BAB 2
LANDASAN TEORI

2.1 Pajak

Penjelasan pajak telah disebutkan dalam Undang-Undang Nomor 28 tahun 2007 tentang Perubahan Ketiga atas Undang-Undang Nomor 6 tahun 1983 tentang Ketentuan Umum dan Tata Cara Perpajakan. Dalam Undang-Undang tersebut dijelaskan bahwa pajak adalah kontribusi wajib kepada negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat [20]. Mengacu pada pada Pasal 1 ayat (2) Undang-Undang Republik Indonesia Nomor 28 Tahun 2007 Tentang Perubahan Ketiga Atas Undang-Undang Nomor 6 Tahun 1983 Tentang ketentuan Umum dan Tata Cara Perpajakan, Wajib pajak adalah orang pribadi atau badan, meliputi pembayar pajak, pemotong pajak, dan pemungut pajak, yang mempunyai hak dan kewajiban perpajakan sesuai dengan ketentuan peraturan prundang-undangan perpajakan [21]. Pajak merupakan aspek penunjang kemampuan finansial sebuah negara sebagaimana pajak menjadi salah satu bentuk pemasukan pemerintah dalam menyelenggarakan pembangunan di segala bidang [22]. Selain menjadi sumber pemasukan sebuah negara yang merupakan fungsi *budgeter*, pajak memiliki fungsi regulasi atau mengatur dengan tujuan untuk mewujudkan keadilan dan stabilitas perekonomian agar terhindar dari distorsi melalui pengenaan dan pemungutan pajak yang mengatur keharmonisan antara kebijakan fiskal, kebijakan moneter dan kebijakan sektor riil [23].

2.2 Twitter

Twitter merupakan salah satu fasilitas *microblogging* secara realtime yang menjadi tempat berbagi pengalaman secara bebas tanpa adanya halangan [10]. Selain itu, Twitter juga diciptakan dengan tujuan untuk memberi kesempatan kepada penggunanya untuk menyampaikan ekspresi, pendapat, keinginan, kritik, serta bertukar informasi mengenai sebuah informasi yang sedang hangat dibicarakan, tanpa adanya batasan waktu dan ruang melalui unggahannya yang dikenal dengan istilah *tweet* [18]. Layanan *microblogging* di jejaring sosial Twitter memungkinkan para pengguna untuk mengunggah pesan pesan singkat

yang dibatasi maksimal 140 karakter, baik dalam bentuk teks, foto, video, maupun audio [24]. Sosial media Twitter juga memiliki fitur trending topik yang bertujuan untuk mengetahui berita yang sedang ramai diperbincangkan di dalam media sosial tersebut [25].

2.3 Analisa Sentimen

Pada dasarnya analisa sentimen adalah klasifikasi. Namun, dalam faktanya analisa sentimen tidak semudah proses klasifikasi biasa karena terikat dengan penggunaan bahasa [26]. Analisa sentimen merupakan salah satu cabang ilmu dari penelitian *text mining* atau *data mining* yang dilakukan untuk mengkekstrak atribut dari sebuah opini, sentimen, dan emosi yang diekspresikan dengan cara tekstual pada sebuah halaman. Kemudian analisa tersebut juga dilakukan dengan tujuan untuk melihat kecenderungan opini seseorang terhadap sebuah masalah atau objek dengan membagi opini tersebut kedalam tiga kategori, yaitu positif, netral, atau negatif [12].

2.4 Text Mining

Bahan dasar yang dimanfaatkan dalam *text mining* merupakan berkas dokumen yang tidak beraturan. Hal ini dikarenakan *text mining* adalah teknik pengolahan informasi berupa teks dengan mengektraksi sumber data dari dokumen dengan tujuan mendapatkan kata yang dapat menyampaikan isi teks ataupun artikel, guna menganalisis hubungan antar dokumen [16]. Secara singkatnya *text mining* adalah proses pencarian informasi dari teks yang ada. Penambangan teks tersebut juga mempunyai tugas untuk pengkategorisasian teks (*Text categorization*) dan pengelompokan teks (*Text clustering*) sebagai tugas khusus [27].

2.5 Natural Language Processing

Natural Language Processing atau biasa yang dikenal dengan singkatan NLP merupakan salah satu ilmu bidang komputer yang merupakan cabang dari keilmuan dari *Artificial Intelligence* (AI) dan bahasa (linguistik) yang berkaitan dengan interaksi antara bahasa manusia dan kompuer. *Natural Language Processing* ini memiliki tujuan utama untuk membuat mesin yang dapat mengerti serta memahami makna bahasa manusia dan kemudian memberikan respon yang sesuai [28]. Dengan kata lain *Natural Language Processing* memiliki fokus

terhadap pengolahan bahasa alami. Bahasa Alami yang dimaksud merupakan bahasa secara umum yang digunakan oleh manusia dalam berkomunikasi. Bahasa yang nantinya akan diterima oleh komputer wajib diproses dan dipahami terlebih dahulu agar komputer dapat memahami maksud dari pengguna dengan baik [29].

2.6 TF-IDF

TF-IDF atau *Term Frequency - Inverse Document Frequency* yang dikenal atas keefisienan, kesederhanaan, serta keakuratannya merupakan suatu metode algoritma yang biasanya dimanfaatkan untuk menghitung bobot setiap kata [30]. Selain itu, TF-IDF juga memiliki fungsi untuk mengetahui seberapa krusial sebuah kata dalam kalimat yang didasari oleh frekuensi kemunculannya. TF atau *term frequency* dimaksudkan untuk menunjukan jumlah kemunculan suatu kata dalam sebuah tweet. Sedangkan IDF atau *inverse document requency* bertujuan untuk menghitung kekerapan munculnya sebuah kata dalam suatu himpunan [31]. Perhitungan TF-IDF diawali dengan menghitung bobot IDF yang dapat dihitung menggunakan rumus berikut [32]:

$$IDF_t = log(\frac{d}{df_t}) \tag{2.1}$$

Keterangan:
IDF_t : Bobot IDF ke t dokumen d.
df_t : Jumlah dokumen yang mengandung *term* t.
d : jumlah dokumen keseluruhan.
Setelah mengetahui jumlah bobot IDF, perhitungan dilanjutkan dengan menghitung bobot *Term Frequency - Inverse Document Frequency* (TF-IDF) dengan menggunakan rumus sebagai berikut [32]:

$$W_{dt} = tf_t \times IDF_t \tag{2.2}$$

Keterangan:
W_{dt} : Bobot *term* ke t pada dokumen d.
tf_t : Jumlah kemunculan *term* t pada dokumen t.
IDF_t : Jumlah bobot IDF

2.7 Naive Bayes

Naive Bayes adalah metode yang dapat dimanfaatkan dalam *data mining* yang bersifat *supervised learning* berlandaskan Teorema Bayes yang memiliki kemampuan klasifikasi yang mirip dengan *decision tree* dan *neural network* [30]. Teorema bayes memodifikasi atau merubah konsep ”probabilitas terjadinya peristiwa Y dengan syarat terjadinya peristiwa X” menjadi ”probabilitas terjadinya peristiwa X dengan syarat terjadinya peristiwa Y”, dimana P(Y|X) mewakili probabilitas *posterior*, P(Y) merupakan probabilitas *prior*, dan P(Y|X) berfungsi sebagai fungsi *likelihood*, yang mana dapat dianggap sebagai faktor penyesuaian, seperti yang dijelaskan dalam persamaan dibawah [33].

$$P(Y|X) = \frac{P(Y) \times P(X|Y)}{P(X)} = \frac{P(Y) \times P(X|Y)}{\sum_{C=1}^{N_y} P(Y = y_c) P(X|Y = y_c)} \tag{2.3}$$

Sangat sulit untuk menemukan probabilitas bersyarat P(Y|X) untuk semua peristiwa dalam teorema bayes. Hal ini dikarenakan secara realita, banyak faktor yang mempengaruhi peristiwa dan peristiwa yang diketahui mempunyai hubungan yang erat satu sama lainnya. Berdasarkan teorema bayes, saling keterkaitan diantara peristiwa-peristiwa yang diketahui tidak lagi dipertimbangkan. Bentuk umum dari model Naive Bayes terbentuk dari kendala independen yang kuat ditambahkan ke dalam set peristiwa X dan setiap peristiwa dalam set tersebut dianggap sebagai independen dari peristiwa lainnya. Adapun persamaan dari bentuk umum model Naive Bayes dapat dilihat pada rumus persamaan dibawah [33].

$$\begin{aligned} P(X|Y = y_c) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n | Y = y_c) \\ P(X|Y = y_c) &= \prod_{i=1}^{N} P(X_i = x_i | Y = y_c) \end{aligned} \tag{2.4}$$

Kemudian deskriminan Naive Bayes dihasilkan dengan memasukan persamaan tersebut ke dalam teorema bayes, seperti yang ditunjukan pada persamaan dibawah [33].

$$\begin{aligned} P(Y = y_c|X) &= \frac{P(Y = y_c) \times P(X|Y = y_c)}{P(X)} \\ P(Y = y_c|X) &= \frac{P(Y = y_c) \times \prod_{i=1}^{N} P(X_i = x_i|Y = y_c)}{P(X)} \end{aligned} \tag{2.5}$$

Untuk suatu urutan fitur, $Feature = \{f_1, f_2, ..., f_L\}$, dimana f_i merupakan nilai dari setiap atribut dalam fitur tersebut. Dalam perhitungan sebenarnya, penyebut diabaikan karena untuk himpunan nilai fitur yang sama nilai penyebut $P(X)$ akan tetap. Untuk kelas dengan nilai pembilang terbesar langsung dipilih sebagai nilai prediksi berdasarkan besarnya pembilang, seperti pada persamaan dibawah [33].

$$\begin{aligned} y &= arg\, max_{y_c} P(X_i = x_i | Y = y_c) \\ y &= \prod_{i=1}^{N} P(X_i = x_i | Y = y_c) P(Y = y_c) \end{aligned} \tag{2.6}$$

Dalam penggunaan Naive Bayes, terdapat tiga varian yang biasa digunakan. Adapun ketiga varian dari Naive Bayes dijelaskan sebagai berikut [17]:

1. Gaussian Naive Bayes

   Perhitungan dari varian ini menggunakan rumus dari densitas gauss dan ditandai dengan rata-rata dan standar deviasi sebagai parameter. Varian Gaussian Naive Bayes dimanfaatkan untuk menghitung probabilitas dari sebuah data kontinu terhadap kelas tertentu [17]. Gaussian menerapkan Gaussian Naive Bayes untuk klasifikasi. Adapun rumus dari algoritma Gaussian Naive Bayes sebagai berikut [34]:

   $$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \tag{2.7}$$

2. Multinomial Naive Bayes

   Varian ini memperkirakan mengenai independensi diantara kemunculan kata-kata pada dokumen tanpa memperhitungkan urutan kata dan konteks informasi yang ada, serta memperhitungkan jumlah kemunculan kata dalam dokumen [17]. Multinomial mengimplementasikan algoritma Naive Bayes untuk data yang terdistribusi secara multinomial dimana data 14 biasanya diwakili sebagai jumlah vektor data. Distribusi dibatasi oleh vektor x pada setiap kelas y yang dimana jumlah fitur pada klasifikasi teks dan ukuran kosa kata adalah probabilitas P(x|y) dari fitur i yang muncul dalam sampel milik kelas y. Adapun perhitungan frekuensi relatif pada multinomial sebagai berikut [34]:

   $$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_{yi} + \alpha n} \tag{2.8}$$

3. Bernoulli Naive Bayes

   Varian ini menggunakan angka biner dalam melakukan klasifikasi. Menggunakan data diskrit dan menerima fitur hanya sebagai nilai biner, seperti ya atau tidak, benar atau salah, berhasil atau gagal, 0 atau 1 dan seterusnya [17]. Bernoulli menggunakan pelatihan Naive Bayes dan algoritma klasifikasi untuk data yang didistibusikan sesuai dengan distribusi bernoulli multivariat. Kelas membutuhkan sampel untuk direpresentasikan sebagai vektor fitur bernilai biner dikarenakan terdapat beberapa fitur yang masing-masing diasumsikan sebagai variabel *binary-valued*. Jika menyerahkan jenis data lainnya, turunan bernoulli dapat membuat binari inputnya. Adapun rumus dasar pada aturan keputusan untuk Bernoulli Naive Bayes sebagai berikut [34]:

$$P(x_I|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_y) \tag{2.9}$$

2.8 *Confusion Matrix*

Confusion matrix merupakan metode yang dipakai guna mengukur kinerja suatu model klasifikasi. Pada saat melakukan pengukuran, teknik *confusion matrix* menggunakan tabel berupa matriks kotak dengan baris dan kolom masing-masing dari kelas yang diklasifikasi. Setiap baris dan kolom pada tabel tersebut dapat dilihat jumlah *true negatif* dan *true positif* hasil klasifikasi [35]. *Confusion matrix* merupakan suatru matriks dengan ukutan n x n yang digunakan guna menggambarkan performa algoritma klasifikasi dengan menunjukkan klasifikasi yang diprediksi dan aktual, di mana n merupakan jumlah kelas yang berbeda. Adapun representasi *confusion matrix* dapat dilihat pada Tabel 2.1 [36].

Tabel 2.1. Representasi *Confusion Matrix*

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	negative	False Positive (FP)	True Negative (TN)

Sumber: [36]

2.8.1 Accuracy

Accuracy memberikan akurasi model secara keseluruhan yang memiliki bagian dari total contoh yang diklasifikasikan dengan benar oleh pengklasifikasi [35]. Nilai *accuracy* didapatkan dengan menghitung pembagian jumlah prediksi yang benar dengan jumlah total sampel. Adapun rumus yang dapat digunakan guna mendapatkan nilai *accuracy* sebagai berikut [36]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

2.8.2 Precision

Precision merupakan tolok ukur seberapa efektif model pada saat memprediksi data benar positif (TP) di antara semua data yang terprediksi positif [35]. Dengan kata lain, *precision* memvisualisasikan tingkat ketepatan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Semakin tinggi nilai dari *precision* menunjukan rendahnya nilai *false positive* (FP). Adapun rumus yang digunakan untuk menghitung nilai *precision* sebagai berikut [36]:

$$precision = \frac{TP}{TP + FP} \tag{2.11}$$

2.8.3 Recall

Recall merupakan indikator yang mengukur sejauh mana model mampu meprediksi data benar positif dari semua data positif sejati [35]. Dengan kata lain, *recall* adalah sebuah perbandingan antara pengamatan yang diprediksi dengan benar terhadap seluruh pengamatan di kelas sebenarnya. Tingginya nilai *recall* yang didapatkan menunjukan rendahnya nilai *false negative*. Adapun rumus yang digunakan untuk menghitung nilai *recall* sebagai berikut [36]:

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

2.8.4 F1-Score

Secara matematis, *f1-score* adalah nilai rata-rata dari *precision* dan *recall* [35]. *F1-score* atau metrik tunggal yang merupakan rata-rata harmonik tertimbang

dari nilai *precisoin* dan *recal* dihasilkan dengan menggabungkan kedua nilai tersebut. Adapun rumus perhitungan yang dapat digunakan guna mendapatkan nilai *f1-score* [36]:

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.13}$$