BAB 2
LANDASAN TEORI

2.1 Tribun News

Tribunnews.com merupakan situs media online nomor satu di Indonesia dikelola oleh PT Tribun Digital Online, serta memiliki media jaringan yang tersebar di penjuru Indonesia, yaitu Tribun Network. Tribunnews.com yang berkantor pusat di Jakarta merupakan media akselerasi transformasi digital Indonesia, hadir untuk menyajikan informasi dari seluruh penjuru Indonesia dari Sabang hingga Merauke melalui jaringan Tribun Network. Jaringan Tribun Network didukung lebih dari 1,500 wartawan yang memberi informasi dengan nilai nilai lokal dari 34 Provinsi, melalui media online yang akan terus berkembang serta media cetak di berbagai daerah, ditambah dengan komunitas online Tribunners yang berada di seluruh penjuru indonesia [11].

2.2 Portal Berita

Portal berita adalah platform, umumnya berbasis web, yang mengumpulkan fakta dari sumber otentik yang dipilih. Portal berita memilih dan bekerja sama dengan sumber tertentu untuk memberikan perspektif yang berbeda dengan menyampaikan informasi yang kredibel dan sah kepada pelanggan. Portal Berita berperan sebagai platform *e-learning* dimana tempat seseorang dapat berbagi berita dari berbagai kategori, seperti administrasi, seni, film, musik, tren pasar, teknologi, dan lainnya. Berita ini diperbarui secara *real time* dan dapat diakses dengan mudah. Dengan kehadirannya secara daring, berita pada portal berita tidak bergantung pada lokasi dan waktu. Selain itu, portal berita dapat mencakup gambar, video, blog, dan lainnya [12].

2.3 Kata Penghubung

Kata penghubung atau kata konjungsi adalah salah satu ragam bahasa tulis yang digunakan pada setiap tulisan termasuk dalam penulisan berita utama dalam koran. Konjungsi adalah kata tugas yang menghubungkan dua satuan bahasa yang sederajat, yaitu kata dengan kata, frasa dengan frasa, atau klausa dengan klausa [13]. Konjungsi dibedakan macamnya sesuai dengan perilaku

sintaksis, ciri khusus, dan fungsinya. Terdapat 4 macam konjungsi yaitu konjungsi koordinatif, konjungsi subordinatif, konjungsi korelatif, dan konjungsi antarkalimat. Berdasarkan penggunaan secara peletakkan, jenis konjungsi yang dapat digunakan sebagai konjungsi awalan yaitu konjungsi subordinatif dan konjungsi antarkalimat. Sedangkan jenis konjungsi yang dapat digunakan sebagai konjungsi antara yang menghubungkan klausa atau frase pada kalimat yaitu konjungsi koordinatif dan konjungsi korelatif [14].

2.4 Praproses Kata

Praproses kata atau *text preprocessing* merupakan proses transformasi teks menjadi format yang lebih bersih dan konsisten sehingga kata dapat dimasukkan ke dalam model untuk analisis dan digunakan untuk pembelajaran lebih lanjut [15]. Praproses kata terdiri dari beberapa teknik yang dapat digunakan sesuai dengan jenis dan kondisi data yang ingin dilakukan praproses. Beberapa teknik pada praproses kata yaitu seperti *case folding* dan *tokenization*.

Case Folding adalah tahap untuk konversi kata menjadi suatu bentuk yang standar. *Case Folding* bertujuan untuk menghilangkan batasan batasan yang terdapat pada kata seperti huruf kecil atau besar, spasi, dan tanda baca. Terdapat beberapa cara yang dapat digunakan dalam tahap *case folding* tergantung dengan tujuan yang ingin dicapai. Cara yang umum dilakukan pada tahap *case folding* yaitu mengubah kata menjadi huruf kecil atau *lowercase* dengan tujuan untuk normalisasi penggunaan huruf kapital pada data [16]. Selain itu, karakter lain yang bukan termasuk huruf dan angka, seperti tanda baca dan spasi dianggap sebagai batasan yang akan dihilangkan pada tahap *case folding*.

Tokenization atau tokenisasi bertujuan untuk memecahkan teks menjadi bagian yang lebih kecil yang disebut sebagai token agar proses analisis dapat lebih mudah dilakukan [17]. Tahap *tokenization* umumnya digunakan untuk memisahkan kata dalam sebuah kalimat. Namun, menggunakan konsep yang sama, *tokenization* juga dapat dilakukan untuk memisahkan kalimat dalam sebuah paragraf.

2.5 Cosine Similarity

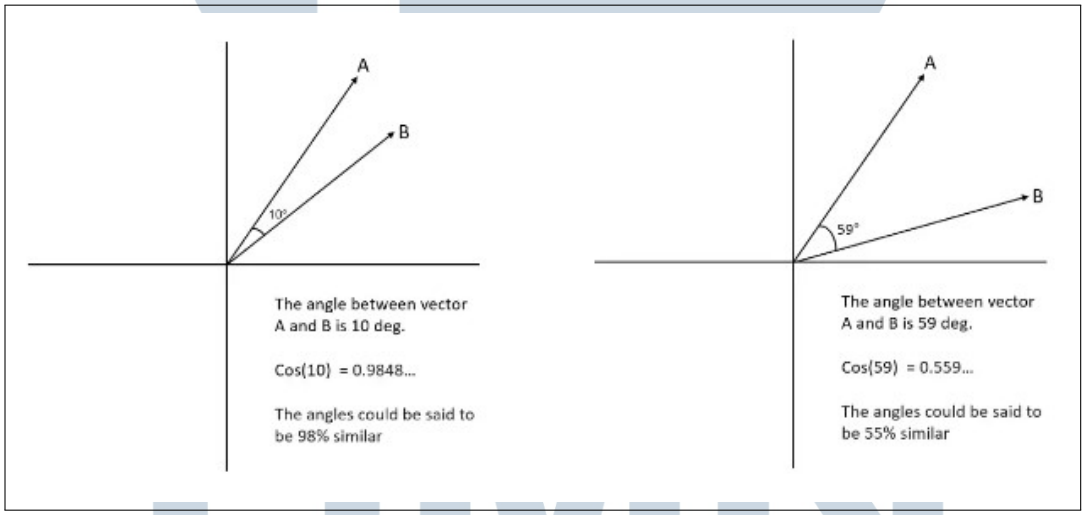
Cosine Similarity adalah ukuran kesamaan antara dua buah vektor dalam sebuah ruang dimensi yang didapat dari nilai cosinus sudut dari pertambahan dua buah vektor yang dibandingkan karena kosinus dari 0 derajat adalah 1 dan kurang

dari 1 untuk sudut sudut yang lain, maka nilai kesamaan dari dua buah vektor dikatakan mirip ketika nilai dari *cosine similarity* adalah 1 [18]. Kosinus dari dua vektor dapat diturunkan dengan menggunakan rumus perkalian titik *Euclidean*.

$$A \cdot B = ||A||\ ||B||\ cos(\theta) \tag{2.1}$$

Berdasarkan rumus 2.1 yang merupakan perkalian titik Euclidean, maka didapatkan rumus *cosine similarity* melalui perkalian titik dan besaran *magnitude* sebagai berikut.

$$Similarity = cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\Sigma_{i=1}^{n} A_i B_i}{\sqrt{\Sigma_{i=1}^{n} A_i^2}\sqrt{\Sigma_{i=1}^{n} B_i^2}} \tag{2.2}$$



Gambar 2.1. Nilai Similaritas Vektor Terhadap Sudut Cosinus
Sumber : Richmond, 2020

Pada rumus 2.2, A_i dan B_i masing-masing merupakan bagian dari vektor A dan B. Hasil dari kosinus umumnya memiliki rentang nilai dari -1 hingga 1, tetapi vektor dokumen umumnya bernilai positif dikarenakan sudut antara dua dokumen tidak pernah lebih besar dari 90 derajat. Gambar 2.1 menjelaskan posisi vektor yang membentuk derajat similaritas. Pengukuran nilai similaritas didapatkan melalui pengukuran nilai kosinus sudut antara dua vektor. Nilai similaritas dari derajat yang dibentuk oleh vektor dokumen memiliki rentang nilai dari 0 hingga 1. Semakin nilai *cosine similarity* mendekati 1, maka sudut antara vektor akan semakin kecil. Nilai 0 menandakan bahwa dokumen bersinggungan dengan kosinus (*orthogonal*) atau

tidak memiliki kesamaan dan nilai 1 menandakan bahwa dokumen tersebut sama [19].

2.6 Confusion Matrix

Pada bidang pembelajaran mesin dan khususnya masalah statistika, *confusion matrix* yang juga dikenal sebagai matriks kesalahan merupakan sebuah tabel yang memvisualisasikan performa dari model algoritma [20]. *Confusion matrix* adalah tabel 2x2 yang terdiri dari 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Terdapat empat variabel yang merupakan hasil proses klasifikasi pada *confusion matrix* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Nilai *True Positive* (TP) merupakan data positif yang terdeteksi benar. *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar. Sementara itu, *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif dan *False Negative* (FN) merupakan data positif namun terdeteksi sebagai data negatif [21]. Berikut adalah contoh tabel *confusion matrix*.

N	Aktual	
Prediksi	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Tabel 2.1. Tabel Confusion Matrix
Sumber: Narkhede, 2018

Berdasarkan hasil dari setiap variabel pada *confusion matrix*, didapatkan beberapa perhitungan penilaian yang digunakan untuk menguji performa model sebagai berikut.

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \tag{2.3}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.4}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.5}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2.6)$$

Penilaian yang dihitung terdiri dari *accuracy*, *precision*, *recall* dan F1 *score*. *Accuracy* merupakan penilaian yang menandakan tingkat kedekatan antara nilai prediksi dengan nilai aktual. *Precision* merupakan penilaian yang menandakan tingkat presisi prediksi model positif dengan nilai total positif prediksi. *Recall* merupakan penilaian yang menandakan tingkat kedekatan prediksi model positif dengan nilai total positif aktual. F1 *score* merupakan penilaian yang menggunakan perhitungan harmonisasi antara *precision* dan *recall* untuk mendapatkan tingkat akurasi yang lebih baik dalam pendistribusian matriks data yang relatif acak.