



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk mengubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

BAB II

LANDASAN TEORI

2.1. *Data mining*

Data mining adalah suatu proses untuk mendapatkan suatu pola dari data dalam jumlah yang besar, dan data tersebut dapat disimpan dalam *database* atau data warehouse, atau tempat penyimpanan lainnya (Ridwan, 2014). *Data mining* merupakan suatu proses untuk mendapatkan pengetahuan (*knowledge*) yang baru, dan dapat digunakan untuk mendapatkan suatu pendukung keputusan. *Data mining* berkaitan dengan bidang ilmu lainnya, yaitu *database system*, *data warehousing*, sehingga *data mining* dapat dikategorikan sebagai metode pengolahan data.

2.1.1 Tahapan *Data Mining*

Menurut M. Ridwan Effendi *data mining* terdapat 6 tahapan sebagai berikut (Ridwan, 2014).

1. Pembersihan Data (*Data Cleansing*)

Data cleansing atau pembersihan data merupakan sebuah proses membersihkan data mentah untuk menghilangkan data yang tidak konsisten atau data yang tidak berkaitan antara satu data dengan data lainnya, contohnya adalah data pelanggan yang didalamnya dengan 2 nama pelanggan yang sama, sehingga harus membersihkan data pelanggan tersebut sampai data tersebut konsisten. Ketika data tersebut sudah dibersihkan maka data tersebut baru bisa dapat diolah.

2. Penggabungan Data (Integrasi Data)

Integrasi data adalah penggabungan antara data dari *database* lain kedalam *database* yang baru. Setelah data tersebut sudah digabungkan, maka data tersebut harus diperiksa sebelum diproses, sehingga data – data yang sudah digabungkan tidak akan muncul data yang tidak konsisten dan data yang tidak relevan.

3. Seleksi Data (*Selection Data*)

Seleksi data merupakan sebuah proses untuk memilih data yang akan dipakai dalam penelitian. Memilih data yang penting berguna untuk memudahkan pengolahan data dalam sebuah mencari pengetahuan baru (*Knowledge*), sehingga data tersebut dapat dipakai sebagai acuan untuk pengambilan keputusan.

4. Transformasi Data (*Data Transformation*)

Transformasi data adalah merubah data atau menggabungkan data satu dengan data lainnya dalam satu format yang sudah ditentukan atau sudah disesuaikan. Merubah data tersebut berfungsi agar lebih mudah dalam pemrosesan dalam *data mining*.

5. *Process Mining*

Process mining merupakan sebuah proses saat metode diterapkan untuk menemukan pola dan pengetahuan (*knowledge*) yang berada didalam data, sehingga data yang sudah diproses tersebut akan menjadi bahan untuk pembuatan pengambilan keputusan yang penting.

6. Evaluasi Pola (*Pattern Evaluation*)

Evaluasi pola berfungsi untuk mengidentifikasi pola – pola menarik dari data - data yang sudah diproses kedalam *knowledge based* yang ditemukan. Evaluasi pola penting untuk menentukan pola – pola dalam data yang sudah diproses agar pola tersebut berguna untuk *decision making*.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

Presentasi pengetahuan merupakan *visualiasi* yang dilakukan untuk memberitahu hasil apa yang didapat dalam penelitian. *Visualisasi* tersebut mengenai tentang metode yang digunakan untuk memperoleh *knowledge*, dan kesimpulan pemrosesan data dengan metode yang digunakan. Presentasi pengetahuan berfungsi untuk memberikan masukan untuk *decision making* dalam permasalahan yang ada, agar keputusan tersebut dapat menjadi salah satu keputusan yang terbaik saat dipilih.

2.2. Text mining

Text mining merupakan salah satu teknik pengolahan data yang digunakan untuk menggali data – data yang tersembunyi dari data yang bersifat teks (Zahrotun & Mushlihudin, 2017). Selain itu *text mining* juga dipergunakan untuk mencari pengetahuan (*knowledge*) yang digunakan untuk mengambil keputusan (*decision making*). *Text mining* termasuk dalam bidang pengolahan data mengenai pengambilan informasi, analisis teks, penggalian informasi, kategorisasi, visualisasi, dan *machine learning* (Dang & Ahmad, 2014).

2.2.1 Tahapan *Text mining*

Ada lima tahap dasar dalam *text mining* sebagai berikut (Dang & Ahmad, 2014):

1. Pengambilan data yang tidak terstruktur

Hal pertama untuk melakukan *text mining* adalah pengambilan data – data dari *source data* yang sudah ditentukan sebelumnya, contohnya : pengambilan data dari *social media instagram* yang berbentuk *comment*.

2. Mengubah data tidak terstruktur menjadi data yang terstruktur

Mengubah data menjadi terstruktur berfungsi untuk memudahkan data untuk diolah dengan metodologi yang sudah ditentukan sebelumnya, sehingga data – data tersebut dapat digali dan mendapatkan informasi yang penting dari data tersebut.

3. Identifikasi pola (*pattern*) dari data

Setelah mengubah data tersebut menjadi data terstruktur, lalu data tersebut diolah dengan menggunakan metodologi yang sudah ditentukan, sehingga pola (*pattern*) yang tersembunyi didalam data tersebut dapat terlihat.

4. Analisis pola

Setelah menemukan pola yang berada didata tersebut, lalu pola tersebut dianalisis sehingga menjadi informasi dan pengetahuan baru untuk pengambilan keputusan. Hasil analisis tersebut dapat berupa *visualisasi* sehingga mudah dibaca bagi yang membutuhkan informasi tersebut.

5. Penyimpanan informasi didalam *database*

Setelah data tersebut sudah diolah dengan metodologi yang sudah ditentukan sebelumnya, sehingga data tersebut menjadi informasi dan pengetahuan yang baru. Lalu informasi dan pengetahuan yang baru tersebut disimpan kedalam *database* agar informasi tersebut tidak hilang dan jika dibutuhkan lagi, informasi dan pengetahuan tersebut dapat diambil dari *database*.

2.3. Analisis Sentimen

Analisis sentimen merupakan pengolahan data tekstual yang didapat dari opini – opini dari setiap manusia yang digunakan untuk mendapatkan informasi sentimen. Analisis sentimen berfungsi untuk penambangan teks yang digunakan untuk komputasi, mengekstrak, mengevaluasi, dan menjelaskan emosi dan informasi dari seseorang dalam bentuk tekstual, yang umumnya berkaitan dengan penghitungan suara misalnya : survei, ulasan di website, sosial media, dan lain sebagainya (Alsaeedi & Khan, 2019).

2.4. Support Vector Machine

Support vector machine adalah salah satu algoritma dengan metode untuk mengenali suatu pola agar dapat digunakan untuk memprediksi untuk klasifikasi suatu kata (Sudin, Junaedi, & Santosa, 2019). Algoritma *support vector machine* merupakan metode yang menggunakan *supervised learning* yang digunakan untuk klasifikasi (Korovkinas & Danėnas, 2017). Dari pengertian tersebut bahwa *support vector machine* dapat digunakan dalam klasifikasi suatu kata. Algoritma *support*

vector machine merupakan algoritma yang sering digunakan untuk melakukan proses klasifikasi, terhadap suatu teks dan memiliki akurasi yang cukup baik (Sudin et al., 2019).

2.5. Naïve Bayes

Algoritma *naïve bayes* merupakan algoritma yang digunakan untuk klasifikasi probabilistik sederhana berdasarkan *Teorema Bayes*, dan khususnya sangat cocok digunakan dalam data yang memiliki input yang tinggi (Korovkinas & Danenas, 2017). Maka dari itu algoritma tersebut dapat digunakan dalam klasifikasi data yang memiliki data input yang tinggi atau yang besar. *Algoritma naïve bayes merupakan algoritma yang baru – baru ini digunakan dalam klasifikasi sentimen.*

2.6. Instagram

Instagram merupakan aplikasi sosial media yang digunakan untuk membagikan hasil foto dan video. Aplikasi *Instagram* juga bisa untuk menyimpan foto dan video yang diabadikan oleh seseorang, dengan cara memposting setiap foto dan video yang ingin dia *share* pada aplikasi tersebut. Pada tahun 2019 aplikasi *Instagram* di Indonesia memiliki pengguna sekitar 62 juta, dengan 49% *user* wanita dan 51% *user* pria (“Indonesia Digital 2019 : Media Sosial - Websindo,” 2019).

2.7. R Programming

R adalah bahasa pemrograman yang berhubungan dengan komputasi dan pengolahan data untuk statistika, dan untuk membuat grafik yang menggunakan tools yang disediakan serta merupakan bahasa pemrograman yang *open source* (Akbar, Arif Deliyus, Adeliyani, & Olviana, 2017). R merupakan bahasa pemrograman yang sering digunakan dalam membuat data statistika, menganalisis

data, dan penelitian lainnya. Selain itu R juga bisa untuk membuat data prediksi yang sangat berguna baik untuk suatu negara, suatu perusahaan ataupun organisasi, sehingga data tersebut dapat digunakan untuk pengambilan keputusan yang tepat.

2.8. Tableau

Tableau merupakan aplikasi atau software yang digunakan untuk menganalisa/menggambarkan suatu data untuk disajikan dalam bentuk yang menarik, seperti contohnya adalah *dashboard* (Akbar et al., 2017). Aplikasi atau software *tableau* sendiri memiliki banyak fungsi untuk mengolah data, seperti contohnya terdapat fungsi *cleansing* yang memudahkan pengguna, memiliki banyak grafik yang dibutuhkan seperti contohnya *scatter plot*, *diagram*, dan grafik lainnya, *tableau* juga memiliki fungsi untuk membuat *source code* sederhana yang digunakan untuk mengolah data.

2.9. Penelitian Terdahulu

Tabel 2.1. Penelitian Terdahulu

No	Nama Penulis	Nama Jurnal	Judul Jurnal	Hasil Penelitian	Pembahasan
1.	Konstantin as Korovkinas, Gintautas Garšva, Paulius Danėnas	Baltic J. Modern Computing, Vol. 5 (2017), No. 4, pp. 398–409	SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis	Membandingkan antara algoritma <i>support vector machine</i> dan <i>naïve bayes</i> untuk sentimen dari <i>review movie</i> , <i>tweets</i> , dan <i>amazon</i>	Yang didapatkan dari jurnal ini adalah penggunaan 2 algoritma <i>support vector machine</i> dan <i>naïve bayes</i> yang digunakan untuk mencari akurasi pada

					sentimen analisis
2.	Marja Härkänen, Jussi Paananen, Trevor Murrels, Anne Marie Rafferty, dan Bryony Dean Franklin	BMC health services research (Tahun 2019, Vol.19, Pages 791)	Identifying risks areas related to medication administrations - <i>text mining</i> analysis using free-text descriptions of incident reports	Hasil dari penelitian ini untuk meningkatkan keamanan pada obat pada pemeriksaan alergi pasien dan dosis obat	Yang didapatkan dari jurnal ini adalah penggunaan <i>text mining</i> analisis yang digunakan untuk menggali informasi pada data berupa teks atau tulisan.
3.	Katrin Scheibe, Janina Nikolic, dan Wolfgang G.Stock	Proceedings of the 5th European Conference on Social Media, ECSM 2018 (Tahun 2018, Pages 274 – 282)	A sentiment analysis on Miley Cyrus' instagram accounts	Hasil dari penelitian ini adalah mencari positif, netral, dan negative komen pada account Instagram Miley Cyrus	Yang didapatkan dari jurnal ini adalah mengenai tentang analisis sentimen yang digunakan untuk menggali informasi tentang positif, netral, dan negatif komen pada akun Instagram Miley Cyrus

4.	Taufik Dwi Saputra Tanwir, Johan Setiawan	Jurnal ULTIMA InfoSys	Visualisasi Perbandingan Anggaran Pendapatan Dan Belanja Daerah (APBD) Pemerintah Provinsi Kabupaten Dan Kota Di Indonesia Periode 2010-2014	Hasil penelitian ini digunakan untuk membantu mengevaluasi anggaran keuangan pada APBD di setiap daerah di Indonesia, dan membantu masyarakat untuk mengetahui anggaran APBD	Yang didapatkan pada journal ini adalah penggunaan visualisasi data untuk memudahkan mendapatkan pengetahuan baru
5.	Friska Natalia, Yustinus Eko, Ferry Vincenttius Ferdinand, I Made Murwantara dan Chang Seong Ko	ICIC Express Letters, Part B: Applications	Interactive dashboard of flood patterns using clustering algorithms	Hasil penelitian ini digunakan untuk memprediksi banjir yang berada di daerah Tangerang.	Yang didapatkan pada jurnal ini adalah penggunaan visualisasi dashboard yang digunakan untuk memudahkan dalam pengguna non-teknis

Dari kelima penelitian dari (Korovkinas & Danénas, 2017), (Härkänen, Paananen, Murrells, Rafferty, & Franklin, 2019), (Scheibe et al., 2018), (Tanwir & Setiawan, 2016) dan (Natalia, Eko, Ferdinand, Murwantara, & Ko, 2019) merupakan acuan untuk menyelesaikan analisis sentimen terhadap pelayanan MRT pada aplikasi *Instagram* dengan menggunakan algoritma *support vector machine* sebagai algoritma utama dan *naïve bayes* sebagai algoritma pembanding, yang bertujuan untuk mengukur akurasi antara komen positif, netral, dan negatif pada

komen dalam post di akun *Instagram @mrtjkt*, sehingga dengan acuan kelima penelitian ini dapat membantu pengolahan data dengan algoritma *support vector machine* dan *naïve bayes* menjadi lebih mudah dipahami.