BAB II

LANDASAN TEORI

2.1. Text Mining

Text merupakan *data* yang tidak memiliki struktur yang terbentuk dari susunan kata. Susunan kata ini memiliki arti dan digabungkan menjadi suatu kalimat. Informasi dalam kalimat tidak dapat dikenali oleh komputer dan dilakukan *text mining* sebagai proses ekstrak informasi melalui *data* teks komputer sehingga dapat mengetahui makna suatu *data* teks [11]. *Text Mining* merupakan suatu proses *mining* atau menambang suatu informasi dari suatu *data* yang disajikan dalam jumlah besar. *Data* yang disajikan ini merupakan teks. Proses *text mining* dilakukan untuk menggali, mengolah, serta mengatur informasi dengan melakukan analisis keterkaitan antara informasi – informasi [12].

Pada *Text Mining* terdapat tahapan yaitu *Pre-Processing* untuk mengubah *data* sesuai dengan *data* yang dibutuhkan. Proses ini akan dilakukan untuk menggali, mengolah, dan mengatur informasi untuk dianalisa hubungan tekstual dari *data* yang terstruktur dan tidak terstruktur. Pada *Pre-Processing* terdapat beberapa tahapan yang dibagi sebagai berikut [8]:

1. *Case Folding*

   Case Folding merupakan tahapan yang bertujuan untuk mengubah setiap kata menjadi bentuk yang sama. Setiap kata akan diubah menjadi huruf kecil tanpa ada huruf kapital [8].

2. *Data Cleansing*

Data Cleansing merupakan tahapan untuk menghilangkan *delimiter* koma, titik, dan tanda baca lainnya serta pada *data cleansing* akan dihilangkan juga *emoticon, mention,* ataupun *link* yang tidak dapat mengganggu. Tujuan dari tahapan ini adalah untuk mengurangi jumlah *noise* [8].

3. Normalisasi Bahasa / *Word Replacer*

Normalisasi Bahasa merupakan tahapan untuk menormalisasi bahasa pada bahasa yang tidak baku. Tujuan tahapan ini untuk mengembalikan penulisan pada kata menjadi ke bentuk kata yang sesuai dengan kata yang baik [8].

4. *Stopword Removal*

Stopword Removal merupakan daftar kata umum yang artinya tidak penting dan tidak digunakan. Proses ini akan mengurangi jumlah kata yang akan disimpan oleh sistem [8].

5. *Stemming*

Stemming merupakan tahapan untuk mencari kata dasar dari hasil *stopword removal*. Aturan dalam *stemming* adalah pendekatan kamus dan pendekatan aturan [8].

6. Tokenisasi

Tokenisasi merupakan tahapan untuk memotong dokumen yang dapat berupa paragraf atau kalimat menjadi beberapa pecahan kecil yang kata [8].

7. *Generate Bigram*

Generate Bigram merupakan proses untuk menggabungkan 2 kata menjadi 1 *token* agar dapat meningkatkan konteks dalam kalimat tersebut dan meningkatkan akurasi dari *model* [13].

2.2. KDD

Knowledge Discovery in Database atau KDD merupakan suatu proses yang terorganisir untuk menemukan pengetahuan yang valid, baru, berguna, dan dapat dipahami dari sekumpulan data. KDD sendiri digunakan untuk memahami suatu fenomena dari data, analisis, dan prediksi [14]. Proses KDD dibagi menjadi beberapa langkah yaitu :

1. *Data Selection*

Data Selection merupakan tahapan untuk melakukan seleksi terhadap *attribute* yang akan digunakan dan sesuai dengan tujuan awal dalam penelitian [14].

2. *Pre-Processing / Cleaning*

Pre-Processing / Cleaning merupakan tahapan yang bertujuan untuk membersihkan *data* dengan melakukan berbagai cara seperti menghilangkan *noise* dalam bentuk simbol, *link, mention* dan menangani *data* yang hilang [14].

3. *Transformation*

Transformation merupakan tahapan untuk mengubah *data* sesuai dengan kebutuhan dan menyamakan dengan tujuan dari penelitian [14].

4. *Data Mining*

Data Mining merupakan tahapan untuk menemukan pola atau informasi yang berguna pada *data* yang digunakan. Terdapat berbagai teknik, metode, atau algoritma yang dapat dipilih. Pemilihan teknik, metode, dan algoritma harus berhubungan dengan tujuan dari penelitian [14].

5. *Interpretation / Evaluation*

Interpretation / Evaluation merupakan tahapan untuk menafsirkan dan mengekstrak pengetahuan yang didapatkan melalui langkah sebelumnya [14].

2.3. Sentiment Analysis

Sentiment Analysis merupakan suatu sistem komputasi yang mempelajari pendapat, sikap, dan emosi terhadap suatu entitas yang dapat mewakili individu, peristiwa, atau topik dari sekumpulan teks [15]. Tujuan dari *sentiment analysis* untuk menentukan isi dari *dataset* yang berbentuk yang berbentuk teks dan bersifat positif, negatif, atau netral karena pendapat menjadi sumber penting dalam pengambilan keputusan seseorang dalam suatu produk [16].

2.4. Instagram

Instagram merupakan sebuah media sosial yang pada dasarnya memiliki fungsi untuk membagi foto dan video kepada sesama pengguna *Instagram*. *Instagram* sudah digunakan oleh berbagai pihak dimulai dari anak-anak hingga dewasa dan sejak September 2017 sudah tercatat lebih dari 800 juta pengguna. Terdapat berbagai

manfaat dari penggunaan *Instagram* baik sebagai akun pribadi hingga menjadi sarana bisnis perseorangan [8].

2.5. Support Vector Machine

Support Vector Machine merupakan teknik untuk melakukan prediksi yang baik dalam melakukan klasifikasi dan regresi. Thorsten Joachim merintis tujuan dari algoritma *SVM* adalah untuk klasifikasi teks dengan menggunakan bobot indeks *term* sebagai fitur. Metode *SVM* mampu menyelesaikan permasalahan *linear* maupun *non – linear*. Pada penyelesaian masalah secara *non – linear* akan menggunakan konsep *kernel* pada ruang kerja dengan dimensi tinggi, dengan mencari *hyperplane* yang dapat membuat maksimal *margin* antar kelas *data*. *Hyperplane* berguna untuk memisahkan 2 kelompok yaitu *class* +1 dan *class* -1 dimana pada setiap *class* memiliki *pattern* - nya tersendiri [8].

2.6. Naïve Bayes

Naïve Bayes merupakan pendekatan statistik yang fundamental dalam pengenalan pola dan didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi menggunakan probabilitas. *Bayesian classification* juga mampu melakukan prediksi probabilitas keanggotaan suatu *class* pada teorema *bayes* dan memiliki kemampuan klasifikasi [9]. Rumus yang digunakan untuk menggunakan algoritma *Naïve Bayes* yaitu:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Rumus 2. 1. Algoritma *Naïve Bayes*

Keterangan :

X = *Data* dengan *class* yang belum diketahui

H = Hipotesis *data* X merupakan suatu *class* spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x

P(H) = Probabilitas hipotesis H

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

2.7. TF-IDF

TF-IDF atau *Term Frequency – Inversed Document Frequency* merupakan metode dalam melakukan pembobotan kata. Pembobotan kata merupakan suatu proses yang memberikan bobot pada setiap kata yang ada di dalam dokumen [6]. Rumus yang digunakan pada *TF-IDF* yaitu [17]:

$$TF - IDF(w,d) = TF(w,d) * (\log\left(\frac{N}{DF(w)}\right))$$

Rumus 2. 2. Metode TF-IDF

Keterangan :

TF – IDF(w,d) = bobot dari sebuah kata dalam dokumen

w = kata

d = dokumen

TF(w,d) = frekuensi dari kemunculan kata di dokumen

IDF(w) = kebalikan dari DF pada kata

N = jumlah keseluruhan dokumen

DF(w) = banyak dokumen yang terdapat kata

2.8. Confusion Matrix

Confusion Matrix merupakan metode yang sudah sering digunakan untuk melakukan perhitungan akurasi dalam *data mining*. Pada *confusion matrix* terdapat informasi mengenai klasifikasi yang telah diprediksi benar oleh sistem. *Parameter* yang akan dihitung pada *confusion matrix* merupakan *accuracy, recall,* dan *precision*. Semakin tinggi dari nilai *parameter* dalam *confusion matrix* maka akan semakin baik [6]. Pada penggunaan *confusion matrix* terdapat hasil berupa tabel yang merepresentasikan hasil klasifikasi biner pada suatu *dataset* [18].

Tabel 2. 1. *Confusion Matrix*

Class	***Positive***	***Negative***
Positive	*True Positive*	*False Positive*
Negative	*False Negative*	*True Negative*

Pada *confusion matrix* terdapat beberapa rumus dalam menghitung performa pada *confusion matrix* [18] :

1. *Accuracy* merupakan perbandingan antara *data* yang telah

diklasifikasikan benar dengan keseluruhan *data* [19].

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

Rumus 2. 3. *Accuracy*

2. *Precision* merupakan perbandingan antara data positif yang diklasifikasikan benar dengan data yang diklasifikasikan positif [19].

$$Precision = \frac{TP}{TP + FP}$$

Rumus 2. 4. Precision

3. *Recall* menunjukan seberapa berhasil data positif terklasifikasi dengan benar positif [19].

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

Rumus 2. 5. *Recall*

2.9. K-Fold Cross Validation

Cross Validation merupakan teknik yang digunakan untuk memvalidasi keakuratan dari sebuah *model* berdasarkan *dataset* tertentu. *Data* yang digunakan untuk pembentukan *model* disebut *data training* dan *data* yang digunakan untuk validasi *model* disebut *data testing.* Pendekatan dalam metode *K-Fold Cross Validation* membagi *dataset* menjadi beberapa *k* [20].

2.10. Rapid Miner Studio

Rapid Miner merupakan *software* atau perangkat lunak untuk mengolah *data* atau *tools* sebuah *tools* yang digunakan untuk teknik *machine learning*, *data mining,*

text mining, dan *predictive analytics. Rapid Miner* mengekstrak pola dalam *dataset* yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan, dan *database* menggunakan prinsip dan algoritma *data mining* [9].

2.11. E-Commerce

E - Commerce merupakan sistem perdagangan atau jual beli yang dilakukan melalui internet dan didapatkan dari inovasi berbagai pihak melalui perkembangan teknologi. Rutinitas dan aktivitas yang meningkat pada masyarakat juga menjadi salah satu inovasi mengapa dibuat *E – Commerce*, agar dapat mempermudah masyarakat dalam mendapatkan produk atau jasa. Manfaat yang diperoleh dari *E – Commerce* dapat dilihat dari sisi pembeli dan sisi penjual. Pada sisi pembeli dapat dengan mudah mendapatkan barang atau jasa dan bisa mendapatkan produk atau jasa dengan harga yang lebih murah. Pada sisi penjual akan bermanfaat karena mampu menjangkau pembeli lebih luas [21].

2.12. PyCharm

Pycharm merupakan *integrated development environment (IDE)* yang digunakan untuk pemrograman komputer, khususnya untuk bahasa pemrograman *Python*. Aplikasi *Pycharm* dikembangkan oleh sebuah perusahaan yang bernama *JetBrains* dan perusahaan ini berasal dari Ceko [22].

Python merupakan bahasa pemrograman yang interpretatif multiguna. Bahasa pemrograman ini lebih menekankan pada keterbacaan kode agar lebih memudahkan

untuk memahami sintaks. Kemampuan yang dimiliki oleh *Python* dapat melakukan *parallel distributed computing* atau teknik komputasi dengan menggunakan beberapa komputer secara bersamaan [23].

2.13. Tableau

Tableau merupakan suatu *software* untuk *business intelligence* yang umum digunakan untuk melakukan visualisasi *data*, analisis *data*, dan membuat laporan dalam bentuk *dashboard* atau *story*. Dengan menggunakan *Tableau* terdapat berbagai hal yang dapat dilakukan seperti menggabungkan *data*. Penggunaan aplikasi *Tableau* menggunakan sistem *drag* dan *drop* [24].

2.14. Cluster Random Sampling

Cluster Random Sampling merupakan suatu proses pengambilan sampel yang memilih secara kelompok atau *cluster* dan bukan memilih secara individu. Setiap kelompok atau *cluster* memiliki kesempatan untuk menjadi anggota atau bagian dari sampel. Teknik ini dapat digunakan dalam penelitian dalam menentukan sampel yang membagi populasi ke dalam beberapa bagian [25].

2.15. Penelitian Terdahulu

Tabel 2. 2. Penelitian Terdahulu

No	Jurnal	Keterangan
1	Penulis	- Siti Masripah - Lila Dini Utami
	Nama Jurnal	*JURNAL SWABUMI, Vol.8 No.2, September 2020*
	Judul Jurnal	*Algoritma Klasifikasi Naïve Bayes untuk Analisa Sentimen Aplikasi Shopee*
	Metode	- Algoritma *Naïve Bayes*

No	Jurnal	Keterangan
		- Analisa *Review* Aplikasi
	Hasil dan Kesimpulan	- Akurasi dari Algoritma *Naïve Bayes* sebesar 71,50%. - Algoritma ini dapat digunakan untuk melakukan analisa sentimen pada aplikasi *shopee*
2	Penulis	- Joviano Siahaan - Wella - Ririn I. Desanti
	Nama Jurnal	*ULTIMA InfoSys, Vol. XI, No. 2, Desember 2020*
	Judul Jurnal	*Apakah Youtuber Indonesia Kena Bully Netizen?*
	Metode	- Algoritma *Support Vector Machine* - Analisa 10 *Youtuber* Indonesia - Pengambilan *data* dari *Instagram*
	Hasil dan Kesimpulan	- Akurasi dari Algoritma *Support Vector Machine* sebesar 81,2% dengan unsur cyberbullying 49,524%
3	Penulis	- Anang Anggono Lutfi - Adhistya Erna Permanasari - Silmi Fauziati
	Nama Jurnal	*Journal of Information Systems Engineering and Business Intelligence, Vol. 4, No. 1, April 2018*
	Judul Jurnal	*Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine*
	Metode	- Algoritma *Support Vector Machine* - Algoritma *Naïve Bayes* - Analisa *Review Sales* dari *Marketplace*
	Hasil dan Kesimpulan	- Akurasi dari algoritma *Support Vector Machine* mencapai 93,65% tertinggi - Akurasi dari algoritma *Naïve Bayes* mencapai 90,65% tertinggi
4	Penulis	- Nitu Kumari - Dr. Shailendra Narayan Singh
	Nama Jurnal	*Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*
	Judul Jurnal	*Sentiment Analysis on E-commerce Application by using Opinion Mining*
	Metode	Analisa Aplikasi *E-Commerce*
	Hasil dan Kesimpulan	Mampu membuat peringkat berdasarkan *data mining* opini berdasarkan peringkat, sistem

No	Jurnal	Keterangan
		bintang, ulasan produk, fitur yang telah di *mining*.

Berdasarkan Tabel 2.1 terdapat 4 jurnal yang telah dijadikan sebagai penelitian terdahulu dengan berbagai macam cara yang berbeda dalam melakukan penelitian. Pada jurnal ke - 1 yang ditulis oleh Siti Masripah dan Lila Dini Utami akan diadopsi penggunaan algoritma *Naïve Bayes*. Pada jurnal ke – 2 yang ditulis oleh Joviano Siahaan, Wella, dan Ririn I. Desanti akan diadopsi penggunaan algoritma *Support Vector Machine* dan pengambilan *data* dari media sosial *Instagram*. Pada jurnal ke – 3 yang ditulis oleh Anang Anggono Lutfi, Adhistya Erna Permanasari, dan Silmi Fauziati akan diadopsi penggunaan algoritma *Support Vector Machine* dan *Naïve Bayes*. Pada jurnal ke – 4 yang ditulis oleh Nitu Kumari dan Dr. Shailendra Narayan Singh akan diadopsi melakukan analisa sentimen terhadap aplikasi *E-Commerce*.

Perbedaan dari penelitian terdahulu dengan penelitian yang akan dilakukan adalah melakukan analisa menggunakan 2 algoritma yaitu *Support Vector Machine* dan *Naïve Bayes*. Algoritma tersebut akan digunakan untuk melakukan klasifikasi positif dan negatif. Sumber *data* yang akan digunakan dalam penelitian ini adalah komentar dari media sosial *Instagram*. Penelitian ini menggunakan *K-Fold Cross Validation* untuk melakukan pelatihan pada *model*.