BAB 2
LANDASAN TEORI

2.1 Robotic Process Automation

Robotic Process Automation (RPA) merupakan suatu pendekatan untuk mengotomatisasi proses dalam sekumpulan teknologi berbeda, yang sesuai dengan tujuan dan proses masing-masing [12]. Dengan RPA, peningkatan kinerja proses bisnis dapat tercapai karena dapat menghilangkan kebutuhan manusia untuk melakukan pekerjaan yang dilakukan secara rutin [13]. Meskipun RPA memiliki berbagai manfaat, tidak semua proses bisnis cocok menggunakan RPA. Pekerjaan dalam proses bisnis yang dapat menerapkan RPA harus memenuhi beberapa kriteria, seperti: pekerjaan yang memerlukan tingkat pengetahuan yang rendah, pekerjaan yang sering dilakukan, atau pekerjaan yang rentan terhadap *human error*. Dengan mempertimbangkan kriteria-kriteria tersebut, dapat dikatakan bahwa RPA akan lebih banyak digunakan untuk keperluan industri daripada keperluan ilmiah [14].

RPA bekerja di atas sistem yang sudah ada sehingga tidak ada *platform* atau sistem baru yang perlu dibuat atau diubah. Robot memiliki akses ke sistem yang sudah ada tersebut seperti pengguna lainnya dan dapat menggunakan *interface* sistem dengan *login credential*-nya sendiri. Kode pemrograman yang digunakan pada sistem tersebut tidak akan berubah karena robot hanya akan menggunakan bagian *interface*-nya saja. Selain itu, robot tidak akan menyimpan data dalam jangka waktu yang lama. Robot hanya akan menyimpan data selama proses berlangsung untuk mengetahui status aktivitas yang sedang dijalankan [15]. Oleh karena itu, pekerjaan yang dilakukan robot biasanya sudah terstruktur dengan baik dan sesuai dengan aturan. Beberapa contoh pekerjaan yang dapat dilakukan robot, antara lain: transfer data antar aplikasi melalui *screen scraping*, pemrosesan *email* secara otomatis, dan pengumpulan data dari berbagai sumber [16].

2.2 Machine Learning

Machine learning merupakan suatu metode yang dapat dimanfaatkan untuk meningkatkan *performance* model yang digunakan untuk melakukan prediksi dan memperbesar akurasi dari prediksi tersebut [17]. *Machine learning* dapat dikelompokkan menjadi tiga macam, yaitu *supervised learning*, *unsupervised*

learning, dan *semi-supervised learning*. *Supervised learning* menggunakan *labeled dataset* untuk melatih algoritma untuk memprediksi hasil atau mengklasifikasikan data [18]. *Unsupervised learning* menggunakan *unlabeled data* sehingga memiliki potensi untuk memahami kumpulan data yang dinamis. Berbeda dengan *supervised learning*, *unsupervised learning* tidak memperhatikan hubungan semantik yang terstruktur sehingga cocok untuk diterapkan pada data heterogen seperti teks, gambar, audio, dan video [19]. *Semi-supervised learning* merupakan gabungan dari keduanya, yaitu menggunakan *labeled data* dan juga *unlabeled data* untuk melakukan tugas tertentu [20].

2.3 Natural Language Processing

Natural Language Processing (NLP) merupakan kemampuan komputer untuk mengenali dan memahami bahasa manusia. NLP termasuk salah satu bidang AI yang bertujuan untuk membuat komputer dapat memahami pernyataan atau kata-kata yang ditulis menggunakan bahasa manusia. NLP dapat mempermudah pekerjaan manusia dan memenuhi keinginan untuk berkomunikasi dengan komputer menggunakan *natural language* [21]. Secara umum, NLP bergantung pada pengetahuan, pemikiran, perspektif, dan interaksi *agent*. *Agent* yang dimaksud dapat berupa manusia atau sistem komputer yang menjadi pembawa informasi, penerjemah, atau yang berpartisipasi sebagai komponen suatu konten informasi [22]. NLP dapat dibagi menjadi tujuh bagian [23], antara lain:

1. Analisis semantik
   Semantik merupakan salah satu cabang ilmu linguistik yang bertujuan untuk mempelajari makna bahasa. Analisis semantik dalam kerangka NLP mengevaluasi dan merepresentasikan bahasa manusia, serta menganalisis teks yang ditulis dalam bahasa manusia dengan interpretasi yang sesuai dengan interpretasi manusia [24].
2. Ekstraksi informasi
   Ekstraksi informasi (*Information Extraction* / IE) adalah proses mengekstraksi informasi terstruktur secara otomatis dari dokumen yang tidak terstruktur dan/atau *semi structured* yang dapat dibaca oleh mesin. Dalam sebagian besar kasus, hal ini melibatkan pemrosesan teks bahasa manusia melalui *text mining*, *pattern matching*, atau teknik serupa lainnya [25].

3. *Text mining*

   Text mining terdiri dari serangkaian teknik yang digunakan untuk mengkarakterisasi dan mengubah teks. *Text mining* menggunakan kata-kata dari suatu teks sebagai unit yang dianalisis [26].

4. *Information retrieval*

   Information retrieval (IR) adalah aktivitas memperoleh sejumlah sumber informasi yang relevan dengan kebutuhan informasi dari suatu kumpulan yang besar. Karena terdapat berbagai sumber yang relevan, hasil yang didapat biasanya akan diberi peringkat berdasarkan gagasan yang relevan [27].

5. *Machine translation*

   Machine translation umumnya menerjemahkan frasa dari satu bahasa ke bahasa lainnya dengan bantuan mesin statistik, seperti Google Translate. Tantangan utama dalam *machine translation* bukanlah menerjemahkan kata secara langsung, namun menghasilkan makna kalimat yang tetap utuh sesuai dengan *grammar* dan *tenses* [21].

6. Sistem penjawab pertanyaan

   Sistem penjawab pertanyaan (*question answering* / QA system) merupakan suatu *platform* yang dapat secara otomatis menjawab pertanyaan yang diajukan dengan bahasa manusia menggunakan *structured database* atau kumpulan dokumen dengan bahasa manusia [28].

7. Sistem dialog

   Sistem dialog sangat populer dan banyak diaplikasikan saat ini, mulai dari menyediakan *support* hingga melakukan *action* tertentu. Pada sistem dialog yang menyediakan *support*, diperlukan pengetahuan mengenai suatu konteks, sedangkan pada sistem dialog yang melakukan *action*, tidak diperlukan banyak pengetahuan mengenai suatu konteks [21].

2.4 Bag of Words

Pada penelitian ini, metode *feature extraction* yang akan digunakan adalah *Bag of Words* (BoW) karena LDA mengasumsikan sebuah dokumen dalam bentuk BoW [29]. *Bag of Words* merupakan salah satu metode yang digunakan untuk *feature extraction* dan klasifikasi dengan cara membuat ”tas” untuk setiap jenis kata. Untuk klasifikasi teks, BoW akan mengkategorikan dan menghitung bobot

setiap kata, yang diambil dari frekuensi kemunculan kata tersebut. Banyaknya kemunculan suatu kata atau istilah dalam suatu dokumen teks disebut sebagai *term frequency*. *Term frequency* ini akan digunakan untuk menentukan kategori atau mengklasifikasikan teks. Metode BoW termasuk ke dalam ekstraksi informasi karena pada dasarnya BoW berfungsi mengekstraksi dan mengklasifikasikan kata-kata dari suatu dokumen [30].

2.5 Topic Modeling

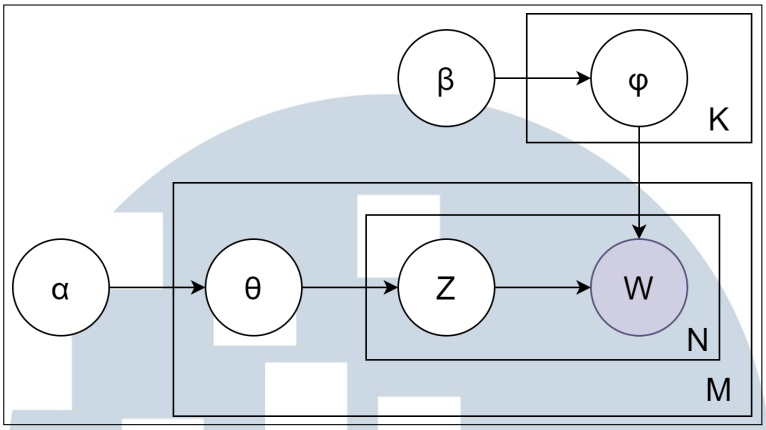
Topic modeling merupakan salah satu metode NLP yang berfungsi untuk menemukan topik tersembunyi dalam suatu data tekstual [31]. *Topic modeling* dapat memberikan gambaran singkat tentang konten utama suatu dokumen dengan cepat tanpa perlu membaca keseluruhan isi dokumen dan memberikan *shortest overview* kepada *pembaca* dengan lebih efisien dibandingkan *text summarization* [32]. *Topic modeling* dianggap sebagai teknik *unsupervised machine learning* karena tidak memerlukan *training* dengan *labeled data* sehingga tidak ada karakteristik khusus dari suatu topik yang perlu dipertimbangkan oleh algoritma sebelum menganalisis data [33].

2.6 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) merupakan suatu metode untuk melakukan *topic modeling* yang bekerja dengan cara menentukan pola pada sebuah dokumen sehingga menghasilkan topik [34]. LDA juga dapat digunakan untuk melakukan *text clustering*, *text summarization*, atau mengolah data yang sangat besar karena LDA menghasilkan sekumpulan topik yang telah diberi bobot untuk setiap dokumen [8].

LDA bekerja dengan mengasumsikan topik sebelum mendapatkan dokumen, kemudian untuk setiap dokumen yang ada akan dilakukan langkah-langkah sebagai berikut [29].

1. Memilih distribusi topik secara acak.
2. Untuk setiap kata dalam dokumen, akan dipilih sebuah topik dari distribusi topik tersebut. Kemudian, dipilih distribusi sebuah kata yang sesuai dari distribusi kosakata.



Gambar 2.1. *LDA plate notation*
Sumber: [29]

Gambar 2.1 merupakan *plate notation* yang menggambarkan model LDA, dengan keterangan sebagai berikut [29].

- β : parameter untuk distribusi kata terhadap topik
- φ : distribusi kata terhadap topik dalam *corpus*
- K : kumpulan topik
- W : kata
- N : kumpulan kata
- M : kumpulan dokumen
- Z : *index assignment*
- θ : dokumen
- α : parameter untuk distribusi topik terhadap dokumen

Nilai α yang tinggi akan menghasilkan distribusi topik per dokumen yang lebih spesifik dan bukan hanya sembarang topik tertentu, sedangkan nilai β yang tinggi berarti setiap topik berisi campuran sebagian besar kata sehingga menghasilkan distribusi kata yang lebih spesifik per topik [35]. Untuk mendapatkan hasil *topic modeling* yang lebih baik dan menghasilkan topik yang lebih koheren, sebaiknya nilai α yang digunakan tidak lebih dari 1 [36].

2.7 Topic Coherence

Topic coherence merupakan suatu metode evaluasi yang menganalisis kata-kata di setiap topik untuk memastikan bahwa kata-kata tersebut dapat menjadi masuk akal jika digabungkan dari sudut pandang manusia [37]. *Topic coherence* memiliki metrik yang konsisten dengan interpretasi manusia. *Topic coherence* dihitung dengan melakukan perbandingan antar kata pada topik secara berpasangan sehingga menghasilkan ukuran standar kualitas suatu topik [38].

Pada penelitian ini, metode *topic coherence* yang digunakan adalah *topic coherence* C_v, yaitu dengan menghitung jumlah kemunculan dua kata atau lebih secara bersamaan dalam suatu dokumen menggunakan *Normalized Pointwise Mutual Information* (NPMI) dari setiap kata yang memiliki jumlah kemunculan tertinggi (w_i) terhadap kata populer lainnya (w_j) dalam dokumen tersebut [39]. Rumus perhitungannya dapat dilihat pada persamaan berikut.

$$NPMI(w_i, w_j) = \sum_{j}^{N-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \tag{2.1}$$

Dengan $P(w_i)$ dan $P(w_j)$ sebagai probabilitas kemunculan kata w_i atau w_j pada suatu dokumen, $P(w_i, w_j)$ sebagai probabilitas kemunculan kata w_i dan w_j secara bersamaan pada suatu dokumen, dan N sebagai jumlah kata dengan tingkat kemunculan tertinggi yang dibandingkan [39].

Nilai yang didapatkan memiliki interval dari -1 hingga 1. Nilai -1 berarti tidak ada kemunculan kata secara bersamaan (tidak berhubungan), 0 berarti tidak ada hubungan antar kata meskipun ada kemunculan secara bersamaan, dan 1 berarti adanya kemunculan kata yang lengkap secara bersamaan dan saling berhubungan [40]. Pada dasarnya, semakin tinggi nilai *topic coherence* (mendekati 1), maka semakin baik topik yang dihasilkan [39]. Nilai *topic coherence* yang optimal tergantung pada data yang digunakan dan interpretasi yang diinginkan. Namun, beberapa penelitian [41], [7], [42] menganggap bahwa skor koherensi di atas 0.3 sudah menghasilkan topik yang relevan dan koheren.