

## BAB 2

### LANDASAN TEORI

#### 2.1 Cyberbullying

*Cyberbullying* adalah sebuah bentuk kekerasan yang dilakukan oleh kelompok atau individu yang menggunakan media elektronik (UNICEF, 2020). Bentuk dari kekerasannya adalah menjelek, menghina, mengintimidasi atau mempermalukan. Contoh dari perilaku *cyberbullying* seperti mengancam melalui surat elektronik, menghina dikolom komentar media sosial, dan mengunggah foto aib seseorang (KEMENDIKBUD, 2021).

#### 2.2 Text Classification

*Text Classification* atau klasifikasi teks adalah proses pemberian label ke dokumen berdasarkan isi dokumennya. Klasifikasi teks dapat dilakukan dengan dua cara yaitu manual dan otomatis. Klasifikasi teks secara manual memerlukan waktu dan biaya yang besar namun mendapatkan hasil lebih akurat, ahli bahasa harus menafsirkan teks yang ingin diklasifikasi dan mengkategorikannya. Klasifikasi teks secara otomatis menerapkan *machine learning*, *natural processing* sehingga lebih cepat dan hemat biaya. Secara umum klasifikasi teks otomatis terbagi menjadi 3 kelompok yaitu yang pertama klasifikasi teks otomatis berbasis *rule*, *machine learning*, *hybrid* (Rahayu & Andi, 2020).

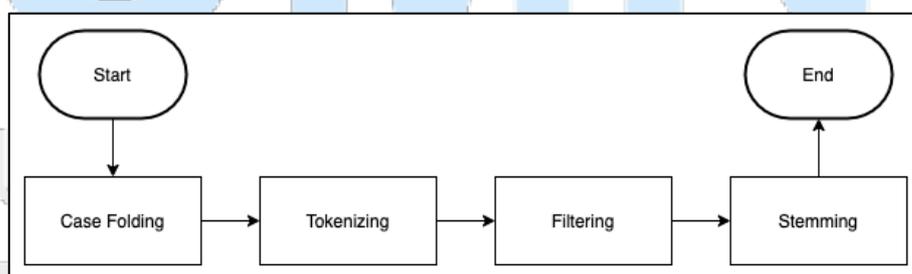
Langkah pertama dalam melakukan *text classification* berbasis *machine learning* adalah ekstraksi fitur untuk merepresentasikan teks menjadi vektor numerik. Selanjutnya vektor numerik akan digunakan oleh sistem untuk melakukan klasifikasi berdasarkan label yang benar (Huilgol, 2020).

### 2.3 Sentimen Analisis

Sentimen analisis adalah proses menentukan emosi atau pendapat seseorang, proses tersebut diekspresikan dalam bentuk teks dan dapat dibedakan menjadi emosi positif atau emosi negatif (Amrullah, dkk., 2016). Sentimen analisis mengacu pada berbagai pemrosesan *natural language*, komputasi linguistik, dan *text mining*. Analisis ini bertujuan untuk menganalisis pandangan, sentimen, evaluasi, sikap, penilaian, dan emosi pembicara atau penulis yang terkait dengan topik, produk, dan layanan, organisasi, individu atau kegiatan tertentu (Dhande & Patnaik, 2014). Sentimen analisis telah banyak digunakan, tidak hanya terbatas pada bidang penelitian ilmiah, tetapi juga untuk memenuhi kebutuhan pemasaran dan teknologi (Habibi, dkk, 2016).

### 2.4 Preprocessing

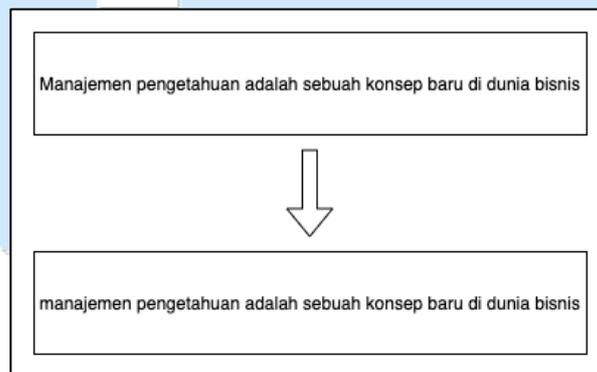
*Preprocessing* adalah teknik *data mining* yang melibatkan transformasi data mentah menjadi format yang mudah dipahami bagi komputer. Langkah *preprocessing* data diperlukan untuk menyelesaikan berbagai jenis masalah, termasuk data *noise*, redundansi data, dan data yang hilang (Gunawan, dkk., 2018). Adapun tahapan dalam melakukan *preprocessing* dapat dilihat pada Gambar 2.1.



Gambar 2.1 Alur *preprocessing*

## 1. *Case folding*

*Case folding* adalah proses mengubah semua huruf dalam dokumen atau kalimat menjadi huruf kecil. *Case folding* digunakan untuk memudahkan pencarian (Valatehan, dkk, 2016). Contoh dapat dilihat pada Gambar 2.2, Mengubah kalimat “Manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis” menjadi “manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis”.

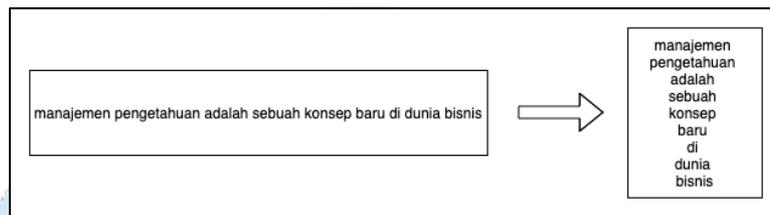


Gambar 2.2 Contoh penggunaan *case folding* (Ulina, 2016)

## 2. *Tokenizing*

*Tokenizing* merupakan sebuah proses memotong teks menjadi perkata berdasarkan teks penyusunnya (Muthia, 2018). Dengan tujuan untuk menghilangkan semua tanda baca atau simbol yang bukan huruf. Contohnya dapat dilihat pada Gambar 2.3, kalimat “manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis” dipecah menjadi “manajemen”, “pengetahuan”, “adalah”, “sebuah”, “konsep”, “baru”, “di”, “dunia”, “bisnis”.

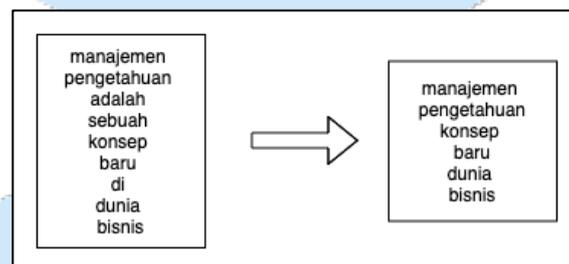
MULTIMEDIA  
NUSANTARA



Gambar 2.3 Contoh penggunaan *Tokenizing* (Ulina, 2016)

### 3. *Filtering*

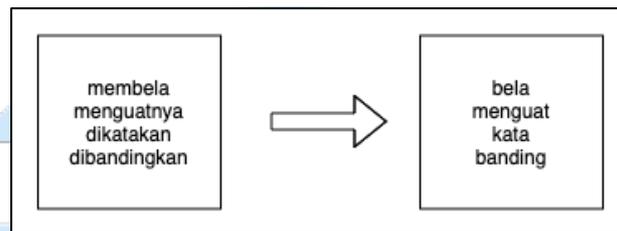
*Filtering* adalah proses menghilangkan kata-kata yang tidak relevan dengan kata sifat sentimen jika berdiri sendiri seperti kata “tetapi”, “untuk”, “adalah” dan “dengan” (Hasan & Wahyudi, 2018). Contoh dapat dilihat pada Gambar 2.4 token yang dimiliki adalah “manajemen”, “pengetahuan”, ”adalah”, “sebuah”, “konsep”, “baru”, “di”, “dunia”, “bisnis”. Maka token “adalah”, “baru”, dan “di” akan dihilangkan.



Gambar 2.4 Contoh penggunaan *filtering* (Ulina, 2016)

### 4. *Stemming*

*Stemming* adalah proses menghilangkan kata imbuhan dan hanya menyisakan kata dasarnya saja (Luqyana, dkk., 2018). Contoh *filtering* dapat dilihat pada Gambar 2.5, token “membela” diubah menjadi “bela” dengan menghilangkan imbuhan “mem-“ dan token “dibandingkan” diubah menjadi “banding” dengan menghilangkan imbuhan “di-kan”.



Gambar 2.5 Contoh penggunaan *stemming* (Ulina, 2016)

## 2.5 Term Frequency – Inversed Document Frequency (TF-IDF)

Term Frequency - Inversed Document Frequency (TF-IDF) adalah proses pemberian bobot pada setiap kata yang ada di sebuah dokumen. Metode TF-IDF mengurutkan kata berdasarkan kata yang paling banyak muncul (Gunawan, dkk., 2018). *Term Frequency* lebih menitik beratkan pada kata yang sering muncul pada dokumen sedangkan *Inversed Document Frequency* lebih berfokus menghitung kata yang sering muncul di sebuah dokumen yang dianggap kata umum yang tidak penting (Luqyana, dkk., 2018).

$$TF(t) = \frac{\text{jumlah kemunculan term } t \text{ dalam dokumen}}{\text{jumlah term dalam dokumen}} \quad (2.1)$$

$$IDF(t) = \log \left( \frac{\text{jumlah seluruh dokumen}}{\text{jumlah dokumen yang mengandung term } t} \right) \quad (2.2)$$

$$W(t) = TF(t) * IDF(t) \quad (2.3)$$

Dimana  $TF(t)$  adalah nilai *Term Frequency* dari *term*  $t$ ,  $IDF(t)$  adalah nilai *Inverse Document Frequency* dari *term*  $t$ , dan  $W(t)$  adalah bobot dari sebuah *term*.

## 2.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) dikembangkan oleh Chen dan Guestrin (2016). XGBoost adalah salah satu dari varian *boosting*. XGBoost dapat berjalan 10 kali lebih cepat dibandingkan dengan implementasi *Gradient Boosting* lainnya, sehingga telah banyak peneliti yang menggunakannya untuk klasifikasi dan regresi

dalam banyak kasus seperti prediksi penjual, prediksi kebiasaan pelanggan, prediksi iklan, dan prediksi web teks. (Chen dan Guestrin, 2016).

*Boosting* merupakan *ensemble technique* dimana model baru ditambahkan untuk memperbaiki kesalahan yang dilakukan oleh model sebelumnya. Model akan ditambahkan secara berurutan hingga tidak ada perbaikan lagi. *Ensemble technique* menggunakan *tree ensemble model* yang merupakan sekumpulan dari pohon klasifikasi dan regresi. Pendekatan *ensemble technique* menjumlahkan prediksi dari beberapa pohon menjadi satu (Sholahuddin & Abdullah, 2021). Hal ini bertujuan untuk mengambil setiap predictor secara berurutan dan memodelkannya berdasarkan *residual error* dari model sebelumnya. Proses awal saat dataset dimasukkan adalah membuat model awal dengan menggunakan dataset yang telah dipilih. Kemudian didapatkan nilai prediksi awal dan *residual error* dari model awal menggunakan persamaan 2.4 dan 2.5. Persamaan 2.4 digunakan untuk membuat model awal, sedangkan persamaan 2.5 digunakan untuk membuat model-model selanjutnya.

$$h_0(x) = \text{mean}(Y) \quad (2.4)$$

$$\hat{Y} = Y - h_0(x) \quad (2.5)$$

Dimana  $h_0(x)$  merupakan nilai prediksi awal dari model pertama dan  $\hat{Y}$  merupakan nilai dari residual error dari model awal. Setelah itu model kedua akan dibentuk menggunakan *residual error* model awal sehingga didapatkan nilai prediksi dari model kedua. Kemudian model ketiga akan dibentuk menggunakan *residual error* model awal dan kedua sehingga didapatkan nilai prediksi dari model ketiga. Proses ini akan terus berulang sebanyak  $n$  estimator yang telah ditetapkan. Algoritma ini

dinamakan *Gradient Boosting* yang bertujuan untuk memperkecil error saat membuat model baru (Handayani, dkk., 2017).

Sama seperti halnya *boosting*, XGBoost membuat kumpulan *decision tree* yang dimana modelnya akan bergantung pada model sebelumnya. Model pertama dalam XGBoost akan lemah dalam inisialisasi nilai prediksi kemudian dilakukan *update* bobot pada setiap model yang dibentuk sehingga menghasilkan nilai prediksi yang kuat. Nilai prediksi dari setiap model akan dijumlahkan kemudian dimasukkan ke dalam Persamaan 2.6 untuk meminimumkan fungsi objektifnya (Syukron, dkk., 2020).

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (2.6)$$

Dimana  $n$  adalah jumlah model yang ingin digunakan,  $l$  adalah fungsi untuk mengukur perbedaan antara prediksi target  $y_i$  dan  $\hat{y}_i$ ,  $f_t(x_i)$  merupakan model baru yang dibangun. Sedangkan  $\Omega$  adalah fungsi untuk membuat model menghindari *overfitting* (Syukron, dkk., 2020). Persamaan 2.6 digunakan saat digunakan untuk mencari nilai keseluruhan.

## 2.7 Evaluasi Performa

Metode yang digunakan dalam mengevaluasi performa dari model yang digunakan adalah *confusion matrix*. Setiap komponen dalam *confusion matrix* menunjukkan jumlah prediksi yang dibuat oleh model mengklasifikasikan dengan benar atau salah (Mohajon, 2020). Gambar 2.5 merupakan contoh *confusion matrix* dengan 2 klasifikasi biner.

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Gambar 2.6 Confusion Matrix klasifikasi biner (Mohajon, 2020)

Pada Gambar 2.5 terdapat 4 komponen utama yaitu TP (*True Positive*) merupakan jumlah data positif yang diprediksi dengan benar, TN (*True Negative*) merupakan jumlah data negatif yang diprediksi dengan benar, FP (*False Positive*) merupakan jumlah data negatif yang diprediksi dengan salah, FN (*False Negative*) merupakan jumlah data positif yang diprediksi dengan salah. Berdasarkan keempat komponen tersebut kita dapat mencari *Accuracy*, *Precision*, *Recall*, dan *F1 score*. Persamaan 2.7 merupakan rumus untuk mencari nilai *Accuracy*, *Accuracy* merupakan rasio dari tingkat seberapa akurat sistem dapat memprediksi dengan benar. Persamaan 2.8 merupakan rumus untuk mencari nilai *Precision*, *Precision* merupakan rasio dari jumlah data yang diprediksi positif dengan benar dibagi dengan semua data yang diprediksi positif. Persamaan 2.9 merupakan rumus untuk mencari nilai *Recall*, *Recall* merupakan ratio data kategori positif yang diprediksi positif dengan benar oleh sistem. Persamaan 2.10 merupakan rumus untuk mencari nilai *F1 score*, *F1 score* merupakan perbandingan rata-rata *Precision* dan *Recall* (Mohajon, 2020).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.7)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.8)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.9)$$

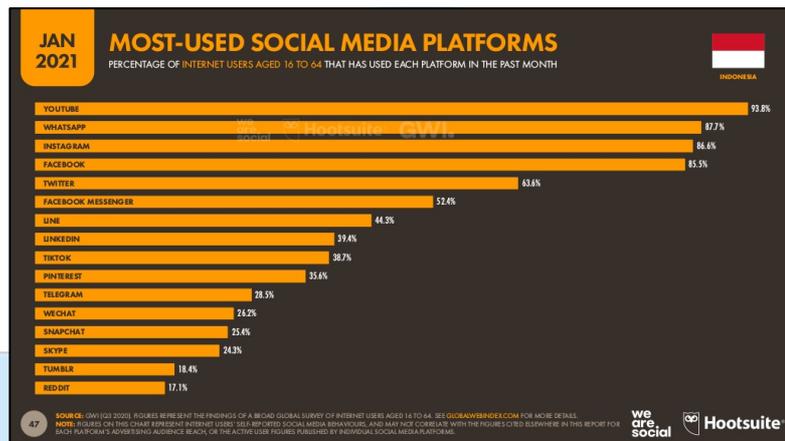
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.10)$$

## 2.8 Media Sosial

Media sosial merupakan media yang digunakan untuk berinteraksi secara *online* yang memungkinkan manusia untuk saling berkomunikasi tanpa dibatasi oleh ruang dan waktu (Rustian, 2012). Pengguna dapat saling berbagi informasi melalui teks, gambar, video, dan audio. Media sosial juga dapat digunakan sebagai sarana untuk membangun profil publik sehingga semakin banyak dikenal oleh orang lain (Dwiyono, 2019).

## 2.9 Instagram

Instagram adalah aplikasi untuk berbagi foto dan video, pengguna dapat membagikan momen miliknya kepada publik, pengguna juga bisa mengomentari postingan berupa foto atau video dari pengguna lainnya. Instagram menjadi sangat populer sehingga dimanfaatkan oleh penggunanya untuk menjadikan Instagram sebagai sarana dalam membangun profil publik (Shafita, 2018). Gambar 2.6 merupakan hasil survei yang dilakukan oleh Global Web Index, Instagram menempati peringkat 3 dengan persentase 86.6% pengguna dengan rentang usia 16 sampai 64 tahun (Kemp, 2021).



Gambar 2.7 Hasil Survei Media Sosial yang Sering Digunakan di Indonesia (Kemp, 2021)

## 2.10 Komentar

Komentar adalah tanggapan atau kesan yang ditulis untuk mengungkapkan pandangan mereka tentang suatu peristiwa atau momen tertentu. Komentar ditulis berdasarkan hati dan pikiran tiap individu ketika mereka berinteraksi dengan orang lain. Komentar dapat berupa tulisan atau lisan dalam bentuk pujian, sanggahan bahkan hinaan (Hartini, dkk., 2017).

UMMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA