" memiliki nilai *tf-idf* yang tinggi untuk dokumen berlabel positif, kata "baik" dianggap paling relevan dan dapat digunakan sebagai kata kunci, sedangkan untuk dokumen berlabel negative, kata "kecewa", yang memiliki nilai *tf-idf* yang tinggi, dianggap paling relevan dan dapat digunakan sebagai kata kunci [25].

2.5 IndoBERTweet

IndoBERTweet adalah model bahasa yang dilatih sebelumnya khusus untuk bahasa Indonesia di platform X. IndoBERTweet merupakan model berskala besar pertama yang dibuat untuk bahasa Indonesia di platform X, dilatih dengan memperluas BERT monolingual Indonesia. Fokus utama adalah penyesuaian model yang efektif di bawah ketidakcocokan kosakata. Para peneliti juga membandingkan berbagai teknik untuk menginisialisasi lapisan embedding BERT untuk jenis kata baru. Mereka menemukan bahwa pra-pelatihan dengan embedding subword BERT rata-rata lima kali lebih baik daripada teknik yang disarankan untuk adaptasi kosakata secara ekstrinsik [26].

Model ini memiliki tiga lapisan feed-forward berdimensi 3.072, dua belas lapisan tersembunyi dengan dimensi 768, dan dua belas kepala perhatian. Jumlah token tertinggi, 128 token, berbeda dengan panjang dokumen media sosial rata-rata. IndoBERTweet memiliki 31.984 token (VIBT), dengan 14.584 token (46%) merupakan kata baru dan 17.400 token (54% dari total) berasal dari IndoBERT. IndoBERTweet adalah model pretrained pertama yang dirancang khusus untuk teks berbahasa Indonesia di media sosial dan Twitter [27]. Berikut Ini adalah rumus pemecahan kata IndoBERTweet.

$$\mathrm{E}_{IBT}(x) = \frac{1}{|T_{IB}(x)|} \sum_{y \in T_{IB}} E_{IB}(y)$$

Keterangan:

$\mathrm{E}_{IBT}(x)$	embedding baru untuk token x dalam IndoBERTweet
$T_{IB}(x)$	himpunan subword dari token x dalam IndoBERTweet
$\lvert T_{IB}(x)\rvert$	jumlah subword dari token x dalam kosakata IndoBERTweet
$E_{IB}(y)$	embedding dari subword y dalam IndoBERT
$\sum_{y \in T_{IB}} E_{IB}(y)$	jumlah embedding dari semua subword y yang membentuk x diambil dari embedding IndoBERT

2.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah model yang terkenal karena tingkat akurasinya yang tinggi dalam melakukan klasifikasi data. SVM adalah algoritma klasifikasi yang kuat, terutama untuk data dengan dimensi tinggi. Dengan mencari hyperplane optimal yang memaksimalkan margin antara kelas, SVM dapat mengklasifikasikan data dengan akurasi tinggi. Dalam analisis sentimen, SVM sering digunakan untuk memisahkan teks menjadi sentimen positif dan negative [24]. SVM Hyperplane dinyatakan dengan persamaan berikut.

$$w \cdot x + b = 0$$

Keterangan :

w : Vektor bobot
x : Vektor data (fitur input)
b : Bias (intersep)

2.7 Augmentasi Data

Salah satu jenis augmentasi yang biasa digunakan untuk data teks yaitu, teknik augmentasi *Synonym Replacement* di mana sebagian besar kata dalam sebuah kalimat (kecuali stopwords) dipilih secara acak dan digantikan dengan sinonim yang dipilih secara acak juga. Tujuan teknik ini adalah untuk mempertahankan makna asli kalimat sambil memperkenalkan variasi kata-kata [28]. Hal ini memungkinkan model untuk belajar dari berbagai bentuk ekspresi yang berbeda tetapi tetap memiliki makna yang sama.

Selain itu, teknik augmentasi data sederhana yang digunakan dalam penelitian ini termasuk Easy Data Augmentation (EDA), yaitu: penggantian kata dengan sinonim *(synonym replacement)*, pengacakan urutan kata *(random swap)*, penyisipan kata secara acak *(random insertion)*, dan penghapusan kata secara acak *(random deletion)* [29]. Fungsi utama EDA, menggabungkan keempat metode ini untuk menghasilkan kalimat baru dari input teks asli.