Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Perpustakaan Univeristas Multimedia Nusantara

BAB III
METODE PENELITIAN

3.1. Objek Penelitian

Objek dari penelitian ini adalah *cyberbullying* yang terjadi di Indonesia. *Cyberbullying* adalah sebuah perundungan yang di lakukan pada media *digital*, perundungan atau *bullying* adalah perilaku atau tindakan ancaman, kekerasan atau paksaan untuk mengintimidasi orang lain dan bisa berulang kali. *Cyberbullying* biasanya terjadi di media sosial, *messaging apps*, *online chatting*, forum *online*, *email*, dan komunitas permainan online. Penelitian ini membahas *cyberbullying* yang dilakukan pada media sosial, yaitu *twitter*.

Twitter merupakan layanan untuk teman, keluarga, dan teman sekerja untuk berkomunikasi dan tetap terhubung melalui pertukaran pesan yang cepat dan sering. Pengguna memposting *tweet*, yang dapat berisi foto, video, tautan, dan teks. *Tweet* ini diposting ke profil, lalu terkirim ke pengikut, dan dapat di cari di pencarian *twitter*.

3.2. Metode Penelitian

Dari banyak metode yang ada, dipilihlah 2 metode yaitu *support vector machine* dan *information gain* yang digunakan untuk melakukan identifikasi pada *tweet* mengenai *cyberbullying* di Indonesia. *Support vector machine* (SVM) digunakan untuk memaksimalkan *dataset* yang berdimensi tinggi, kemudian untuk *information gain* jika data semakin banyak maka nilai *entropy* yang semakin rendah dan dapat meminimalisir kejutan. Jika dijelaskan lebih rinci, yaitu :

3.2.1. Support Vector Machine

Support Vector Machine (SVM) memiliki konsep yaitu mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua kelas data. SVM memaksimalkan *margin*, yang merupakan jarak pemisah antara kelas data.

SVM juga mampu berkerja pada dataset yang berdimensi tinggi dengan menggunakan *kernel trick*. Ada beberapa macam fungsi *kernel* SVM, yaitu *Linear, Polynomial, Gaussian* RBF, *Sigmoid*, Invers Multi Kuadratik, dan *Additive.*

Pada penelitian ini fungsi *kernel* yang digunakan adalah SVM *Polynomial.* SVM *linear* digunakan ketika ada data yang akan diklasifikasi terpisah dengan sebuah *hyperplane*, sedangkan SVM non-*linear* digunakan ketika data hanya dapat dipisahkan dengan garis lengkung. SVM *Polynomial* memiliki definisi fungsi dengan persamaan $K(\vec{x}_{\mathrm{I}}, \vec{x}_j) = (\vec{x}_{\mathrm{i}} \cdot \vec{x}_j + 1)^d$ yang dimana $K(\vec{x}_{\mathrm{I}}, \vec{x}_j)$ merupakan fungsi *kernel*, x merpuakan fitur dan *d* merupakan ordo.

Hyperplane dalam SVM yang optimal didapatkan dengan cara merumuskan ke dalam *quadratic programing problem* (QP *problem*) dan dapat di selesaikan menggunakan library yang banyak tersedia dalam *numeric analyst*. Tetapi terdapat sebuah alternative yaitu menggunakan metode sequential. Metode ini dikembangkan oleh Vijayakumar untuk mencari nilai α, yang dapat di uraikan, sebagai berikut :

1. Inisialisasi $\alpha_i = 0$

Menghitung nilai matriks *hessian* dengan menggunakan persamaan :

$$D_{ij} = y_i y_j (K(x_i . x_j) + \lambda^2)$$

Rumus 3.1. Menghitung Nilai Matriks *Hessian*

Dimana y merupakan kelas dari data ke - i dan ke – j, $K(x_i . x_j)$ merupakan fungsi kernel *polynomial* yang digunakan.

2. Menghitung setiap level dengan tahapan menggunakan persamaan :

$$E_i = \sum_{j=1}^{n} \alpha_i D_{ij}$$

$$\delta\alpha_i = \min\{ \max[\gamma\ (1 - E_i), C - \alpha_i\}$$

$$\propto_i = \propto_i + \delta\alpha_i$$

Rumus 3.2. Menghitung Persamaan

3. Melakukan pengulangan ke tahap 2 sampai nilai $\propto$ mencapai konvergen.

3.2.1. Information Gain

Information Gain (IG) merupakan salah satu metode untuk melakukan seleksi fitur, yang biasa digunakan oleh para peneliti untuk menentukan batas dari kepentingan sebuah atribut. Nilai IG diperoleh dari nilai *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan. Nilai ini digunakan untuk penentuan atribut mana yang akan dibuang dan digunakan. Atribut yang memenuhi kriteria pembobotan nantinya akan digunakan untuk proses klasifikasi.

Dalam pemilihan fitur dengan IG dilakukan dalam 3 tahapan, yaitu :

1. Menghitung nilai *Information Gain* untuk setiap atribut.
2. Menentukan *threshold* atau batasan. Hal ini untuk menentukan atribut yang bobotnya lebih kecil dari *threshold* akan dibuang.
3. Memperbaiki dataset dengan pengurangan atribut.

Seleksi fitur *Information Gain* dirumuskan menjadi :

$$IG(t) = -\sum_{i=1}^{|C|} P(C_i) \log P(C_i) +$$

$$P(t) \sum_{i=1}^{|C|} P(C_i|t) \log P(C_i|t) +$$

$$P(\bar{t}) \sum_{i=1}^{|C|} P(C_i|\bar{t}) \log P(C_i|\bar{t})$$

Rumus 3.3. Seleksi Fitur *Information Gain*

Dimana C_i merupakan kelas data, $P(C_i)$ merupakan peluang dari kelas data, $P(t)$ dan $P(\bar{t})$ merupakan peluang *term* t yang muncul atau tidak muncul dalam dokumen.

3.3. Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan hanya satu, yaitu melalui *text mining* yang dilakukan pada media sosial *twitter*.

3.4. Teknik Pengolahan Data

3.4.1. *Text Pre-processing*

Text preprocessing merupakan proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses yang lebih lanjut. *Text preprocessing* sendiri di bagi menjadi beberapa tahap, yaitu :

5. *Case Folding*

   Tahap ini mengubah kata pada document menjadi huruf kecil atau *lowercase*.

6. *Tokenizing*

   Pada tahap *tokenizing* atau tokenisasi melakukan proses pemotongan *string* atau kata dan penghilangan tanda baca pada kalimat yang sudah di jadikan huruf kecil pada tahap *case folding*.

7. *Filtering*

   Lalu pada tahap *filtering* dilakukan pengambilan kata – kata penting dari hasil *tokenizing*. Pada tahap ini bisa digunakan algoritma *stoplist* atau *wordlist*. Contoh kata – kata *stoplist* adalah “yang”, “dan”, “di”, dan sebagainya.

8. *Stemming*

   Teknik *stemming* merupakan tahap lanjutan dari tahap *filtering* yaitu untuk mengetahui akar sebuah kata. *Stemming* dilakukan agar memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga dapat digunakan untuk mengelompokan kata – kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk yang berbeda karena memiliki imbuhan yang berbeda.

9. *Term Weighting*

Pada tahap *term weighting* dilakukan proses penghitungan jumlah bobot setiap kata hasil *stemming* yang akan digunakan untuk membandingkan *frequency* dari jumlah kemunculan token tersebut. Rumus *term weighting* yang di gunakan yaitu TF-IDF (*Term frequency – Inverse Document Frequency*). Metode TF-IDF ini merupakan metode untuk menentukan seberapa pentingnya sebuah kata terhadap sebuah dokumen, dan terhadap sebuah dokumen korpus. Metode ini bekerja dengan cara menghitung jumlah kata unik dalam satu dokumen, dan dibandingkan dengan jumlah total kata pada document dimana metode ini diterapkan.

$$W_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d+} \\ 0, \end{cases}$$

$$if\ tf_{t,d} > 0$$

$$if\ tf_{t,d} = 0$$

Rumus 3.4. Rumus TF-IDF untuk menghitung persamaan (1)

Dimana tf adalah *term frequency* yang menyatakan berapa banyak jumlah suatu term dalam sebuah dokumen.

$$idf_t = \log_{10}(\frac{N}{df_t})$$

Rumus 3.5. Rumus TF-IDF untuk menghitung persamaan (2)

Dimana N merupakan jumlah banyaknya dokumen, karena terkadang suatu *term* muncul pada beberapa dokumen sehingga terkadang proses pencarian *term* dapat terganggu.

10. Hasil Keluaran

Hasil keluaran dari penelitian ini adalah hasil identifikasi *tweet* yang mengandung konten *cyberbullying*.