BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Penelitian yang dilakukan oleh Wulandari dkk (2021) yang mengkaji analisis sentimen media sosial *Twitter* terhadap reaksi masyarakat terhadap RUU Cipta Kerja menggunakan metode klasifikasi algoritma *Naïve Bayes* ditemukan bahwa penelitian ini didasarkan pada kerangka teoritis analisis sentimen publik di media sosial. Metode *Naïve Bayes* dipilih karena tingkat akurasi yang tinggi dalam mengklasifikasikan sentimen, dan penelitian ini melibatkan tahapan *Preprocessing* data, pengolahan data, klasifikasi dan evaluasi. Hasil penelitian menunjukkan bahwa rasio *training* dan *testing* 75:25 menghasilkan akurasi tertinggi sebesar 88%, dengan hasil prediksi yang lebih dominan pada label positif. Performa model yang dibuat menggunakan algoritma *Naïve Bayes* memiliki nilai *precision* sebesar 92%, *recall* sebesar 84%, dan *F1-score* sebesar 86%. Oleh karena itu penelitian ini memberikan landasan yang kuat untuk menyusun kerangka pikir, hipotesis penelitian dan metode penelitian dalam proposal skripsi ini.

Selain itu juga, penelitian dilakukan oleh Augustia dkk (2021) terkait analisis sentimen *Omnibus law* pada *Twitter* dengan algoritma klasifikasi berbasis *Particle Swarm Optimization*, ditemukan bahwa media sosial *Twitter* menjadi wadah untuk mengekspresikan dan mengapresiasi pendapat secara *online*. *Omnibus law* yang merupakan undang-undang cipta kerja yang ditetapkan oleh pemerintah Indonesia telah menjadi topik yang hangat di media sosial *Twitter*. Masyarakat memberikan beragam tanggapan baik dukungan maupun penolakan terhadap

undang-undang tersebut. Penelitian ini bertujuan untuk memisahkan sentimen positif dan sentimen negatif serta mengukur pendapat terhadap *Omnibus law* menggunakan algoritma klasifikasi. Sentimen positif dan negatif dikumpulkan melalui proses *crawling*. Algoritma klasifikasi yang digunakan adalah *Support Vector Machine* (SVM), *Naïve Bayes* (NB) dan *Particle Swarm Optimization* (PSO) digunakan untuk meningkatkan akurasi. Hasil pengujian menggunakan metode *K-fold Cross Validation* menunjukkan akurasi sebesar 84,95% dan 87,53% dengan nilai *Area Under the Curve* (AUC) masing-masing 0,958 dan 0,754. Selain itu pengujian menggunakan metode SVM dan NB secara terpisah menghasilkan akurasi sebesar 86,53% dan 90,12% dengan nilai AUC masing-masing 0,948 dan 0,816. Temuan-temuan ini memberikan landasan yang kuat dalam menyusun kerangka teoritis, kerangka pikir serta metode penelitian untuk proposal skripsi ini.

Pada penelitian lainnya yang dilakukan oleh Pane dkk (2021) terkait analisis sentimen UU *Omnibus law* pada *Twitter* menggunakan metode *Support Vector Machine* ditemukan bahwa media sosial *Twitter* menjadi platform yang memungkinkan individu secara bebas memberikan opini dan *tweet* yang bermanfaat bagi pengguna lainnya. Namun, dalam memberikan opini penting bagi masyarakat untuk dapat membedakan apakah opini tersebut bersifat positif, negatif atau netral. Permasalahan yang dihadapi adalah kurangnya sistem yang dapat memberikan sentimen secara otomatis dalam konteks tertentu. Oleh karena itu penelitian ini bertujuan untuk mengembangkan sistem yang dapat memberikan sentimen secara otomatis agar masyarakat dapat mengetahui opini yang bersifat positif, negatif atau netral terkait UU *Omnibus law*. Dalam analisis sentimen ini, metode yang

digunakan adalah *Support Vector Machine*, yang merupakan metode pengklasifikasian *supervised learning* yang mampu membedakan opini positif, negatif dan netral. Penelitian ini menggunakan bahasa pemrograman *Python* dan data yang diambil berasal dari *Twitter* sebanyak 150 data. Tahapan penerapan metode *Support Vector Machine* meliputi pengambilan data opini masyarakat Indonesia tentang UU *Omnibus law* melalui proses *scraping* kemudian dilanjutkan dengan tahap *Text Preprocessing* dan *Feature Extraction*. Penelitian ini menghasilkan akurasi sebesar 83% menggunakan teknik *K-fold Cross-Validation*, sehingga hasil yang diperoleh dapat dianggap cukup akurat.

Pada penelitian yang dilakukan oleh Tyo dkk (2021) menggunakan metode analisis sentimen dengan menggunakan metode *Naïve Bayes* dan *Relevance Frequency Feature Selection*. Dataset yang digunakan adalah 300 data opini masyarakat di *Twitter* terkait *Omnibus law*, dengan pembagian data menggunakan *k-fold cross validation* dengan k=5. Proses analisis sentimen terdiri dari *pre-processing* untuk pemrosesan opini, *Relevance Frequency Feature Selection* untuk mengurangi jumlah fitur, dan klasifikasi menggunakan metode *Naïve Bayes*. Hasil dari pengujian sebanyak 5 pengujian menggunakan klasifikasi *Naïve Bayes*, diperoleh rata-rata akurasi sebesar 62,6%, sementara hasil pengujian akurasi klasifikasi dengan penambahan *Relevance Frequency Feature Selection* diperoleh rata-rata akurasi sebesar 65,3%.

Dalam Penelitian lainnya yang dilakukan oleh Fanny dkk. (2022) menyimpulkan bahwa Klasifikasi *Naïve Bayes* dapat digunakan untuk mengklasifikasikan sentimen pengguna *Twitter* terhadap *Omnibus law*. Penelitian

ini menemukan bahwa metode *Naïve Bayes Classifier* memberikan hasil pengujian klasifikasi dengan akurasi yang tinggi untuk kedua kategori *Hashtag* Pro dan *Hashtag* Kontra. Nilai akurasi rata-rata untuk *Hashtag* Pro adalah 92,1%, dengan nilai presisi rata-rata 94,8% dan nilai recall rata-rata 90,7%. Sedangkan untuk *Hashtag* Kontra, nilai akurasi rata-rata adalah 98,3%, dengan nilai presisi rata-rata 97,6% dan nilai recall rata-rata 98,7%. Penelitian ini juga menemukan bahwa kata yang dominan muncul dalam analisis *text cloud* adalah "*Omnibus law*", menunjukkan bahwa semua hashtag dalam scrap membahas topik utama *Omnibus law*. Penelitian ini menggunakan bahasa pemrograman *Python* dan *web scraping*.

Temuan-temuan ini menjadi dasar yang relevan dalam menyusun kerangka teoritis. Berdasarkan penemuan-penemuan tersebut, *Naïve Bayes* digunakan untuk menganalisa sentimen yang dilakukan pada penelitian ini. Penelitian ini memiliki perbedaaan dengan penelitian lainnya dalam hal penanganan ketidakseimbangan kelas yang menggunakan SMOTE dan pembagian data yang dimana terdapat data uji dan data validasi yang bertujuan untuk mendeteksi jika terjadi *overfitting*, dengan menggunakan tehnik tersebut klasifikasi sentimen memiliki akurasi pada data uji 90,27% dan pada data validasi memiliki akurasi 89,92%. Dengan hasil tersebut memberikan informasi bahwa model berjalan secara baik dan tidak terdeteksi *overfitting*.

Tabel 2. 1 Perbandingan Penelitian

No	Penulis	Objek	Metode	Hasil
1	Wulandari, D.A., dkk (2021)	RUU Cipta Kerja	*Naïve Bayes*	Akurasi 88% dengan sentimen dominan positif
2	Augustia, A.E., dkk. (2021)	*Omnibus law*	*Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) dengan optimasi *Particle Swarm Optimization* (PSO)	Akurasi SVM: 84,95% dan AUC: 0,958, Akurasi NB: 87,53% dan AUC: 0,754
3	Pane, S.F., dkk(2021)	UU *Omnibus law*	*Support Vector Machine* (SVM)	Akurasi 83%
4	Tyo, K. V. S., dkk(2021)	Opini Masyarakat mengenai Kebijakan New Normal	*Naive Bayes* dengan *Relevance Frequency Feature Selection*	*Naïve Bayes* akurasi rata-rata 62,6% dan rata-rata akurasi *Relevance Frequency Feature Selection* yakni 65,3%
5	Fanny, O., dkk (2022)	*Omnibus law*	*Naïve Bayes*	Akurasi *Hashtag* Pro 92,1% dan *Hashtag* Kontra 98,3%
6	Suara, L.D.J., (2023)	*Omnibus law*	*Naïve Bayes*	Akurasi data uji 90,27% dan data validasi 89,92%

2.2. Dasar Teori

2.2.1. *Omnibus law*

Omnibus law atau Rancangan Undang-Undang (RUU) Cipta Kerja adalah sebuah rancangan metode untuk membuat suatu regulasi atau undang-undang yang terdiri atas banyaknya subjek atau materi pokok tertentu yang bertujuan agar menyimpangi suatu norma peraturan (Mayasari. I., 2020). Rancangan undang-undang ini dirancang untuk memangkas sebagian norma yang dikira tidak cocok

dengan perkembangan zaman serta merugikan kepentingan bagi negeri sehingga pemerintah memandang kalau RUU Cipta Kerja ini perlu dicoba karena tingginya angka pengangguran di Indonesia yang telah mencapai 7 juta jiwa serta diharapkan RUU Cipta Kerja dapat membuka lapangan kerja baru (Kurniawan, F., 2020).

2.2.2. *Twitter*

Dalam media sosial orang dapat mengutarakan opininya secara bebas sesuai dengan UU ITE, baik itu buruk maupun baik. Salah satu media sosial yang dipakai oleh banyak orang adalah *Twitter*. Jumlah pengguna *Twitter* sebanyak 21.05% dari total seluruh pengguna internet di Indonesia (Widyanto., et al., 2019). *Tweet* dapat berisi teks, gambar, video, atau tautan tentang berbagai topik, seperti informasi, opini, hiburan atau promosi (Murthy. D., 2018). *Twitter* juga memiliki tantangan-tantangan seperti batasan karakter, bahasa gaul dan ironi yang harus diatasi dalam pengambilan data tersebut (Pratama. A., 2020).

2.2.3. Analisis sentimen

Analisis sentimen adalah sesuatu yang menganalisa pendapat, sentimen, evaluasi dan emosi orang tentang suatu topik yang yang diekspresikan melalui teks (Herwijayanti. B., et al., 2018). Tujuan analisis sentimen adalah menemukan pendapat atau ulasan berdasarkan pendapat orang yang diklasifikasikan pada berbagai polaritas seperti positif, negatif atau netral. Analisis sentimen juga dapat menyatakan perasaan seperti emosional sedih, gembira atau marah (Rustiana. D., et al., 2017).

2.2.4. Pra-pemrosesan

Pra-pemrosesan (*pre-processing*) merupakan tahapan awal yang dilakukan

dalam penerapan *text mining*. *Pre-processing* pada *Text Mining* bertujuan agar memperoleh informasi menarik berdasarkan data yang tidak terstruktur serta menghapus kata tidak penting pada dokumen (Pakpahan., et al., 2019). Penelitian ini akan melakukan beberapa tahapan dalam pra-pemrosesan diantaranya sebagai berikut :

a. Tokenisasi

Tokenisasi adalah proses membagi teks menjadi unit-unit yang lebih kecil yang disebut token (Aini. R., et al., 2017). Tokenisasi adalah langkah awal dalam banyak aplikasi pengolahan bahasa alami seperti analisis sentimen, pemahaman teks, pencarian informasi dan penerjemahan mesin (Hadi. M. I., et al., 2018).

b. Normalisasi

Normalisasi adalah proses penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu misalnya antara 0 dan 1 (Aini. R., et al., 2017).

c. Stemming

Stemming adalah proses yang mengubah kata-kata dalam teks menjadi bentuk dasarnya atau kata akar (Pramudita. H. R., 2016). *Stemming* bertujuan untuk mengurangi variasi kata yang tidak perlu dan meningkatkan efisiensi pemrosesan teks (Nugroho, B., et al., 2020).

d. Stopword removal

Stopword removal adalah proses yang menghapus kata-kata yang tidak relevan atau tidak berpengaruh dalam suatu teks berdasarkan daftar

Stopword (Rinandyaswara. R., et al., 2021). *Stopword* adalah kata-kata yang sering muncul dalam teks tetapi tidak memberikan informasi penting atau makna, seperti kata sambung, kata depan atau kata ganti (Anwar. M. S., et al., 2019). *Stopword removal* dapat meningkatkan efisiensi dan akurasi dari sistem pencarian informasi, analisis sentimen, klasifikasi teks dan aplikasi pengolahan bahasa alami lainnya (Aini. R. N., et al., 2016).

e. *Feature Extraction*

Feature Extraction adalah proses yang mengubah data mentah menjadi bentuk yang lebih ringkas dan informatif yang dapat digunakan oleh algoritma pembelajaran mesin (Hidayat. R., et al., 2021). *Feature Extraction* bertujuan untuk mengurangi dimensi data, meningkatkan kualitas data, meningkatkan kinerja dan akurasi dari model pembelajaran mesin (Nugroho, B., et al., 2020).

2.2.5. Evaluasi performansi

Evaluasi Performansi adalah suatu tahapan untuk mengukur performansi dari *Classifier* dalam melakukan klasifikasi yang digunakan dalam penelitian (Simorangkir. H., et al., 2018). Dalam membuat evaluasi performansi matriks perlu adanya *Confusion matrix* yaitu suatu alat ukur berupa matriks yang digunakan buat memperoleh jumlah ketepatan klasifikasi terhadap kelas dengan algoritma yang dipakai (Sasongko. T. B., 2016).

Tabel 2. 2 Struktur *Confusion matrix*

	Actual Positive	***Actual Negative***
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Jika model melakukan prediksi terhadap komentar dan menebak label positif dengan benar, maka pada Tabel 2.2 dapat dihitung sebagai *True Positive* (TP). Begitu juga dengan label negatif jika ditebak dengan benar maka dapat dihitung sebagai *True Negative* (TN). Sebaliknya, jika model melakukan prediksi terhadap komentar namun menebak label positif dengan salah maka dapat dihitung sebagai *False Positive* (FP). *False Negative* (FN) juga merupakan hasil prediksi yang salah terhadap label negatif. Dengan adanya *Confusion Matrix* performa sebuah model akan dihitung dari seberapa banyak model tersebut melakukan prediksi dengan benar (Wulandari. D. A., et al., 2021). Pada penelitian ini, penulis akan melakukan tahapan-tahapan evaluasi performansi sebagai berikut :

a. *Precision*

Precision adalah tingkat ketepatan suatu informasi sistem dalam memprediksi target positif terhadap jumlah data yang diprediksi positif (Agastya. I. M., 2018). Berikut adalah rumus untuk menghitung *precision* (Wulandari. D. A., et al., 2021) :

$$Precision = \frac{TP}{TP+FP} \qquad \text{.....(1)}$$

b. *Recall*

Recall adalah tingkat ketepatan suatu informasi sistem dalam memprediksi target positif terhadap jumlah data yang riil positif (Agastya. I. M., 2018). Berikut adalah rumus untuk menghitung *recall* (Wulandari. D. A., et al., 2021) :

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(2)}$$

c. *F1-score*

F1-score adalah salah satu perolehan dari konstanta dua dikali nilai *recall* dan *precision*, dan dibagi jumlah keduanya (Agastya. I. M., 2018). Rumus *F1-score* yaitu sebagai berikut (Wulandari. D. A., et al., 2021) :

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad \text{.....(3)}$$

e. *Accuracy*

Accuracy adalah keberhasilan suatu algoritma dengan seberapa akurat sistem dapat mengklasifikasikan secara benar (Agastya. I. M., 2018). Berikut adalah rumus untuk menghitung *accuracy* (Wulandari. D. A., et al., 2021) :

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad \text{.....(4)}$$

2.2.6. *Synthetic minority oversampling technique*

Synthetic Minority Oversampling Technique (SMOTE) adalah teknik *oversampling* yang secara sintetis membentuk data baru dengan melanjutkan vektor titik minoriti ke arah titik dekatnya yang telah ada dalam dataset. Algoritma ini

dilaksanakan melalui beberapa langkah utama (Astuti, F. D., 2021):

a. Memilih titik minoritas dan titik dekat yang termasuk dalam k-titik dekat terdekat sesuai dengan metrik jarak yang ditentukan.
b. Menghitung posisi baru titik sintesis dengan koordinat yang disusun antara titik minoritas asli dan titik dekat yang dipilih.
c. Menerapkan transformasi normalisasi agar titik sintesis tidak melewati batasan nilai atau ukuran spasi dimensi lain.

Pemilihan titik dekat sangat penting dalam proses SMOTE, karena hasil akhir akan terdampak oleh distribusi titik dekat yang dipilih. Teknik SMOTE telah diimplementasikan dalam berbagai algoritma klasifikasi, termasuk *K-Nearest Neighbor* (K-NN). Proses SMOTE dapat memberikan manfaat seperti (Astuti, F. D., 2021):

a. Mengatasai ketidakseimbangan kelas dalam dataset, yang akan mempengaruhi akurasi klasifikasi.
b. Mempermudah algoritma klasifikasi untuk mempelajari karakteristik kelas minoritas, yang seringkali hanya terdapat dalam jumlah sedikit.
c. Mempersiapkan dataset lebih baik bagi algoritma klasifikasi, yang akan mempengaruhi akurasi klasifikasi global.

2.2.7. *Naïve bayes classifier*

Naïve Bayes Classifier adalah metode klasifikasi yang berdasarkan teorema *Bayes* dan asumsi *Naive* bahwa setiap fitur (kata) dalam dokumen independen satu sama lain (Sari. R. P., et al., 2021). Dalam konteks analisis sentimen publik di *Twitter* terhadap *Omnibus law*, *Naïve Bayes Classifier* dapat digunakan untuk

mengklasifikasikan sentimen dokumen berdasarkan probabilitas kemunculan setiap kata dalam dokumen (Sari. R. P., et al., 2021). Rumus *Naïve Bayes Classifier* untuk mengklasifikasikan sentimen dokumen adalah sebagai berikut:

$$P(c|dokumen) = \frac{\mathrm{P}(c).\prod_{i=1}^{n} P(w_i|c)}{P(dokumen)} \quad \ldots\ldots\ldots\ldots(5)$$

$P(c|dokumen)$ adalah probabilitas bahwa dokumen memiliki kelas sentimen c setelah melihat kontennya, P(c) adalah probabilitas sebelum melihat konten dokumen (probabilitas prior) dari kelas sentimen c. $P(w_i|c)$ adalah probabilitas bahwa kata $(w_i|c)$ muncul dalam dokumen jika dokumen tersebut termasuk dalam kelas sentimen c. Ini biasanya dihitung dengan mengacu pada data pelatihan. $\prod_{i=1}^{n} P(w_i|c)$ adalah hasil perkalian probabilitas setiap kata dalam dokumen dengan kelas sentimen c. $P(dokumen)$ adalah probabilitas dokumen itu sendiri yang merupakan faktor normalisasi yang memastikan semua probabilitas kelas sentimen bersama-sama menjadi 1 (Sari. R. P., et al., 2021).

2.2.8. Data validasi

Dalam pembelajaran mesin data validasi digunakan untuk mengevaluasi kinerja model pada data yang belum pernah dilihat sebelumnya. Data validasi biasanya merupakan subset dari data latih yang digunakan untuk melatih model tetapi tidak digunakan dalam proses pelatihan. Tujuan dari penggunaan data validasi adalah untuk memastikan bahwa model yang dikembangkan dapat menghasilkan prediksi yang akurat pada data baru dan tidak hanya pada data latih (Muller. A., et al., 2016).

Terdapat beberapa teknik yang dapat digunakan untuk membagi data menjadi subset latih, uji dan validasi seperti validasi silang (*cross-validation*) dan

hold-out validation. Validasi silang adalah teknik yang umum digunakan dalam pembelajaran mesin untuk mengevaluasi kinerja model dengan membagi data menjadi subset yang saling tumpang tindih. Dengan cara ini, setiap subset data digunakan secara bergantian sebagai data uji dan data latih, sehingga memungkinkan evaluasi yang lebih akurat terhadap kinerja model. Sedangkan *hold-out validation* adalah teknik yang membagi data menjadi dua subset yaitu data latih dan data uji. Data latih digunakan untuk melatih model sedangkan data uji digunakan untuk mengevaluasi kinerja model (Muller. A., et al., 2016).

2.2.9. *Library* sastrawi

Library Sastrawi digunakan dalam tahapan *stop removal* untuk menghapus kata-kata yang dianggap tidak berhubungan (Putri., et al., 2022). Selain itu, *library* sastrawi dapat memudahkan pengolahan kata dalam bahasa Indonesia dan membantu mempersiapkan data sebelum digunakan. Setiap kata hasil *stopwords* akan dihilangkan semua imbuhan kata dan diubah ke dalam kata bentuk dasar menggunakan bantuan *library* sastrawi (Cahyani, S. N., et al., 2023).