BAB II

LANDASAN TEORI

2.1 *Text Mining*

Text mining atau penambangan teks merupakan metode penambangan yang dilakukan oleh perangkat komputer yang bertujuan untuk mendapatkan suatu informasi atau pengetahuan yang tidak diketahui sebelumnya. Informasi atau pengetahuan baru tersebut berasal dari berbagai jenis sumber teks yang berbeda yang diekstrak secara otomatis oleh perangkat komputer [5]. Jika dilihat pada inti dasarnya, konsep teknik *text mining* ini memiliki cara kerja atau metode yang kurang lebih sama dengan konsep metode *data mining,* perbedaannya terletak pada pola datanya. Pada umumnya metode *text mining* ini mengambil objek dari *database* yang belum terstruktur, sedangkan metode *data mining* pada umumnya mengambil objek dari *database* yang sudah terstruktur [6]. Oleh karena metode *text mining* ini biasanya mengambil objek dari *database* yang belum terstruktur, maka tahap-tahap dalam proses melakukan *text mining* secara umum biasanya melakukan *text preprocessing* terlebih dahulu. Tahap ini merupakan tahap awal dari *text mining.* Proses ini bertujuan untuk memperoleh kumpulan data terstruktur yang akan digunakan dalam proses berikutnya. Pada umumnya, tahap *text preprocessing* ini memiliki beberapa tahap didalamnya, yaitu *cleansing, case folding, tokenizing, stopwords removing,* serta *stemming* [7]. Penjelasan dari tahap-tahap tersebut adalah sebagai berikut:

2.1.1 Cleansing

Tahap *cleansing* ini merupakan tahap dimana seluruh kata di dalam dokumen yang tidak diperlukan seperti *hashtag,* alamat *email*, *emoticon, username* dihilangkan.

2.1.2 Case Folding

Tahap *case folding* ini merupakan tahap dimana seluruh karakter huruf di dalam dokumen dijadikan *lowercase.* Tahap ini menghilangkan seluruh karakter selain dari karakter huruf a hingga z, seluruh karakter huruf selain karakter tersebut akan dihilangkan karena dianggap sebagai *delimiter.*

2.1.3 Tokenizing

Tahap *tokenizing* merupakan tahap dimana setiap kalimat yang ada di dalam dokumen dipotong setiap kata-katanya.

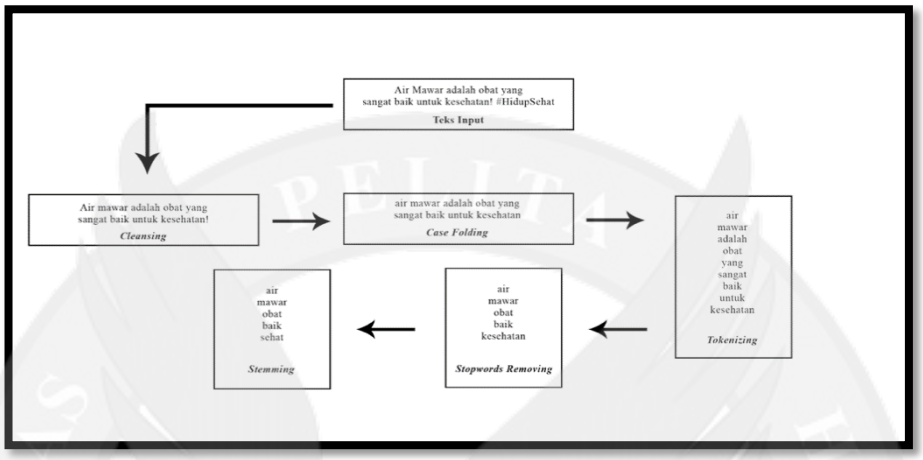
2.1.4 Stopwords Removing

Tahap *stopwords removing* merupakan tahap yang bertujuan untuk membuang kata-kata yang dianggap tidak penting dari hasil *tokenizing*. Sebelum tahap *stopwords removing* dilakukan, biasanya pengguna akan membuat sebuah daftar *stopword* (*stoplist*). Kata-kata akan dihapus jika kata-kata tersebut termasuk dalam daftar *stoplist* yang telah dibuat sebelumnya, sehingga di dalam dokumen hanya akan tersisa kata-kata yang mencirikan suatu dokumen atau yang dianggap penting saja. Terdapat 2 cara umum yang biasanya digunakan untuk melakukan proses *stopwords removing* ini, yang pertama yaitu *stoplist* (membuang kata yang kurang penting), dan ada juga *wordlist* (menyimpan kata penting). *Stoplist* atau *stopword* merupakan kata-kata yang tidak deskriptif sehingga dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopword* dalam bahasa Indonesia adalah "yang", "dan", "di", "dari" dan lain – lain. Untuk saat ini, *stoplist* atau *stopword* versi bahasa Indonesia masih tergolong sedikit dibandingkan dengan bahasa yang sudah umum dikenal contohnya seperti bahasa inggris.

2.1.5 Stemming

Tahap *stemming* merupakan tahap yang bertujuan untuk mengubah seluruh kata-kata menjadi kata-kata akarnya (*root word*) di dalam dokumen. Intinya, Imbuhan-imbuhan yang ada pada kata-kata seperti prefiks, sufiks, maupun konfiks

yang ada pada dokumen dihilangkan pada tahap ini. Namun maka proses *stemming* ini juga harus berdasarkan aturan morfologi yang berlaku untuk bahasa Indonesia karena bahasa Indonesia mempunyai aturan morfologi sendiri.



Gambar 2.1 Contoh alur proses *text preprocessing*

2.2 R Studio

R studio merupakan *software open-source* yang memungkinkan banyak pihak untuk dapat memberikan kontribusi dalam komputasi statistik serta grafis [8]. Penggunaan *software* R studio saat ini pun dapat dikatakan cukup luas, baik di bidang akademis, industri, dan lainnya. R studio menggunakan bahasa pemrograman R. Bahasa R ini pun tidak dapat dikatakan sebagai bahasa pemrograman yang baru, namun bahasa pemrograman R ini merupakan pengembangan bahasa pemrograman "S" di *Bell Laboratories*. Bahasa pemrograman R ini merupakan bahasa pemrograman yang telah dirancang ulang untuk keperluan dalam hal kepraktisan analisis statistika. Walaupun pada awalnya, bahasa pemrograman R dikembangkan untuk analisis statistik, namun saat ini telah berkembang aplikasinya cukup signifikan hingga dapat melakukan manipulasi data spasial serta menampilkannya secara dinamis dalam situs web, hal tersebut dikarenakan meningkatnya kebutuhan-kebutuhan organisasi dalam melakukan analisis statistika. Perkembangan R ini pun menjadi tidak terbendung lagi karena teknologi *big data* sedang berkembang dengan pesatnya saat ini. Saat ini, bahasa R

ini telah menyediakan perintah dasar untuk melakukan analisis *time-series*, pemodelan statistik linear dan nonlinear, klasifikasi, analisis klaster, dan analisis grafis dan masih banyak lagi. Setiap tahunnya, bahasa R pun secara konsisten dengan mengembangkan fungsi dan kemampuannya dengan adanya ribuan paket tambahan yang diunggah ke server CRAN [9].

2.3 Metode *Term Frequency – Inverse Document Frequency* (TF-IDF)

Metode *Term Frequency Inverse – Document Frequency* (TF-IDF) merupakan metode yang digunakan dengan memberikan bobot setiap kata untuk menentukan seberapa jauh keterhubungan kata (*term*) terhadap dokumen. Metode TF-IDF ini cukup sering digunakan ketika melakukan *text mining.* Metode ini terkenal dengan efisiensinya, kemudahan dalam penerapan, serta hasil yang baik dan akurat. Inti dari metode TF-IDF adalah metode ini merupakan penggabungan dari dua konsep yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut [10] [11]. *Term frequency* (TF) adalah frekuensi dari kemunculan sebuah kata dalam dokumen yang bersangkutan, semakin tinggi nilai dari *term frequency* (TF) suatu kata, maka kata tersebut dianggap semakin penting. Sedangkan *inverse document frequency* (IDF) mengukur seberapa pentingnya sebuah kata dihitung dari seluruh dokumen yang dimiliki. Sebagai contoh, dalam sebuah kumpulan dokumen dengan topik *Internet of things*, jika setiap dokumen berbicara mengenai *Internet of things,* maka istilah tersebut tidak akan membantu untuk mengidentifikasi setiap dokumen karena istilah tersebut muncul pada frekuensi yang sama di semua dokumen. Intinya nilai TF akan semakin besar jika jumlah kemunculan suatu kata atau istilah juga besar, sedangkan nilai IDF ini merupakan sistem pembobotan yang akan mengurangi pentingnya istilah tertentu ketika jumlah dokumen yang muncul bertambah. Fungsi untuk menghitung nilai TF dapat dilihat dalam rumus 2.1 berikut ini:

$$TF(d_j, t_k) = f(d_j, t_k) \qquad (2.1)$$

Dimana:

- *f(d,t)* = kemunculan kata *t* dalam dokumen *d*.

Pembobotan IDF ini membuat bobot sebuah kata menjadi semakin besar jika kata tersebut semakin sering muncul dalam satu dokumen, tidak dalam banyak dokumen. Fungsi untuk menghitung nilai IDF dapat dilihat dalam rumus 2.2 berikut ini:

$$IDF\ (t) = log\ \frac{N}{df(t)} \qquad (2.2)$$

Dimana:

- *df(t)* = jumlah dokumen yang memiliki kata *t*.

Sedangkan fungsi untuk menghitung TF-IDF dapat dilihat dalam rumus 2.3 berikut ini:

$$TF\ IDF = TF\ (d_j, t_k)\ .IDF(t_k) \qquad (2.3)$$

2.4 *K-Means Clustering*

K-means clustering merupakan salah satu metode data *clustering* non hirarki yang bertujuan untuk membagi atau mempartisi data yang telah dikumpulkan ke dalam bentuk satu atau lebih *cluster* atau kelompok yang ditentukan. Kecenderungan pengelompokan tersebut didasarkan pada kemiripan karakteristik individu-individu data yang ada [12]. Inti dari metode *k-means clustering* adalah metode ini akan mengelompokkan data yang ada menjadi beberapa kelompok berdasarkan dengan kesamaan karakteristik yang dimiliki oleh satu sama lainnya atau setidaknya memiliki kemiripan antar data. Setiap kelompok

tentu akan saling memiliki karakterstik atau ciri yang memiliki perbedaan satu sama lainnya. Tujuan dari algoritma ini adalah untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah *cluster* yang diinginkan sesuai dengan keinginan *user*. Berikut merupakan tahap-tahap umum algoritma dari metode *k-means clustering* [13]:

a. Masukkan data yang akan diklaster.
b. Tentukan jumlah klaster.
c. Ambil sebarang data sebanyak jumlah klaster secara acak sebagai pusat klaster (sentroid).
d. Hitung jarak antara data dengan pusat klaster, dengan menggunakan rumus *k-means* dapat dilihat dalam rumus 2.4 berikut ini:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} \qquad (2.4)$$

e. Hitung kembali pusat klaster dengan keanggotaan klaster yang baru

Jika pusat klaster tidak berubah maka proses klaster telah selesai, jika belum maka ulangi langkah ke (d) sampai pusat klaster tidak berubah lagi.

2.5 Twitter

Twitter merupakan sebuah media sosial dengan layanan *microblogging* yang mengizinkan penggunanya untuk saling mengirimkan pesan-pesan secara *real-time*. Pesan-pesan yang dikirim oleh pengguna melalui media sosial twitter ini populer dengan sebutan *tweet*. *Tweet* merupakan sebuah pesan pendek dan panjang karakternya hanya dibatasi sampai dengan 140 karakter. Sejak awal, media sosial twitter memang diciptakan sebagai layanan berbasis *mobile* yang dirancang sesuai dengan batasan karakter pada sebuah pesan teks (SMS), dan hingga hari ini, media sosial twitter masih dapat digunakan pada *handphone* apa saja yang memiliki kemampuan untuk mengirim dan menerima pesan teks (SMS) [14].

Media sosial twitter diciptakan sebagai tempat saling berbagi pengalaman (*tweet*) antar sesama penggunanya secara *online* serta bersifat publik sehingga penggunanya dapat mengaksesnya kapanpun dan dimanapun asal terhubung dengan jaringan Internet. Hal ini membuat media sosial twitter kini menjadi salah satu media sosial yang populer di penjuru dunia termasuk di Indonesia karena media sosial twitter sangat mempermudah penggunanya untuk saling terhubung dengan relasi dekatnya. Media sosial twitter juga memudahkan penggunanya untuk mendapatkan informasi-informasi terkini dari seluruh penjuru dunia. Dengan menggunakan tanda # (*hashtag*), pengguna bisa menulis *tweet* berdasarkan topik tertentu. Pengguna juga bisa menggunakan tanda @ untuk menyebut atau membalas pesan dari pengguna lain.

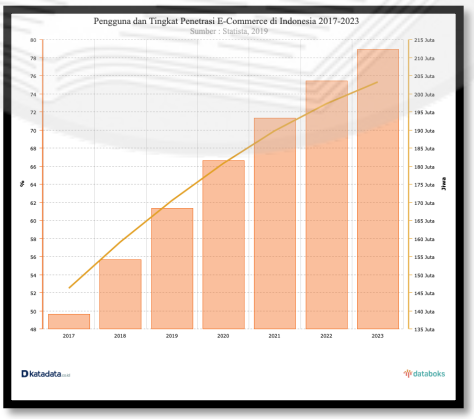
Saat ini, pemakaian media sosial twitter pun terus berkembang dan bertambah, media sosial twitter saat ini tidak hanya digunakan oleh individu saja namun organisasi atau perusahaan pada saat ini kebanyakan sudah memiliki akun media sosial twitter. Hal tersebut tentu dapat memberikan pihak tersebut informasi mengenai pelayanan atau kinerja pihak tersebut sehingga dapat dijadikan sebagai bahan evaluasi, karena media sosial twitter memungkinkan pengguna untuk dapat mengemukakan pendapat secara bebas, kapanpun, dan dimanapun. Media sosial twitter saat ini terdapat cukup banyak *feedback*, baik positif maupun negatif. Media sosial twitter saat ini pun dapat dikatakan menjadi salah satu sarana publikasi yang terbaik karena banyaknya pengguna.

2.6 *Online Marketplace*

Industri *online* di Indonesia kini sedang berkembang sangat pesat, terutama perkembangan industri *online marketplace. Online marketplace* merupakan tempat yang memfasilitasi jual beli secara *online* yang dikelola oleh satu pihak tetapi produk dan informasinya bisa disediakan oleh produsen lain sebagai pihak ketiga [2]. *Online marketplace* ini memiliki konsep yang kurang lebih sama dengan pasar tradisional. *Online marketplace* ini hanya menjadi perantara antara pembeli dan penjual yang bertujuan untuk melakukan transaksi, barang-barang yang dijual di

online marketplace merupakan milik penjual, bukan milik *online marketplace* tersebut. *Online marketplace* hanya menjadi sarana untuk mempermudah transaksi antara penjual dan pembeli dengan memanfaatkan majunya teknologi. Pada umumnya, transaksi yang terjadi pada *marketplace* sendiri memang diatur oleh *marketplace* itu sendiri. Kemudian setelah menerima pembayaran, penjual akan mengirim barang ke pembeli.

Salah satu alasan mengapa *online marketplace* sangat dikenal oleh masyarakat adalah *online marketplace* ini sangat memudahkan masyarakat dalam hal berbelanja, penjual pun tidak perlu membayar biaya sewa sehingga hal tersebut menjadi daya tarik bagi penjual untuk menjual barangnya di *online marketplace*. Jangkauan produk yang ditawarkan di *online marketplace* pun sangat luas, bahkan hampir setiap barang dapat ditemukan di *online marketplace*. Di Indonesia, *marketplace* terbukti telah menjadi *startup* yang paling sukses, dilihat dari jumlah *startup* yang menjadi *unicorn* di Indonesia, seperti Tokopedia dan Bukalapak. Dalam dunia industri, *marketplace* adalah sebuah web *e-commerce*, namun tidak semua situs *e-commerce* adalah *marketplace*. *E-commerce* merupakan sebuah toko virtual yang melakukan kegiatan jual beli produk suatu perusahaan menggunakan media Internet. Sedangkan *marketplace* terdiri dari banyak penjual dan pembeli, serta terdapat berbagai macam produk. Pengguna *e-commerce* di Indonesia menurut situs katadata mengalami pertumbuhan yang cukup besar beberapa tahun terakhir, dan diprediksi akan terus meningkat [15].



Gambar 2.2 Grafik pertumbuhan *e-commerce* di Indonesia
Sumber: situs katadata [15]

2.7 Penelitian Terdahulu

Peneliti mengumpulkan sebanyak 7 penelitian terdahulu serupa dan mengambil keterangan dari dari setiap penelitian yang dikumpulkan.

Tabel 2.1 Penelitian terdahulu

No.	Judul	Peneliti / Tahun	Keterangan
1.	*Opinion Mining On E-Commerce Data Using Sentiment Analysis And K-Medoid Clustering*	Untung Rahardja, Taqwa Hariguna, Wiga Maulana Baihaqi / 2019	Penelitian ini menggunakan konsep *text mining* untuk memilah kata-kata yang tidak terstruktur menjadi informasi sehingga penulis dapat membedakan antara opini positif dan negatif berdasarkan pendapat pelanggan. Penelitian ini mengambil data dari komentar di situs resmi 2 *e-commerce*, yaitu Tokopedia dan Shopee. Penelitian ini menggunakan algoritma TF-IDF dan *K-Medoids.* Penelitian ini memilih menggunakan metode *K-Medoids* karena metode ini dapat membuat lebih dari 2 kategori secara otomatis tanpa bantuan manusia [16].
2.	Penerapan *Text Mining* untuk Melakukan *Clustering Data Tweet* Shopee Indonesia	Dwi Smaradahana Indraloka, Budi Santosa / 2017	Penelitian ini menggunakan konsep *text mining* untuk mencari informasi mengenai jenis konten *tweet* yang banyak dilakukan *retweet* oleh pengikutnya. Dengan adanya informasi tersebut, pelaku bisnis dapat menggunakan data tersebut sebagai salah satu faktor pendukung untuk melakukan *tweet* berikutnya. Penelitian ini mengambil data sebanyak 2000 *tweets* dari akun twitter milik Shopee. Penelitian ini menggunakan metode *clustering* dengan *k-means.* Penelitian ini memilih menggunakan metode *clustering* dengan *k-means* karena metode ini tidak memerlukan jumlah iterasi yang banyak sehingga sangat cocok untuk diterapkan untuk jumlah data yang sangat besar [17].

Tabel 2.2 Penelitian terdahulu (Lanjutan)

No.	Judul	Peneliti / Tahun	Keterangan
3.	*Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes*	Achmad Bayhaqy, Sfenrianto Sfenrianto, Kaman Nainggolan, d Emil R. Kaburuan / 2018	Penelitian ini membahas bagaimana melakukan *sentiment analysis* dari opini pelanggan *e-commerce* di twitter yang akan digunakan untuk menentukan apakah data opini termasuk sentimen yang positif atau negatif. Penelitian ini mengambil data dari akun twitter milik Tokopedia dan Bukalapak. Penelitian ini menggunakan tiga metode berbeda, dan kemudian ketiga hasil tersebut dibandingkan untuk mengetahui *classifier* yang memberikan hasil terbaik dalam aspek *recall ratio* dan *accuracy* dalam *data mining.* Ketiga metode tersebut yaitu *decision tree, k-nearest neighbor,* dan juga *naïve bayes* [18].
4.	*Sentiment Analysis System for Indonesia Online Retail Shop Review Using Hierarchy I Bayes Technique*	Cut Fiarni, Herastia Maharani, Rino Pratama / 2016	Penelitian ini bertujuan untuk merancang sebuah sistem yang dapat membantu mendapatkan informasi mengenai pendapat pelanggan mengenai produk atau servis distro yang ada di Indonesia. Penelitian ini menggunakan metode *naïve bayes classifier.* Metode klasifikasi ini dipilih penulis karena metode ini cukup sederhana dan memiliki performa yang baik. Penelitian ini mengambil data sebanyak 1442 dari o*nline review* pada akun facebook Distro XYZ dengan periode awal bulan januari hingga akhir bulan juli [19].
5.	*Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies*	Siti Ernawati, Eka Rini Yulia, Frieyadie, Samudi / 2018	Penelitian ini untuk meneliti apakah algoritma *naïve bayes s*etelah ditambahkan *genetic algorithm feature selection* dapat meningkat akurasinya untuk s*entiment analysis fashion online companies review. P*enelitian ini mengambil data sebanyak 200 diambil dari situs web yang menyediakan *online review*. Hasil dari penelitian ini yaitu akurasi dari algoritma *naïve bayes* s*e*telah ditambahkan g*enetic algorithm feature selection* dapat meningkat [20].

Tabel 2.2 Penelitian terdahulu (Lanjutan)

No.	Judul	Peneliti / Tahun	Keterangan
6.	*Real-time Sentiment Analysis On E-Commerce Application*	Jahanzeb Jabbar, Iqra Urooj, Wu JunSheng, Naqash Azeem / 2019	Penelitian ini mengidentifikasi pendapat dan fitur secara individual dengan melakukan *sentiment analysis.* Penelitian ini mengambil data dari salah satu situs *e-commerce* terbesar di dunia yaitu amazon, namun tidak disebutkan berapa jumlah data yang dikumpulkan pada penelitian ini. Penelitian ini menggunakan metode *support vector machine* (SVM) untuk melakukan klasifikasi [21].
7.	*Measuring e-Commerce service quality from online customer review using sentiment analysis*	Puspita Kencana Sari, Andry Alamsyah, Sulistyo Wibowo / 2018	Tujuan dari penelitian ini adalah untuk mengetahui bagaimana persepsi konsumen menggunakan *service quality measurement* terhadap salah satu *e-commerce* terbesar di Indonesia, yaitu Tokopedia berdasarkan data dari yang diperoleh dari TrustedCompany, sebuah situs *review.* Penelitian ini menggunakan metode *naïve bayes classifier*. Hasil penelitian ini dapat digunakan sebagai evaluasi mendatang untuk Tokopedia mengenai kualitas layanan mereka [22].

Berdasarkan penelitian-penelitian terdahulu tersebut, dapat disimpulkan bahwa metode *text mining* banyak digunakan oleh peneliti terdahulu untuk tujuan tertentu, peneliti terdahulu juga banyak mengambil data-data teks dari media sosial dan kebanyakan dari penelitian terdahulu tersebut bertujuan untuk melakukan *sentiment analysis. Sentiment analysis* memungkinkan peneliti untuk mengidentifikasi sentimen atau pandangan masyarakat terhadap *e-commerce*. Algoritma ataupun metode yang digunakan pada *text mining* peneliti terdahulu juga bervariasi, menyesuaikan dengan tujuan penelitian. Hasil penelitian tersebut pun diharapkan dapat menjadi bahan evaluasi.