BAB II

LANDASAN TEORI

2.1. Text Mining

Text mining adalah proses ekstraksi informasi dan proses *knowledge discoveries* dari kumpulan teks yang tidak terstruktur [11]. *Text mining* sendiri berguna untuk menemukan *pattern* dari sebuah *resource* yang sebelumnya belum terdeteksi [12].

Oleh karena data yang tidak terstruktur, maka tahap *preprocessing* berperan penting dalam melakukan *feature selection* dan *feature extraction* sehingga data tersebut memiliki bentuk yang lebih terstruktur dan dapat digunakan untuk analisis.

Text mining dapat diimplementasikan ke berbagai bidang, seperti *academic applications*, *bioinformatics*, *copyright and customer profile analysis* dan *internet security* [13]. Selain yang telah disebutkan, *text mining* juga dapat diaplikasikan kedalam bidang lain seperti *email spam filtering*, *social media data analysis* dan *business intelligence* [14].

2.2. Sentiment Analysis

Sentiment analysis atau yang juga dikenal dengan *opinion mining* adalah salah satu metode analisa dalam data analisis dimana membutuhkan usaha lebih karena memiliki langkah yang cukup banyak seperti *sentiment extraction and classification*, *subjectivity detection*, *opinion summary*, dan *opinion spam detection* [15].

Sentiment analysis juga dapat dimanfaatkan kedalam berbagai bidang, seperti bisnis, politik, *public action* dan keuangan [16] untuk mengetahui opini dari sebuah topik guna menentukan langkah yang akan diambil agar memiliki dampak yang lebih signifikan.

2.3. Naïve Bayes Classifier

Naïve Bayes Classifier adalah salah satu dari algoritma *supervised*, dimana cara kerja algoritma ini mengasumsikan bahwa setiap *feature* yang terdapat didalam data adalah variabel independen [17]. Algoritma ini mengimplementasikan bayes theorem dengan tingkat asumsi yang tinggi(naïve) bahwa setiap variabel tidak ada yang berkorelasi dan merupakan variabel independen. Walaupun mengasumsikan bahwa setiap variabel adalah variabel independen adalah asumsi yang buruk, namun pada kenyataannya Naïve Bayes memiliki performa yang tidak kalah baik dengan algoritma yang lain [17]. Berikut rumus dasar dari Naïve Bayes dari Rumus 2.1:

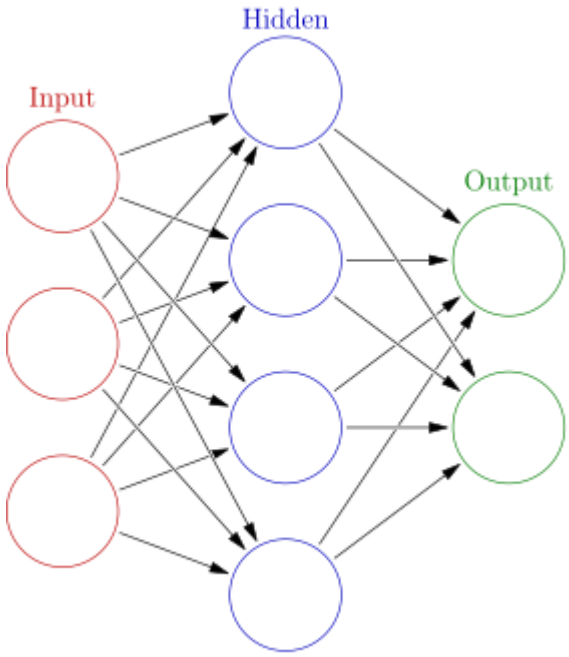
$$\text{argmax} P(x|G = k) \cdot \pi_k$$

Rumus 2.1 Rumus Naïve Bayes [18]

2.4. Neural Network

Neural Network adalah algoritma *deep learning* yang termasuk kedalam kategori *unsupervised learning*. Algoritma ini terdiri dari beberapa lapisan neuron yang dapat dipecah menjadi 3 bagian, yaitu *input layer*, *hidden layer*,

output layer [19]. Berikut ilustrasi dari dari *Neural Network* pada Gambar 2.1:



Gambar 2.1 Ilustrasi Neural Network

Neural Network sendiri memiliki karakteristik dimana ia memiliki keunggulan dimana algoritma ini akan bersinar dengan jumlah dataset yang besar [20]. Karena cara kerja dari algoritma ini sendiri adalah mempelajari data dan menemukan *pattern* dari yang ada pada data tersebut tanpa kita harus memberikan label layaknya algoritma lain pada *supervised learning*.

2.5. Cross Validation

Cross Validation adalah salah satu metode yang digunakan dalam pembangunan model untuk memeriksa akurasi dari model *classifier* yang telah dibangun terutama jika jumlah *data train* dan *test* dalam jumlah yang terhitung kecil. Pada *cross validation*, data dibagi menjadi kedalam beberapa *fold*, yang diandaikan variabel k yang kemudian data dipisah menjadi beberapa *fold* sesuai dengan k yang sudah ditentukan. Misal, k = 5 dengan jumlah data 100 *row*, maka setiap *fold* akan memiliki 20 baris data yang kemudian akan pengulangan sebanyak 5 kali sampai semua *fold* mendapatkan kesempatan menjadi test data dan bagi *fold* yang tidak menjadi *test* data akan menjadi *train* data [21].

Cara standar yang biasa digunakan dalam mengukur tingkat *error* dari *classifier* adalah *stratified 10-fold cross validation*. *Stratified* yang artinya setiap kelas dibagi secara tepat dalam data *train* dan *test*. Sementara angka 10 dianggap sebagai angka yang paling baik dalam mendapatkan perkiraan error terbaik [22]. *Stratified 10-fold cross validation* juga direkomendasikan karena memiliki tingkat bias dan nilai variansi yang terbilang rendah [23].

2.6. Confusion Matrix

Confusion matrix adalah alat yang digunakan untuk mengukur tingkat akurasi dari sebuah model *classifier* yang dibetuk kedalam bentuk tabel n x n. Berikut contoh ilustrasi dari *confusion matrix*:

	Actual Positive	Actual Negative	Actual Neutral
Predicted Positive	True Positive	False Positive	False Positive
Predicted Negative	False Negative	True Negative	False Negative
Predicted Neutral	False Neutral	False Neutral	True Neutral

Gambar 2.2 Ilustrasi Confusion Matrix

Pada Gambar 2.2, akan dijumlahkan antara *True Positives* dan *True Negatives*, dimana *True Positives* adalah nilai positif yang benar diprediksi positif oleh model dan *True Negatives* adalah nilai *negative* yang benar diprediksi *negative* oleh model. Berikut rumus untuk menghitung nilai dari *confusion matrix*:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positive + True\ Negative + True\ Neutral}}{\boldsymbol{n}}$$

Rumus 2.2 Rumus Confusion Matrix [18]

True Positive dan *True Negative* yang dimaksud disini adalah prediksi model dengan fakta atau label dari data. Jika seandainya label data adalah positif dan diprediksi positif berarti dapat dikatakan hasil prediksi tersebut merupakan True Positive begitu juga dengan negative dan netral.

2.7. Python

Python adalah bahasa pemrograman inteprestasi, *high-level programming language* dan *general-purpose* yang diciptakan oleh Guido van Rossum pada tahun 1980 sebagai turunan dari bahasa pemrograman ABC [25]. *Python* diciptakan dengan tujuan untuk membuat bahasa pemrograman yang mudah untuk dibaca dengan menggunakan sistem *whitespace* dimana bahasa pemrograman lainnya menggunakan *curly brackets*.

Python juga dapat digunakan untuk melakukan analisis statistik, pembangunan model pembelajaran mesin atau hal yang berhubungan dengan *data science* lainnya berkat kontribusi dari komintas kedalam *library open-source* yang dibuat untuk digunakan oleh bahasa pemrograman *Python*.

2.8. Sastrawi

Sastrawi adalah *library* yang digunakan untuk mengembalikan sebuah kata berimbuhan kembali ke bentuk dasarnya atau yang bisa juga disebut sebagai *stemming*. *Stemming* merupakan bagian yang terpenting dari proses *text mining*, karena *stemming* dapat menentukan baik atau tidaknya hasil dari *text mining* [26].

2.9. spaCy

spaCy adalah *library* dari *Python* yang digunakan untuk melakukan *tokenization*, *library* ini sudah mendukung untuk pemecahan kata dalam berbagai bahasa, salah satunya adalah bahasa Indonesia. Tujuan dari

tokenization adalah memecah kalimat menjadi perkata sehingga akan lebih mudah untuk dipahami.

2.10. Tweepy

Tweepy adalah adalah *open-source library* yang diciptakan untuk mengakses *Application Programming Interface* (API) dari Twitter. *Library* ini memungkinkan kita untuk mengakses *tweet* pada Twitter dan melakukan *streaming* data dengan menggunakan *query* kedalam *Tweepy*.

Untuk menggunakan *Tweepy*, diperlukan *access token* yang bisa didapatkan dengan mendaftarkan diri sebagai *developer* di halaman Twitter Developer. Setelah mendaftarkan diri, kita akan mendapatkan *consumer key* (*API key*), *consumer secret* (*Secret API key*), *access token (authorization key to use the api*) dan *access token secret (secret authorization key to use the api)*.

2.11. Penelitian Terdahulu

2.11.1. Daftar Penelitian Terdahulu

Tabel 2.1 Tabel Penelitian Terdahulu

1.	Penulis	Brian Keith Norambuena, Exequiel Fuentes Lettura and Claudio Meneses Villegas
	Nama Jurnal	*Intelligent Data Analysis 23 (2019) 191–214 DOI 10.3233/IDA-173807 IOS Press*
	Judul	*Sentiment analysis and opinion mining applied to scientific paper reviews*
	Permasalahan	Ekstraksi sentimen dan opini dari koleksi artikel review yang dibuat oleh International Conference on Computing di Chile Utara
	Metode	- *Machine Learning – Supervised Learning*

		- *Unsupervised methods: Part-Of-Speech tagging* - *Hybrid approach SVM*
	Hasil & Simpulan	Model hybrid SVM yang dibuat memiliki hasil yang lebih baik dibandingkan dengan model lainnya seperti SVM dan Naïve Bayes. Namun untuk akurasi sendiri akan ikut semakin menurun jika jumlah kelas (label) yang semakin banyak
	Adopsi	Model yang digunakan serupa dalam penelitian ini
2.	Penulis	Oueslati, Oumaima, Cambria, Erik, HajHmida, Moez Ben, Ounelli, Habib
	Nama Jurnal	*Future Generation Computer Systems 112 (2020) 408–430 Contents*
	Judul	*A review ofsentiment analysis research in Arabic language*
	Permasalahan	Menganalisa *sentiment analysis* pada umumnya dilakukan dalam bahasa Inggris, namun dalam paper ini akan menganalisa *sentiment analysis* dalam bahasa arab
	Metode	- *Lexicon Construction* - *Monolingual Sentiment Classificaton* - *Corpus Based Approach* - *Feature Extraction* - *Machine Learning-based sentiment classification* - *Hybrid*
	Hasil & Simpulan	*Sentiment analysis* dalam bahasa arab mulai menarik perhatian komunitas NLP dan kebanyakan pendekatan yang digunakan adalah monolingual dan bilingual. Analisa sentimen dalam bahasa arab kebanyakan masih menggunakan Naïve Bayes dan SVM karena *deep learning* masih belum begitu di eksplor sedalam *sentiment analysis* dalam bahasa inggris. Selain itu, untuk *sentiment analysis* yang bersumber dari media sosial masih belum disarankan karena memiliki perbedaan dialek yang cukup luas.
	Adopsi	Pendekatan dalam menganalisis *sentiment* masyarakat terutama yang bukan berasal dari bahasa inggris
3.	Penulis	Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi
	Nama Jurnal	*International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016*

	Judul	Sentiment Analysis of Twitter Data for Predicting Stock Market Movements
	Permasalahan	Memprediksi harga saham merupakan *well-known problem of interest* sehingga peneliti ingin melakukan analisa menggunakan data tweet
	Metode	- *Data Collection* - *Data pre-processing* - *Feature extraction* - *Model training* - *Correlation analysis*
	Hasil & Simpulan	Terdapat korelasi yang kuat antara tweet dengan harga saham. Namun terdapat *future work* yang karena tweet dari twitter memiliki kemungkinan bias yang tinggi maka dapat menggunakan data dari stocktwits dan/atau news
	Adopsi	Alur penelitian dari penelitian ini digunakan dan dimodifikasi untuk disesuaikan dengan penelitian ini
4.	Penulis	Sneh Kalra, Jay Shankar Prasad
	Nama Jurnal	*2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019*
	Judul	Efficacy of News Sentiment for Stock Market Prediction
	Permasalahan	Analisa terhadap *stock market* akan tetap sulit karna *naturenya* yang sulit diprediksi
	Metode	- *Stock numeric data collection* - *Stock news dataset* - *Text preprocessing* - *Sentiment analysis using naïve bayes, knn, svm* dan *neural network*
	Hasil & Simpulan	Akurasi prediksi nilai saham paling tinggi dihasilkan oleh KNN dengan nilai 91.2%. *Future work* menyarankan untuk menggunakan data dari sosial media, review atau blog yang mempengaruhi *stock market* jangka panjang
	Adopsi	Sentiment analysis pada media sosial twitter
5	Penulis	Nico Nathanael Wilim, Raymond Sunardi Oetama
	Nama Jurnal	*IJNMT (International Journal of New Media Technology), Vol. 8, No. 1 | June 2021*
	Judul	Sentiment Analysis about Indonesian Lawyers Club Television Program Using K-Nearest Neighbor, Naïve Bayes Classifier, and Decision Tree

	Permasalahan	Melakukan *sentiment analysis* untuk menganalisis opini masyarakat di Twitter tentang Indonesia Lawyers Club dan Mata Najwa pada tahun 2018 dan 2019
	Metode	- *Sentiment Analysis* - *Data Collection* - *Preprocessing* - *Classification*
	Hasil & Simpulan	Opini masyarakat pada twitter dapat digunakan untuk memprediksi pemenang Panasonic Gobel Award. 3 model algoritma yang digunakan untuk memvalidasi label manual dengan akurasi tertinggi sebesar 76.94% oleh K-NN. Namun tidak ada algoritma yang memiliki performa terbaik secara keseluruhan karena untuk data pada tahun 2018 akurasi tertinggi dimiliki oleh Naïve Bayes sementara 2019 dimiliki oleh K-NN.
	Adopsi	Sentiment analysis pada media sosial twitter yang menggunakan Bahasa Indonesia

2.11.2. Kesimpulan

Dari penelitian terdahulu, terdapat kesimpulan bahwa media sosial khususnya twitter memiliki peran penting dalam membangun model *machine learning* untuk mengetahui dan memahami secara otomatis pendapat dari masyarakat, dan melakukan *forecasting* menggunakan model *machine learning* tersebut. Namun pada penelitian sebelumnya masih menggunakan bahasa Inggris, maka dari itu dalam penelitian ini ingin melihat apakah model-model *machine learning* tersebut dapat digunakan untuk membangun model dalam bahasa Indonesia.