BAB II
TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan penilaian publik, penilaian mengenai kesukaan atau ketidaksukaan dan dapat menentukan terhadap komentar mengenai permasalahan, apakah masalah tersebut memiliki kecenderungan positif, negatif, maupun netral. Tujuan dari teks mining adalah memproses unstructured data (tekstual) guna dicari pola makna serta ditindaklanjuti dengan pengambilan keputusan yang terbaik. (Sari, 2020).

2.2 Aplikasi Pedulilindungi

Aplikasi pedulilindungi terbilang baru karena diresmikan pada tahun 2020 yang lalu. Aplikasi ini merupakan gagasan dari Kominfo, Gugus Tugas Covid-19 yang mengkoordinasikan Kementerian BUMN, BNPB, Kemenkes, TNI, Polri, dan Kementerian Pendayagunaan Aparatur Negara dan Reformasi Birkorasi yang dalam hal ini didasari oleh Keputusan Menteri Kominfo Nomor 171 Tahun 2020 dimana memuat tentang Penetapan Aplikasi Pedulilindungi Dalam Rangka Pelaksanaan Surveilans Kesehatan Penanganan Corona Virus Disease 2019 (COVID-19). Aplikasi pedulilindungi pada awalnya digunakan agar dapat membantu pemerintah dalam melaksanakan tracing dan tracking guna mengidentifikasi siapa saja orang-orang yang memerlukan penanganan akibat Coronavirus.

Aplikasi Pedulilindungi ini masih terus dikembangkan dengan melengkapi beragam fitur yang dapat mempermudah masyarakat di era kenormalan baru, salah satunya dengan akan ditambahkannya fitur e-passport sebagai dokumen perjalanan dalam bentuk data pengguna yang telah dinyatakan negatif tes Covid-19. Pemerintah juga akan meningkatkan aplikasi Pedulilindungi agar bisa

digunakan oleh perangkat telepon non-smartphone. Sehingga nantinya aplikasi ini bisa juga digunakan oleh pengguan telepon melalui teknologi SMS. Selain itu pemerintah juga membuat dashboard monitoring yang digunakan untuk *tracing*, *tracking*, dan *fancing*. Dalam hal ini dashboard *tracing* dan *tracking* untuk melihat user yang pernah *closed-contact* dengan pasien positif. Dashboard fencing untuk melihat pergerakan orang dalam karantina mandiri. Untuk memontoring tersedia di Kementerian Kesehatan yang dapat digunakan khususnya untuk memonitor pasien dan orang yang melakukan karantina mandiri. Adapun cara kerja aplikasi PeduliLindungi bagi masyarakat yang sudah mengunduh aplikasi ini diminta untuk mengaktifkan bluetooth pada ponselnya. Diharapkan dengan kondisi beluetooth aktif, maka secara berkala aplikasi akan melakukan identifikasi ponsel pengguna PeduliLindungi lainnya, yang berada di radius bluetooth untuk merekam lokasi dan waktu kontaknya. Mekanismenya ponsel-ponsel yang berdekatan kemudian akan saling merekan ID anonim masing-masing, data Anonim ID tersebut akan disimpan dalam rentan waktu 14 hari, sehingga apabila ada seseorang yang dinyatakan sakit oleh petugas kesehatan (bukan oleh aplikasi) dan diinput kedalam sistem database, maka sistem akan memfilter ID-ID anonim lain yang terekam pernah melakukan kontak dengan penderita Covid-19 dalam waktu 14 hari terakhir. (Fastyaningsih et al., 2021).

2.3 Confusion Matrix

Confusion matrix adalah matrix 3x3 yang merepresentasikan hasil klasifikasi biner pada suatu dataset. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai *accuracy*, *precision* dan *recall* bisa ditampilkan dalam persentase (Andika, 2019).

a. *Accuracy*. *Accuracy* adalah jumlah proporsi prediksi yang benar. Adapun rumus perhitungan akurasi dapat dilihat dari persamaan (2.1) di bawah ini:

$$\boldsymbol{Accuracy = \frac{TP+TN}{TP+FP+FN+TN}} \qquad \textbf{(2.1)}$$

b. *Precision*. *Precision* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks yang terpilih oleh sistem. Rumus precision dapat dilihat pada persamaan (2.2) di bawah ini:

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \qquad \textbf{(2.2)}$$

c. *Recall*. *Recall* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi. Rumus recall dapat dinyatakan dengan persamaan (2.3) di bawah ini::

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \qquad \textbf{(2.3)}$$

Tabel 2.1 Confusion Matrix

	KLASIFIKASI SISTEM		
KLASIFKASI MANUAL	Positif	Negatif	Netral
Positif	TP	FN	FN
Negatif	FP	TN	TN
Netral	FP	TN	TN

Dimana :

1. TP adalah True Positif, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
2. TN adalah True Negatif, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
3. FN adalah False Negatif, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
4. FP adalah False Positif, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

2.4 Pre-processing Data

Dokumen pada umumnya mempunyai struktur yang sembarangan atau tidak terstruktur. Oleh karena itu, diperlukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur ke dalam bentuk data yang terstruktur. Tahap preprocessing memiliki beberapa proses, yaitu case folding,stopwords removing, tokenizing, dan stemming. Selanjutnya data yang sudah mengalami preprocessing akan diubah menjadi bentuk numerik dengan tahap term weighting. Pada penelitian ini terdapat tiga metode term weighting yang digunakan, yaitu term frequency, inverse document frequency, dan term frequencyinverse document frequency (Jumeilah, 2017). Tahapan dalam pre-processing adalah sebagai berikut:

2.4.1 Cleansing

Proses membersihkan dokumen dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikokn emosi, hashtag (#), username (@username), url (http://situs.com), dan email (nama@situs.com).

2.4.2 Case folding

Case folding yaitu pengubahan bentuk huruf menjadi huruf kecil.

2.4.3 Tokenizing

Tokenizing adalah proses memecah teks menjadi kata tunggal dan penghapusan tanda baca serta angka, sesuai dengan kamus data yang telah ditentukan. Pada penelitian ini fitur yang digunakan dalam memecah teks adalah unigram yaitu token yang terdiri hanya satu kata.

2.4.4 Stopword removal

Stopwords removal adalah proses menghilangkan kata tidak penting dalam text. Hal ini dilakukan untuk memperbesar akurasi dari pembobotan term. Untuk proses ini, diperlukan suatu kamus kata-kata yang bisa dihilangkan. Dalam Bahasa Indonesia, misalnya kata: dan, atau, mungkin, ini, itu, dll adalah kata-kata yang dapat dihilangkan.

2.4.5 Stemming

Stemming adalah pengubahan kata ke bentuk kata dasar atau penghapusan imbuhan. *Stemming* disini menggunakan kamus daftar kata berimbuhan yang mempunyai kata dasarnya dengan cara membandingkan kata-kata yang ada di dalam kokmentar dengan daftar kamus stem.

2.5 Term Frequency Inverse Document Frequency (TF – IDF)

Metode TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen (Nurjannah & Fitri Astuti, 2013). Pembobotan menggunakan TF-IDF dijelaskan pada Persamaan (2.4) dan (2.5).

$$IDF\ (w) = log\left(\frac{N}{DF(w)}\right) \qquad \textbf{(2.4)}$$

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \qquad \textbf{(2.5)}$$

Keterangan:

TF-IDF (w,d)	: bobot suatu kata dalam keseluruhan dokumen
W	: suatu kata (word)
d	: suatu dokumen
Tf (w,d)	: frequensi kemunculan sebuah kata w dalam dokumen
IDF (w)	: inverse DF dari kata W
N	: jumlah keseluruhan dokumen
DF (w)	: jumlah dokumen yang mengandung kata w

2.6 Cosine Similarity

Metode *Cosine Similarity* adalah mengukur kemiripan antara dua dokumen atau teks. Pada *Cosine Similarity* dokumen atau teks dianggap sebagai vector. Pada penelitian ini, *Cosine Similarity* digunakan untuk menghitung jumlah kata istilah yang muncul pada halaman-halaman yang diacu pada daftar indeks. Semakin banyak jumlah kata istilah yang muncul pada suatu halaman semakin tinggi nilai *Cosine Similarity* yang diperoleh (Wahyuni et al., 2017). Rumus yang digunakan untuk menghitung cosine similarity dapat kita lihat pada persamaan (2.6) di bawah ini:

$$cosSim(X, dj) = \frac{\sum_{i}^{m}=1 xi.dji}{\sqrt{(\sum_{i}^{m}=1\ xi).\ \sqrt{(\sum_{i}^{m}=1\ dji)}}} \qquad \textbf{(2.6)}$$

Metode pengklasifikasian yang digunakan pada sistem ini adalah dengan cara membandingkan kesamaan atau similaritas antara judul dokumen dengan kata kunci pertama, kemudian cara membandingkan kesamaan atau similaritas antara judul dokumen dengan kata kunci kedua, begitu seterusnya hingga kata kunci kedelapan. Kemudian dicari jumlah similaritas yang tertinggi antara kedelapan kata kunci. Apabila total similaritas yang didapatkan adalah nol (0) maka dokumen akan masuk ke dalam kategori kesembilan. Dokumen yang diolah nantinya akan diklasifikasikan secara otomatis oleh sistem ke dalam 9 kategori yang berbeda.

2.7 Improved K-nearest neighbor

Penentuan k-values yang tepat diperlukan agar didapatkan akurasi yang tinggi dalam proses kategorisasi dokumen uji. Algoritma Improved k-Nearest Neighbor melakukan modifikasi dalam penentuan k-values. Dimana penetapan k-values tetap dilakukan, hanya saja tiap-tiap kategori memiliki k-values yang berbeda. Perbedaan k-values yang dimiliki pada setiap kategori disesuaikan dengan besar-kecilnya jumlah dokumen latih yang dimiliki kategori tersebut. Sehingga ketika k-values semakin tinggi, hasil kategori tidak terpengaruh pada kategori yang memiliki jumlah dokumen latih yang lebih besar. Selanjutnya pada algoritma *Improved k-Nearest Neighbor*, *k-values* yang baru disebut dengan n (Windiarti, 2018). Persamaan (2.7) menjelaskan mengenai proporsi penetapan *k-values* (n) pada setiap kategori.

$$n = \left[\frac{k \div N(cm)}{maks\{N(c_m) | j=1 \ldots N_C}\right] \quad \textbf{(2.7)}$$

Keterangan:
n: k-values baru,
k: k-values yang ditetapkan,
N (Cm): jumlah dokumen latih di kategori/kategori m,
Maks {N(Cm){j=1...Nc}: jumlah dokumen latih terbanyak pada semua kategori.

Dalam menentukan kategori untuk dokumen uji menggunakan algoritma *Improved k-Nearest Neighbors*, maka dilakukan perbandingan kemiripan dokumen uji dengan dokumen latih pada tiap kategori. Persamaan (2.8) menyatakan nilai maksimum perbandingan antara kemiripan dokumen X dengan dokumen latih dj sejumlah top n tetangga pada suatu kategori dengan kemiripan dokumen X dengan dokumen latih di sejumlah top n tetangga pada training set.

$$\rho(\chi, c_m) = argmax_m \frac{\sum d_j \in top\ n\ k\ NN(c_m) sim(x,d_j) y(d_j,c_m)}{\sum d_j \in top\ n\ k\ NN(c_m) sim(x,d_j)} \quad \textbf{(2.8)}$$

Dimana:

p(x,cm) : probabilitas dokumen X menjadi anggota kategori cm

sim(x,dj) : kemiripan antara dokumen X dengan dokumen latih dj

top n kNN : top n tetangga

y(dj.cm) : fungsi atribut dari kategori yang memenuhi persamaan.

Adapun langkah-langkah untuk klasifikasi dokumen X menggunakan algoritma Improved K-Nearest Neighbor adalah sebagai berikut:

1. Melakukan tahapan pre-prosesing sehingga didapatkan representasi dari dokumen X dan semua dokumen latih.
2. Setelah terbentuk vektor, hitung bobot masing-masing dokumen menggunakan TF-IDF.
3. Hitung nilai cosine similarity dokumen X dengan semua dokumen latih.
4. Selanjutnya urutkan hasil dari perhitungan nilai cosine similarity secara menurun. Nilai yang lebih tinggi menunjukkan bahwa di antara dokumen X dan dokumen latih tersebut memiliki kemiripan.
5. Mengelompokkan hasil dari perhitungan nilai cosine similarity berdasarkan kategorinya.
6. Menentukan k-values kemudian melakukan perhitungan penetapan k-values baru (n) pada masing-masing kategori menggunakan persamaan (2.5).
7. Setelah didapatkan nilai n yang menyatakan sebagai top tetangga dari langkah 6, maka langkah selanjutnya adalah menentukan kategori dokumen X berdasarkan hasil perhitungan menggunakan persamaan (2.6).
8. Berdasarkan perhitungan pada persamaan (2.6), maka dokumen X akan dikategorikan ke dalam kategori yang memiliki P(x,cm) terbesar.

2.8 Tinjauan Pustaka

Dalam penelitian skripsi ini, peneliti terinspirasi dan mereferensi dari penelitian-penelitian sbelumnya yang berkaitan dengan skripsi ini. Daftar penelitian terkait ialah sebagai berikut:

1. **Tinjauan Literatur 1**

   Willa Oktinas (2017) menguraikan bahwa, Sentimen masyarakat dapat dijadikan sebagai salah satu indikator oleh stasiun televisi untuk menentukan kualitas suatu acara. Pada twitter dapat dilakukan proses penggalian informasi mengenai sentimen masyarakat terhadap kualitas acara yang ditayangkan. Salah satu teknik penggalian informasi pada twitter adalah analisis sentimen. Pada penelitian ini terdiri dari 3 tahapan proses analisis sentimen. Tahap pertama yaitu proses pre-pocessing yang terdiri dari *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, dan *filter redudansi*. Selanjutnya pada tahap kedua yaitu proses perhitungan bobot pada setiap kata menggunakan metode TF-IDF. Tahap terakhir yaitu proses klasifikasi sentimen menjadi 2 kategori yaitu sentiment positif dan negatif menggunakan metode *improved k-nearest neighbor*. Hasil yang diperoleh dari pengujian analisis sentimen berbahasa Indonesia dengan metode *Improved K Nearest Neighbor* menghasilkan akurasi tertinggi dengan nilai k=10 sebesar 90%.

2. **Tinjauan Literatur 2**

   Nurjannah & Fitri Astuti (2013) menguraikan, Algoritma Term Frequency Inverse-Document Frequency merupakan suatu algoritma yang menggalikan antara Term frequency dengan Inverse Document Frequency. Term frequency yaitu jumlah kemunculan sebuah term pada sebuah dokumen. Inverse Document Frequency yaitu pengurangan dominasi term yang sering muncul diberbagai dokumen, dengan memperhitungkan kebalikan frekuensi dokumen yang mengandung suatu kata. Text Mining pada umumnya adalah unstructured data, atau minimal semistructured. Maka merupakan tantangan tambahan pada text mining yaitu struktur teks yang kokmpleks dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah translasi yang tidak akurat. Hasil dari penelitian menunjukan bahwa, penerapkan algoritma term

frequency inverse-document frequency untuk text mining sangat membantu pengguna. Untuk mendapatkan informasi pada kumpulan dokumen. Dengan format file txt berdasarkan kata kunci yang dimasukan oleh pengguna pada sistem. Dengan kokleksi uji kata ‘upaya’ pada query maka didapatkan keluaran dengan bobot nilai 86% yang merupakan jumlah kata terbanyak sesuai dengan query.

3. Tinjauan Literatur 3

(Wahyuni (2017) menguraikan, banyaknya arsip dokumen skripsi yang terkumpul dalam bentuk soft file yang tidak terklasifikasi dengan baik mengakibatkan proses pencarian kembali menjadi sulit. Untuk mengakses informasi yang dibutuhkan menjadi kurang cepat dan tepat apabila keseluruhan dokumen disimpan dalam satu folder database. Maka dari itu diperlukan suatu sistem yang dapat mengklasifikasikan dokumen secara otomatis ke dalam folder berbeda pada database agar lebih mudah dalam mengelola dokumen yang ada. Metode TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode cosine similarity merupakan metode untuk menghitung kesamaan antara dua buah objek yang dinyatakan dalam dua buah vector dengan menggunakan keywords (kata kunci) dari sebuah dokumen sebagai ukuran. Metode pengembangan sistem yang digunakan dalam penelitian ini adalah model waterfall, sedangkan metode penelitian yang digunakan adalah metode Research and Development (R&D). Hasil penelitian menunjukkan bahwa persentase tingkat ketepatan klasifikasi sistem adalah sebesar 98%.

Tabel 2.2 Tinjauan Pustaka

Nomor	Penulis	Tahun	Metode	Judul	Hasil Penelitian
1	Willa Oktinas	2017	*Improved K-Nearest Neighbor*	Analisis sentimen pada acara televisi menggunakan *Improved k-nearest neighbor*	90%
2	Nurjannah & Fitri Astuti	2013	*Term Frequency inverse document* (tf-idf)	Penerapan algoritma *term frequency-inverse document frecuency* (tf-idf) untuk *text mining*	86%
3	Wahyuni	2017	*Research and Development* (R&D).	Penerapan algoritma *Cosine Similarity* dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi.	98%