BAB II

LANDASAN TEORI

2.1 Bank Digital

Bersamaan dengan perkembangan teknologi informasi, semua sektor industri terus berlomba-lomba untuk mengimplementasikan teknologi agar dapat meraih keuntungan. Salah satunya adalah industri perbankan yang sebelumnya sudah menawarkan berbagai produk yang berbasis internet untuk melakukan proses bisnisnya, seperti *internet banking* dan *mobile banking* [26].

Mengutip peraturan dari Otoritas Jasa Keuangan (OJK) RI No. 12/POJK.03/2021, bank digital adalah Bank Berbadan Hukum Indonesia (BHI) yang menyediakan dan menjalankan kegiatan usaha terutama melalui saluran elektronik tanpa kantor fisik, selain minimal satu Kantor Pusat (KP) atau menggunakan kantor fisik terbatas [27].

Dalam proses pembuatan bank digital di Indonesia, OJK memperbolehkan dua cara berdasarkan Pasal 25 OJK RI No. 12/POJK.03/2021. Cara pertama adalah pembuatan bank BHI baru yang didaftarkan sebagai bank digital [27]. Sedangkan cara kedua adalah dengan menggunakan bank BHI yang sudah resmi kemudian diubah menjadi bank digital [27].

2.2 Text Mining

Text mining mengambil konsep dasar dari *data mining* dalam mengolah informasi, yaitu mengambil suatu unit data teks baik dalam bentuk kata, kalimat, atau dokumen yang berguna dari sekumpulan unit lainnya [28]. Dokumen berguna yang dimaksud dapat diambil polanya kemudian digunakan untuk memprediksi dokumen lainnya, baik dengan cara klasifikasi atau cara *clustering* yang masih mengolah teks dalam wujud tidak terstruktur [28]. Dikarenakan bentuk data yang akan diolah tidak selalu dalam bentuk terstruktur, langkah-langkah *preprocessing* perlu dilakukan sebelum memulai proses *mining.*

2.3 Analisis Sentimen

Analisis sentimen atau *opinion mining* merupakan tahap analisis lebih dalam lagi dari *text mining* yang berguna untuk menemukan pendapat atau kecenderungan suatu hal yang ada di unit teks tersebut, seperti emosi yang positif, negatif, ataupun netral. [29]. Analisis sentimen merupakan salah satu implementasi metode klasifikasi yang populer untuk dilakukan, karena dapat mengetahui pandangan, pendapat, reaksi, dan emosi seseorang hanya berdasarkan nada bicara dari teks yang tertulis [30].

Tetapi terdapat banyak tantangan yang dihadapi ketika melakukan analisis sentimen, utamanya dikarenakan adanya banyak konteks ambigu dalam kalimat yang hanya dapat dinilai oleh manusia. Oleh karena itu analisis sentimen juga disebut sebagai proses analisis yang membutuhkan teknik komputasi tetapi analisisnya perlu dibimbing oleh *user* yaitu manusia [28].

Proses analisis sentimen biasanya bekerja lebih baik pada unit teks yang bersifat subjektif, yaitu yang mengandung banyak emosi dan perasaan dalam teksnya [29]. Pemanfaatannya biasanya dapat digunakan pada analisis survei, data media sosial, serta berbagai macam jenis ulasan suatu produk, barang, ataupun jasa.

2.4 Pemodelan Topik

Proses untuk mendapatkan topik dari beberapa dokumen disebut sebagai pemodelan topik *(topic modelling)* yang berdasarkan dari konsep *unsupervised* [30]. Pemodelan topik dapat digunakan untuk merangkum dokumen berupa teks dalam jumlah besar menjadi sekelompok topik. Topik yang dimaksud merupakan sekumpulan kata-kata yang sering kali muncul bersama dalam dokumen, dimana setiap kata juga memiliki bobotnya dalam suatu topik [31]. Oleh karena itu, satu kata dapat ditemukan dalam sekumpulan kata (topik) lainnya yang berbeda [31]. Beberapa algoritma yang digunakan untuk *topic modelling* adalah Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), dan Non-Negative Matrix Function (NNMF).

2.5 Naïve Bayes

Naïve Bayes adalah jenis algoritma klasifikasi yang cocok memproses data teks dalam jumlah banyak karena bekerja menggunakan teori probabilitas, misalnya, berdasarkan probabilitas keberadaan suatu kata pada salah satu target output [30]. Algoritma ini juga sering dibandingkan dengan beberapa algoritma lainnya yang lebih rumit, namun algoritma ini memiliki kelebihan yaitu dapat memproses data kategorikal dengan beberapa level dalam waktu yang sebentar dan mampu memproses data yang *null* [31]. Kelemahan dari algoritma ini adalah data yang digunakan harus bervariabel independen, serta sering terjadi *overfitting* yaitu saat model sangat menggambarkan data *training*, sehingga tidak dapat memprediksi data *testing* secara benar [31].

2.6 Latent Dirichlet Allocation (LDA)

Algoritma Latent Dirichlet Allocation adalah algoritma berbasis statistik yang menghitung peluang terkumpulnya suatu kelompok kata-kata menjadi satu topik yang mewakili sekumpulan tersebut [30]. Dalam mekanisme algoritma LDA ini, setiap dokumen dapat terhitung ke beberapa topik yang berbeda. Sebagai salah satu algoritma dengan teknik *unsupervised clustering*, hasil luaran dari algoritma ini harus dievaluasi kembali kebenarannya oleh manusia yang lebih mengerti konteksnya secara mendalam [32].

2.7 Confusion Matrix

Confusion matrix adalah bentuk visualisasi tabel yang menggambarkan hasil performa model klasifikasi. Berikut adalah tampilan tabel *confusion matrix* beserta penjelasan masing-masing *metrics* yang berada di dalamnya [31], [29]:

Tabel 2.1 Tabel Confusion Matrix

Actual Class	Predicted Class		
		Positive	**Negative**
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

1. Akurasi: nilai akurasi sering dianggap menjadi tingkat kesuksesan model secara keseluruhan, walaupun tidak selalu menggambarkan hal tersebut. Hal yang didapatkan dari perhitungan adalah jumlah kelas yang berhasil diprediksi ke kelas yang aslinya (*true positive* dan *true negative*) [29]. Berikut rumus 2.1 adalah rumus dari nilai akurasi:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Rumus 2.1 Rumus Nilai Akurasi

2. Presisi: nilai ini menggambarkan jumlah data positif yang terprediksi benar positif, dengan total data yang diprediksi sebagai positif. Oleh karena itu, nilai presisi bergantung pada banyaknya data di kelas positif [29]. Berikut adalah rumus dari presisi pada rumus 2.2:`

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

Rumus 2.2 Rumus Nilai Presisi

3. *Recall / Sensitivity:* nilai yang mengukur persentase jumlah kelas positif yang benar terprediksi positif [29]. Berikut rumus 2.3 adalah rumus dari *recall*:

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

Rumus 2.3 Rumus Nilai *Recall*

4. *F1 Score:* digunakan untuk untuk mendapatkan nilai presisi dan *recall* yang seimbang [29]. Berikut rumus dari nilai F1 pada rumus 2.4:

$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

Rumus 2.4 Rumus Nilai F1-*score*

2.8 Penelitian Terdahulu

Tabel 2.2 berikut ini merupakan tabel yang merangkum penelitian terdahulu mengenai analisis sentimen, teknik pemodelan topik, serta yang melakukan keduanya:

Tabel 2.2 Tabel Penelitian Terdahulu

Judul Artikel	**Nama Jurnal / Vol / *Issue***	**Penulis / Tahun**	**Metode (Algoritma, Tools)**	***Findings***
Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter [23]	The 6th International Conference on Cyber and IT Service Management (CITSM 2018)	Kristiyanti, D., Umam, A., Wahyudi, M., et al. / 2018	• Naive Bayes • SVM • 10-*fold cross validation* • RapidMiner	• Menggunakan data *tweet* tentang 4 pasang calon gubernur Jawa Barat periode 2018-2023, masing-masing pasangan sebanyak 100 *tweet* berlabel positif dan 100 *tweet* negatif. • Pada setiap *dataset* pasangan calon gubernur, algoritma Naive Bayes selalu memperoleh nilai akurasi yang lebih tinggi dibanding SVM. • Model Naive Bayes yang dibuat memiliki performa paling baik dengan akurasi 94% pada *dataset* 2DM.
Support Vector Machine versus Naive Bayes Classifier: A Juxtaposition of Two Machine Learning Algorithms for Sentiment Analysis [33]	International Research Journal of Engineering and Technology (IRJET) / 7 / 7	Arora A., Patel P., Shaikh S., et al. / 2020	• Naïve Bayes • SVM	• Menggunakan 3 *dataset* yaitu ulasan film dari IMDB, ulasan produk dari Amazon, dan ulasan *service* dari Yelp. Masing-masing *dataset* terdiri dari 500 data bersentimen positif dan 500 data bersentimen negatif. • Model Naive Bayes memperoleh rata-rata nilai akurasi dan F1 *score* tertinggi dibandingkan dengan SVM, yaitu sebesar 0.794 untuk kedua nilai.

Judul Artikel	Nama Jurnal / Vol / *Issue*	Penulis / Tahun	Metode (Algoritma, Tools)	*Findings*
Apakah *Youtuber* Indonesia Kena *Bully* Netizen? [34]	ULTIMA InfoSys / XI / 2	Joviano S., Wella, W., Desanti, R.I./2020	• SVM	• Menganalisis *unlabelled* data dari media sosisal Instagram milik 10 *youtuber* di Indonesia. Masing-masing 1200 komentar per akun. • Model dibangun dari *labelled dataset* lainnya, yang terdiri dari 7440 data positif yang tidak mengandung *cyberbullying*, serta 8390 data negative yang mengandung *cyberbullying.* • Untuk mengklasifikasikan *tweet* yang mengandung *cyberbullying* dan tidak, algoritma SVM menghasilkan akurasi tertinggi 81.2% menggunakan *cross validation.* • Disarankan penelitian selanjutnya menggunakan langkah *preprocessing* seperti *stemming*, *tri-gram*, dan sinonim.
Topic Modeling Twitter Data Using Latent Dirichlet Allocation and Latent Semantic Analysis [19]	AIP Conference Proceedings	Qomariyah, S., Iriawan, N., Fithriasari, K. / 2019	• LDA • LSA	• Menggunakan data *tweet* sebanyak 12534 *tweet* dan 14205 kata setelah diseleksi, untuk menganalisis opini masyarakat Surabaya terhadap pemerintahan. • Berdasarkan nilai *coherence* pada semua skenario data, algoritma LDA selalu memiliki nilai yang lebih tinggi dibanding LSA. • Nilai *coherence* paling tinggi sebesar 0.1376 menggunakan algoritma LDA, pada data dengan skenario 2 dimana setiap tweet memiliki lebih dari 2 kata, dan setiap kata harus muncul lebih dari 4 kali. • Nilai *coherence* paling tinggi dengan LDA menghasilkan 4 topik, yaitu tentang pemilu, parkir, padatnya lalu lintas, serta hujan yang menyebabkan lalu lintas padat.

Judul Artikel	Nama Jurnal / Vol / *Issue*	Penulis / Tahun	Metode (Algoritma, Tools)	*Findings*
				• Hasil topik dapat digunakan oleh pemerintah kota untuk memperbaiki kotanya.
Data Mining of New Snack E-commerce Reviews Based on Text Sentiment Analysis and Latent Dirichlet Allocation Topic Model [16]	Proceedings of the 3rd International Conference on Economy, Management and Entrepreneurship (ICOEME 2020)	Yang, Qian. / 2020	• LDA • *Library* SnowNLP untuk analisis sentimen teks berbahasa mandarin. • *Web data crawling* menggunakan *Python*.	• Menggunakan data ulasan produk "*new snack*" jenis kacang dari 5 toko produk makanan ringan pada *e-commerce* T-Mall. • Sebanyak 435350 data ulasan digunakan dalam periode Januari 2018-2020. • Semua toko memiliki persentase sentimen positif di atas 50%, secara umum pelanggan merespon produk "*new snack*" dengan baik. • Toko Wolong memiliki kepuasan pelanggan yang paling tinggi dengan persentase ulasan positif paling tinggi sebesar 71%. • Toko BESTORE memiliki kepuasan pelanggan yang paling rendah karena mendapat persentase ulasan negatif yang paling tinggi sebesar 28%. • Penentuan topik dilakukan pada semua ulasan sekaligus hanya pada dua toko (Wolong dan BESTORE). Berdasarkan itu, hampir semua topik menghasilkan kata-kata yang positif. • Didapatkan faktor yang memengaruhi nilai beli pelanggan dan kepuasannya yaitu rasa, kualitas, layanan, logistik, dan promo. • Berdasarkan hasil penentuan topik dari ulasan produk, dihasilkan 4 rekomendasi strategi untuk pemilik toko.
Analyzing Brexit's impact using	Pervasive eHealth: Pervasiv	Ilyas, S., Soomro, Z.,	• Vader untuk	• Menggunakan data *tweet* tentang 'brexit', harga indeks saham

Judul Artikel	Nama Jurnal / Vol / *Issue*	Penulis / Tahun	Metode (Algoritma, Tools)	*Findings*
sentiment analysis and topic modeling on Twitter discussion [35]	e Computing Technologies for Healthcare	Anwar, A. et al. / 2020	analisis sentimen • LDA • Korelasi spearman • Twitter API untuk *scraping* • NLTK: *preprocess data*	Inggris, dan data nilai mata uang *pound sterling* dalam 51 hari. • Pergerakan skor sentimen per hari sesuai dengan peristiwa yang terjadi mengenai Brexit. • Pola pergerakan skor sentimen dan harga saham hampir semuanya tidak sama, didukung dengan nilai korelasi spearman sebesar -0.162 yang berarti korelasi linear lemah. • Nilai korelasi harga *pound* dengan skor sentimen 0.589 dan *p-value* sebesar 2.59e-0, menandakan korelasi positif yang cukup kuat dan signifikan. • Validasi jumlah topik ditentukan manual sebanyak 10, dengan *learning decay* 0.7, serta dicari topiknya per hari. • Validasi topik yang didapat secara manual, dengan membandingkannya dengan peristiwa yang terjadi mengenai Brexit per hari. • Penelitian selanjutnya dapat menggunakan nilai log *likelihood* dan *coherence* untuk evaluasi model LDA.

Penulis akan menggunakan beberapa penelitian terdahulu pada tabel 2.2 sebagai referensi untuk penelitian ini mengenai, analisis sentimen dan pemodelan topik pada data media sosial dari tiga bank digital di Indonesia. Penelitian ini akan melakukan klasifikasi sentimen data media sosial dengan algoritma Naïve Bayes sebagaimana digunakan pada referensi di tabel 2.2 yang berjudul *Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter* [23], serta artikel *Support Vector Machine versus Naive Bayes Classifier: A Juxtaposition of Two Machine Learning Algorithms for Sentiment Analysis* [33].

Pada tugas pemodelan topik akan digunakan algoritma Latent Dirichlet Allocation, dengan referensi artikel jurnal berjudul *Topic Modeling Twitter Data Using Latent Dirichlet Allocation and Latent Semantic Analysis* [19], artikel jurnal *Data Mining of New Snack E-commerce Reviews Based on Text Sentiment Analysis and Latent Dirichlet Allocation Topic Model* [16], serta artikel berjudul *Analyzing Brexit's impact using sentiment analysis and topic modeling on Twitter discussion* [35] yang juga melakukan analisis sentimen sebelum pemodelan topik.

Semua referensi artikel jurnal pada tabel 2.2 akan digunakan untuk melakukan analisis sentimen dan pemodelan topik pada media sosial dari tiga bank digital di Indonesia, yaitu Jenius, Jago, dan Line Bank menggunakan algoritma Naïve Bayes dan Latent Dirichlet Allocation.