BAB III
LANDASAN TEORI

3.1. *Machine Learning*

Machine learning merupakan metode dari *artificial intelligence* yang dapat didefinisikan sebagai aplikasi komputer dan algoritma matematika dengan cara pembelajaran yang berasal dari data yang akan menghasilkan prediksi di masa yang akan datang. Proses untuk memperoleh pembelajaran diperlukan dua tahap yaitu *training* (latihan) dan *testing* (pengujian)[10]. *Machine learning* mempunyai empat metode antara lain sebagai berikut.

1. *Supervised Machine Learning Algorithms* adalah penerapan informasi yang telah ada pada data dengan memberikan label. Algoritma ini mampu memberikan hasil dengan cara membandingkan pengalaman belajar di masa lalu.
2. *Unsupervised Machine Learning Algorithms* adalah untuk memproses data yang tidak mempunyai informasi, dapat digunakan secara langsung. Algoritma ini memiliki hasil untuk menemukan struktur tersembunyi pada data yang tidak berlabel.
3. *Semi-supervised Machine Learning Algortithms* adalah algoritma untuk melakukan pembelajaran data yang tidak memiliki label maupun memiliki. Algoritma ini menggunakan teknik *unsupervised* untuk menemukan struktur dalam variable input. Setelah itu, menggunakan teknik *supervised* untuk membuat prediksi terbaik.
4. *Reinforcement Machine Learning Algorithms* memiliki kemampuan berinteraksi dengan proses belajar. Algoritma ini akan memberikan *reward (*poin) jika model semakin baik atau pengurangan poin (*error)* saat model semakin buruk [11].

3.2. *Text Mining*

Text Mining merupakan analisis teks yang sumber datanya didapatkan dari dokumen. *Text mining* melingkupi proses ekstrasi dari informasi yang berasal dari data teks seperti dokumen, artikel, naskah, atau bahkan komentar pada video YouTube. Pada *text mining* data teks dapat diproses menjadi *Classification*, *Clustering* maupun hanya untuk dianalisis *wordcloud* [12].

3.3. *Cluster Analysis*

Cluster Analysis adalah teknik *multivariate* yang prosesnya menggunakan algoritma *clustering*, dengan tujuan utama menyortir data berdasarkan karakteristik sehingga data yang telah dikelompokkan memiliki kemiripan. *Clustering Analysis* memiliki beberapa seperti *Fuzzy C-Means Clustering*, *Self-Organizing*, *K-means*, dll [13].

3.4. *Elbow Method*

Metode *Elbow* merupakan salah satu metode untuk menentukan jumlah *cluster* yang optimal melalui hasil dari perbandingan jumlah *cluster* yang akan menghasilkan bentuk siku pada suatu titik. Titik *cluster* optimal ditentukan dengan cara membandingkan nilai *cluster*[14].

3.5. *K-means*

K-means merupakan salah satu algoritma untuk melakukan *clustering*. *K-means* didasari oleh penentuan jumlah awal kelompok dengan cara mendefinisikan nilai *centroid* awalnya. Algoritma *K-means* menggunakan proses secara berulang untuk mendapatkan basis data *cluster*. *K-means* akan memilih titik awal *centroid* secara acak. Jumlah iterasi akan dipengaruhi oleh *cluster centroid* secara acak. Setelah titik *centroid* dan iterasinya ditemukan maka proses *clustering* akan berjalan dan akan menghasilkan *cluster* untuk setiap objek dalam *dataset* setelah proses iterasi berhenti [15].

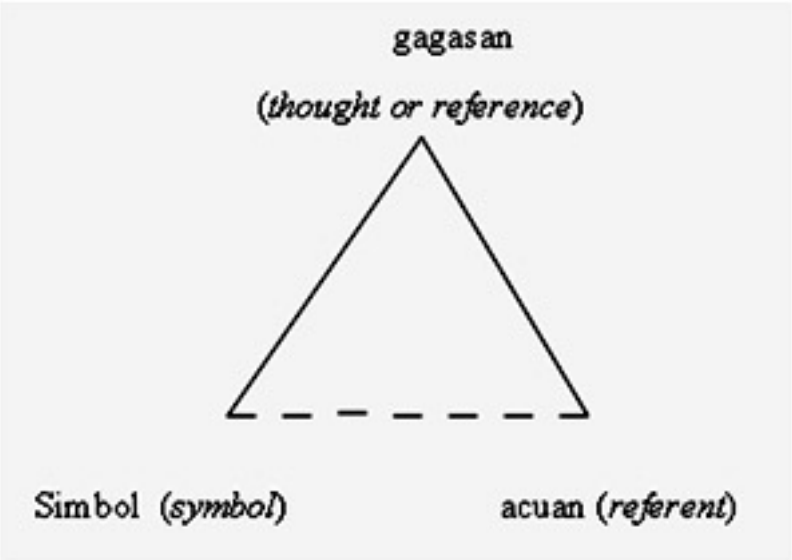
3.6. *Word2vec*

Word2vec salah satu algoritma yang digunakan untuk *word embedding*. Algoritma *Word2vec* mengubah kata menjadi *vector* yang nantinya dapat membawa makna semantik. Algoritma *Word2vec* mempunyai dua algoritma yang berbeda yaitu *Continuous Bag-of-Word* (*CBOW)* dan *Skip-gram*[16].

3.7. Semantik

Semantik adalah cabang ilmu bahasa yang membahas tentang arti kata atau makna. Bahasa merupakan alat untuk berkomunikasi yang digunakan manusia untuk menyampaikan gagasan, perasaan, dan pikiran. Semantik dalam bahasa Indonesia berasal dari bahasa Yunani yaitu *sema* yang berarti "tanda". Kelas kata kerjanya yaitu *semaino* yang berari "menandai". Semantik merupakan salah satu cabang ilmu yang linguistik yang memiliki kaitan erat dengan ilmu-ilmu sosial[17].

3.8. Ogden *and* Richard's *Semantic Triangle*



Gambar 3.1 Segitiga Semantik Ogden dan Richard's

Ogden dan Richard's memperkenalkan teori segitiga makna. Dalam pemaknaan mereka membagi tiga unsur yaitu lambang, konsep, dan panda

referen. Ketiga unsur tersebut saling berhubungan satu sama lain. Pada teori tesebut, hubungan di antara lambang dan konsep terdapat hubungan secara langsung sedangkan lambang bahasa dengan referen atau objek tidak berhubungan langsung[18].

3.9. *Pandas*

Pandas adalah sebuah *library* pada python yang berlisensi *BSD* dan *open-source* yang digunakan untuk analisis data. *Pandas* memiliki struktur dasar yaitu *DataFrame*. *Pandas* memiliki dua tipe struktur data yaitu *Series* dan *DataFrame*. *Series* adalah satu dimensi struktur data *array*. *DataFrame* adalah dua dimensi struktur data atau bisa gabungan dari beberapa *Series*[19].

3.10. *Numpy*

Numpy adalah *library* yang berfokus pada *scientific computing*. *Numpy* sering digunakan untuk membentuk objek N-dimensional *array*. *Numpy* juga digunakan untuk memudahkan operasi Aljabar linear terutama operasi *vector* dan *matrix*[20].

3.11. *Scikit-learn*

Scikit-learn adalah *library* yang berfokus pada pembelajaran mesin. *Scikit-learn* menyediakan beberap algoritma seperti untuk pembelajaran mesin seperti *classification*, *regression*, *clustering*, dan masih banyak lagi[21].

3.12. *NLTK*

NLTK (*Natural Language ToolkitI*) adalah *library* yang berfokus pada pengolahan data *text*. *Platform* ini berbasis *python*. *NLTK* meyediakan beberapa *library* untuk pengolahan *text processing* seperti klasifikasi, tokenisasi, *stemming*, *tagging*, *parsing*, dan sebagainya[22].

3.13. *Gensim*

Gensim adalah *library open-source python* yang diperuntukkan untuk

memproses data teks yang tidak terstruktur. *Gensim* mengubah data teks menjadi *vector* atau yang lebih dikenal dengan *word embedding*. *Gensim* menyediakan beberapa algoritma untuk *word embedding* yaitu *Word2vec*, *FastText*, *Latent Semantic Indexing* (*LSI*), *Latent Dirichlet Allocation* (*LDA*) dan sebagainya. Algoritma tersebut merupakan algoritma bertipe *unsupervised*[23].