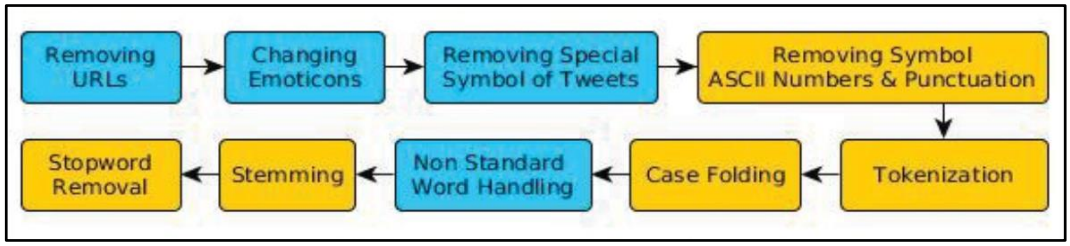
BAB 2
LANDASAN TEORI

2.1 Preprocessing

Preprocessing adalah proses membawa teks ke dalam bentuk yang dapat diprediksi dan dianalisis, menyesuaikannya dengan skema tertentu (DIMITRIADIS, 2020). Pada Gambar 2.1 menggambarkan keseluruhan langkah dan alur kerja terkait *preprocessing* untuk *tweet* yang didapatkan dari Twitter, kotak kuning menunjukkan langkah *preprocessing* yang umum, sedangkan kotak biru menunjukkan langkah *preprocessing* tertentu (Aditya, 2017).



Gambar 2.1 Alur kerja preprocessing (Aditya, 2017)

Menurut DIMITRIADIS (2020), *preprocessing* terdiri dari berbagai langkah yang berbeda dari bahasa ke bahasa dan kasus penggunaan ke kasus penggunaan. Untuk penelitian ini tahap *preprocessing* disesuaikan kembali sehingga tugas yang akan dilakukan yaitu menghapus *line break*, menghapus URL, menghapus simbol dan *emoticon*, menghapus karakter khusus Twitter, menghapus angka, melakukan *case folding*, menghapus tanda baca, menghapus *extra white spaces*, melakukan tokenisasi, menghapus *stop words*, dan melakukan *stemming*.

2.2 Bag of Words

Bag of Words (BoW) adalah metode umum yang digunakan untuk membangun sebuah representasi vektor dari teks dokumen (B, Datko and Maciejewski, 2019). BoW merupakan kumpulan kata pada teks dokumen untuk membentuk suatu urutan sehingga frekuensi kemunculan kata dalam domain tersebut dapat dihitung (Rakhmawati, Basuki and Wicaksono, 2020). Menurut Trisari *et al.* (2020), definisi BoW adalah sebuah model yang mempelajari sebuah kosakata dari seluruh dokumen, kemudian setiap dokumen akan dimodelkan dengan menghitung total kemunculan setiap kata.

2.3 TF-IDF

Metode *Term Frequency - Inverse Document Frequency* (TF-IDF) merupakan pembobotan hubungan dan ukuran statistik yang digunakan untuk mengevalusai seberapa penting sebuah kata dalam sebuah dokumen (Puspita, Muhajir and Aliady, 2020). *Term Frequency* (TF) digunakan untuk mengukur seberapa sering sebuah istilah muncul dalam sebuah dokumen, sedangkan *Inverse Document Frequency* (IDF) digunakan untuk mengukur seberapa penting suatu istilah (Sumalatha, 2018). Sumalatha (2018), mendefinisikan fungsi TF-IDF pada Persamaan 2.1, untuk mendapatkan TF pada Persamaan 2.2, dan untuk mendapatkan IDF pada Persamaan 2.3.

$$TF - IDF(t, d, n, N) = TF(t, d) \times IDF(n, N) \qquad (2.1)$$

$$TF(t, d) = \sum_{word \in d} \begin{matrix} 1 & if\ word = t \\ 0 & else \end{matrix} \qquad (2.2)$$

$$IDF(n, N) = \log\left(\frac{N-n}{n}\right) \qquad (2.3)$$

Keterangan:

a. t = Istilah t atau istilah yang dicari
b. d = Dokumen
c. n = Jumlah dokumen yang mengandung istilah t
d. N = Total dokumen

2.4 Topic Modeling

Topic modeling atau pemodelan topik merupakan salah satu teknik paling kuat dalam penambangan teks untuk penambangan data, penemuan data laten, dan menemukan hubungan antar data dan dokumen teks (Jelodar *et al.*, 2018). *Topic modeling* dianggap sebagai teknik *unsupervised learning* karena tidak melakukan pelatihan dengan data yang sudah diklasifikasikan ('berlabel') (DIMITRIADIS, 2020). Menurut DIMITRIADIS (2020), inti ide dari *topic modeling* adalah bahwa semantic dokumen diatur oleh beberapa variabel tersembunyi atau variabel laten yang tidak diamati. Tujuan dari *topic modeling* adalah untuk mengungkap variabel laten yang membentuk makna dokumen dan korpus dengan membuat topik atau kumpulan kata dalam dokumen yang diamati. *Input* dari *topic modeling* berupa korpus atau satu set dokumen yang kemudian menyediakan *output* sebagai klaster set dari kata-kata, pada setiap klaster tersebut mewakili sebuah topik yakni tema dari diskusi yang muncul dalam dokumen yang diamati. Pada penelitian ini, metode yang digunakan untuk melakukan *topic modeling* adalah dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Menurut Puspita, Muhajir and Aliady

(2020), Konsep *topic modeling* terdiri dari beberapa entitas yaitu kata-kata, dokumen, dan korporasi.

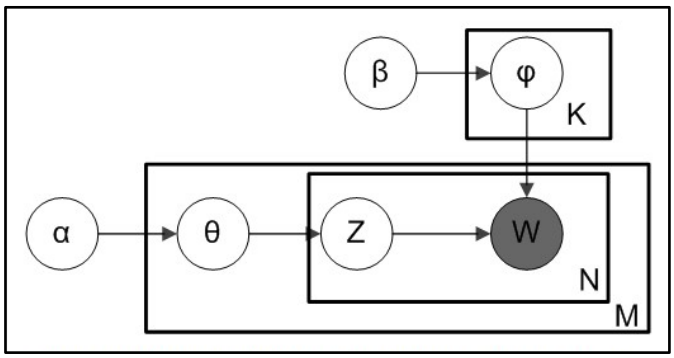
2.5 Latent Dirichlet Allocation

Latent Dirichlet Allocation atau LDA adalah model statistik dalam *Natural Language Processing*. Peran utama LDA yaitu *topic modeling*. LDA memandang setiap dokumen sebagai kumpulan topik dan setiap dokumen memiliki sekumpulan topik tertentu (Mishra, Rajnish and Kumar, 2020). LDA merupakan metode *topic modeling* yang banyak digunakan saat ini, ini adalah teknik faktorisasi matriks dan model statistik yang mengekstraksi kumpulan kata sebagai topik dari kumpulan dokumen teks. Dibutuhkan dokumen *input* sebagai vektor dengan panjang yang tetap (*Bag of Words*). Karena LDA merupakan model statistik dengan kemampuan menghasilkan distribusi probabilitas kata, model ini dapat digunakan untuk memecahkan masalah *machine learning* lainnya selain *topic modeling*, seperti dapat digunakan untuk mengekstrak fitur dari dokumen untuk atribusi kepenulisan (Hasan *et al.*, 2019).

Menurut Setijohatmo *et al.* (2020), cara kerja dari model LDA yakni dengan mengasumsikan topik yang telah dispesifikasikan sebelum dokumen didapatkan. Kemudian pada setiap dokumen yang terdapat di dalam koleksi dilakukan proses sebagai berikut:

1. Distribusi atas topik dipilih secara acak
2. Di dalam dokumen, setiap kata dilakukan proses sebagai berikut:
    a. Sebuah topik dipilih secara acak dari distribusi atas topik

b. Distribusi sebuah kata dipilih secara acak dari distribusi yang sesuai atas kosa kata



Gambar 2.2 Plate notation metode LDA (Setijohatmo *et al.*, 2020)

Pada Gambar 2.2 merupakan *plate notation* untuk visualisasi dari metode LDA. Berikut merupakan keterangan dari *plate notation* metode LDA:

1. β adalah distribusi kata pada setiap topik
2. φ adalah distribusi kata untuk topik k
3. k adalah kumpulan topik
4. α adalah distribusi topik pada setiap dokumen
5. θ adalah distribusi topik untuk dokumen m
6. z adalah topik untuk kata ke-n pada dokumen m
7. w adalah kata spesifik
8. n adalah kumpulan kata
9. m adalah kumpulan dokumen

Dari *plate notation* beserta keterangan di atas menandakan bahwa semakin tinggi parameter β, semakin banyak kata yang terdapat dalam sebuah topik dan sebaliknya, sehingga semakin rendah parameter β maka topik tersebut mengandung kata-kata yang lebih spesifik. Semakin banyak parameter α, maka semakin banyak

topik yang dibahas. Untuk parameter θ, semakin tinggi maka topik yang terdapat dalam dokumen semakin banyak dan sebaliknya, sehingga semakin rendah parameter θ maka topik yang terdapat pada dokumen tersebut semakin spesifik. Pada Gambar 2.1 juga diketahui bahwa parameter w yang merupakan kata spesifik merupakan variabel yang diamati, sedangkan lainnya adalah variabel laten (φ dan θ) dan parameter hiper atau *hyperparameter* (α dan β) (Jelodar *et al.*, 2018). Jelodar *et al.* (2018), mendefinisikan persamaan untuk menghitung probabilitas data pengamatan *D* yang diperoleh dari korpus dengan metode LDA seperti pada Persamaan 2.4.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\,\alpha) \left(\prod_{n=1}^{Nd} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}\,,\beta)\right) d\theta_d \quad (2.4)$$

Keterangan:

1. *D* = Data pengamatan D
2. α = *hyperparameter* untuk distribusi Dirichlet
3. β = *hyperparameter* untuk distribusi Dirichlet
4. *M* = Jumlah dari dokumen.
5. *d* = Dokumen
6. θ = Distribusi Multinomial
7. *N* = Ukuran dari *vocabulary* (kosakata)
8. *n* = kosakata
9. *z* = topik
10. *w* = kata

2.6 Topic Coherence

Dalam penelitian ini *Topic coherence* digunakan untuk mengukur performa dari hasil *topic modeling* yang dilakukan menggunakan metode LDA untuk *tweet* pada sosial media Twitter. *Topic coherence* memberikan ukuran yang tepat untuk menilai seberapa baik model topik (Annisa *et al.*, 2019). *Topic coherence* mencetak topik dengan mengukur tingkat kemiripan semantik dari setiap kata-kata pada topik (Habibi *et al.*, 2021). Menurut Habibi *et al.* (2021), *topic coherence* adalah cara lain untuk mengevaluasi model topik dengan jaminan analisis manusia yang jauh lebih tinggi. Menurut Syed and Spruit (2018), *topic coherence* telah diusulkan oleh peneliti sebagai pendekatan kualitatif untuk secara otomatis dapat mengungkap koherensi dari topik. Pengukuran *topic coherence* bertujuan untuk menemukan ukuran yang berkorelasi tinggi dengan evaluasi topik yang dilakukan manusia.