***Term Frequency - Inverse Document Frequency* (TF-IDF)**

Pembobotan TF-IDF adalah perhitungan yang menggambarkan pentingnya sebuah kata (term) dalam sebuah dokumen dan sebuah corpus. Proses ini memungkinkan untuk menilai relevansi terminologi dokumen dengan semua dokumen dalam korpus [12]. Secara intuitif, perhitungan ini menentukan seberapa relevan kata yang diberikan dalam dokumen tertentu. Kata-kata yang umum dalam satu atau sekelompok kecil dokumen cenderung memiliki angka TF-IDF yang lebih tinggi daripada kata-kata umum seperti artikel dan preposisi [11]. Frekuensi kemunculan kata dalam dokumen tertentu menunjukkan pentingnya kata tersebut dalam dokumen. Frekuensi dokumen yang berisi kata tersebut menunjukkan seberapa populer kata tersebut. Bobot kata lebih besar jika sering muncul dalam dokumen dan lebih kecil jika muncul di banyak dokumen [13]. Algoritma TF-IDF menggunakan rumus untuk menghitung bobot (W) setiap dokumen terhadap kata kunci dengan rumus yaitu :

$$W_{ij} = t\, f_{ij} * Idf_j \text{, dimana } Idf_j = (\log\left(\frac{N}{df}\right)) \quad (1)$$

Dengan W_{ij} merupakan bobot dokumen ke-i terhadap kata ke-j, tf_{ij} merupakan banyaknya kata yang dicari dalam sebuah dokumen, Idf_j merupakan Inversed Document Frequency, N merupakan total dokumen, dan df merupakan banyak dokumen yang mengandung kata yang dicari. Setelah bobot W dari setiap dokumen diketahui, maka proses penyortiran dilakukan dimana semakin besar nilai W, semakin besar tingkat kemiripan dokumen dengan kata kunci, begitu sebaliknya.

***Support Vector Machine* (SVM)**

Support Vector Machine merupakan metode supervised learning. SVM dapat digunakan untuk regresi dan klasifikasi [3]. Penggunaan algoritma SVM, yang bertujuan untuk klasifikasi teks menggunakan bobot indeks istilah sebagai fitur pertama kali dirintis oleh Thorsten Joachim. Pembelajaran SVM telah dipopulerkan sejak tahun 1992 oleh Boser, Guvon dan Vapnik [14].

SVM dapat menyelesaikan permasalahan secara linier maupun non-linier. Memecahkan masalah nonlinier menggunakan konsep kernel di ruang kerja berdimensi tinggi, pencarian *hyperplane* mampu memaksimalkan margin antara lapisan data. Hyperplane berguna dalam memisahkan 2 kelompok kelas +1 dan kelas -1 dimana setiap kelas memiliki *pattern* masing-masing. Terdapat fungsi kernel yang digunakan untuk mengambil keputusan dengan metode SVM. Secara matematis Berikut merupakan fungsi kernel yang digunakan dalam metode SVM:

1. Kernel *Lini*er

   Kernel *linier* merupakan fungsi kernel sederhana. Kernel *linier* digunakan ketika data yang dianalisis sudah terpisah secara linear. Kernel *linier* pantas digunakan ketika terdapat banyak fitur. Berikut merupakan persamaan dari linear kernel [15].

   $$\text{Kernel } Linier = x^T x \quad (2)$$

2. Kernel *Polynomial*

   Kernel *polynomial* merupakan fungsi kernel yang pantas digunakan ketika data tidak terpisah secara *linear*. *Polynomial* kernel sangat pantas untuk permasalahan dimana semua *training* data dinormalisasi [15].

   $$\text{Kernel } polynomial = (x^T x + 1)^p \quad (3)$$

3. Radial Basis Function (RBF)

Kernel RBF adalah fungsi kernel yang digunakan ketika data tidak dapat dipisahkan secara linier. RBF membutuhkan dua parameter yaitu Gamma dan Cost (C) untuk menghasilkan klasifikasi yang optimal dan menghindari kesalahan klasifikasi. *Gamma* digunakan untuk mengukur seberapa jauh pengaruh dari satu sempel data latih. Nilai *gamma* rendah berarti "jauh" dan nilai tinggi berarti "dekat" [15].

$$\text{Kernel RBF} = \exp(\gamma||x - x||^2) \quad (4),$$

4. *Tangent hyperbolic* (sigmoid)

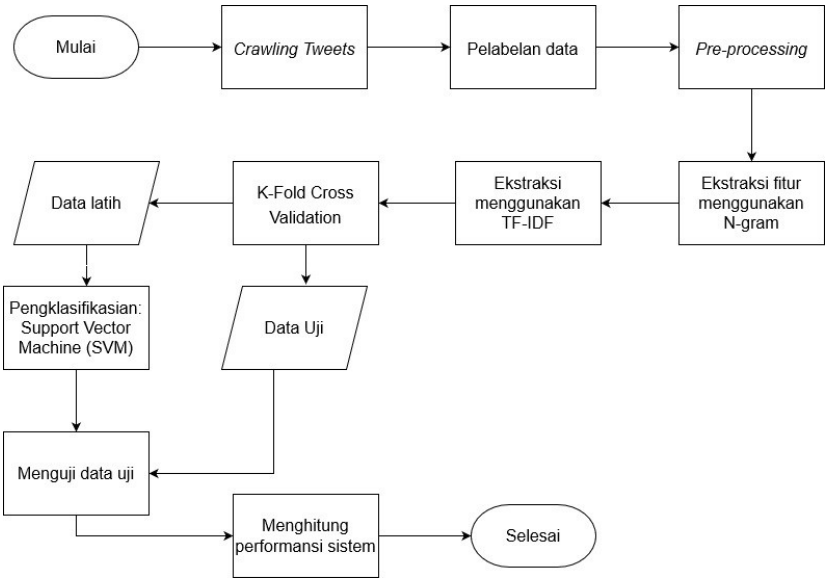
   Persamaan dari sigmoid kernel sebagai berikut.

$$\text{Kernel Sigmoid} = \tanh(\beta_0 x^T x_i + \beta_1) \quad (5)$$

3. Sistem yang Dibangun

Gambaran Umum Sistem

Sistem yang dibangun pada penelitian ini merupakan sistem yang dapat menganalisis sentimen terhadap *cyberbullying* pada media sosial Twitter. Berikut Gambar 1 merupakan *flowchart* atau alur yang berjalan pada sistem.

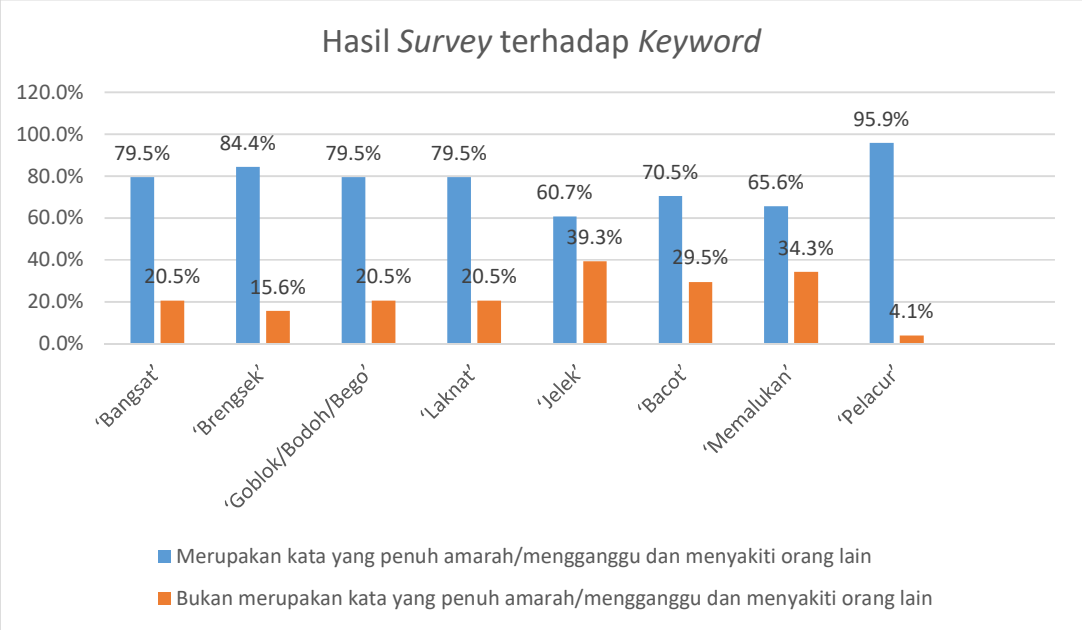


Gambar 1 *Flowchart* Perancangan Sistem

Pengumpulan Data

Pengumpulan data (*crawling*) dilakukan dengan mengambil data dari Twitter sejumlah 10.000 data, 80% sebagai data latih dan 20 % sebagai data uji. Proses *crawling* data dilakukan dengan menggunakan *Application Program Interface* (API) yang telah disediakan oleh Twitter. *Keyword* yang digunakan yaitu 'jelek', 'bodoh', 'goblok', 'brengsek', 'bangsat', 'memalukan', 'laknat', 'bacot' dan 'pelacur'. Berdasarkan [1] contoh tindakan *cyberbullying* adalah *flaming* (terbakar) yaitu mengirimkan pesan teks yang isinya merupakan kata-kata yang penuh amarah dan frontal, dan juga *harassment* (gangguan) yaitu pesan-pesan yang berisi gangguan pada email, sms, maupun pesan teks di jejaring sosial. Maka dari itu dilakukan *survey* kepada 122 orang mengenai *keyword* yang digunakan untuk mengetahui kata tersebut merupakan kata yang penuh amarah dan dapat menyakiti perasaan orang lain atau tidak. Proses pemilihan *keyword* untuk *crawling* data dilakukan dengan membagikan pertanyaan *survey* dalam bentuk Google Form kepada responden, dan dari hasil *survey* ditentukan apabila lebih dari 50%

responden menyatakan bahwa *keyword* tersebut merupakan kata yang penuh amarah atau mengganggu dan menyakiti orang lain. Gambar 2 merupakan hasil dari *survey*.



Gambar 2 Hasil *Survey* terhadap Pemilihan *Keyword* yang digunakan untuk *Crawling*

Dalam pelaksanaan *survey* berikut merupakan pertanyaan yang diajukan:

1. Apakah menurut kalian kata 'Bangsat' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
2. Apakah menurut kalian kata 'Brengsek' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
3. Apakah menurut kalian kata 'Goblok/Bodoh/Bego' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
4. Apakah menurut kalian kata 'Laknat' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
5. Apakah menurut kalian kata 'Jelek' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
6. Apakah menurut kalian kata 'Bacot' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
7. Apakah menurut kalian kata 'Memalukan' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?
8. Apakah menurut kalian kata 'Pelacur' merupakan kata penuh amarah atau mengganggu dan menyakiti perasaan orang lain?

Pelabelan Data

Dalam mempersiapkan datas untuk sistem klasifikasi, dengan itu proses agar dataset yang telah dikumpulkan memiliki label kelas yang benar [12]. Pada penelitian ini tahap pelabelan data dilakukan dengan membagikan data yang sudah dikumpulkan sebelumnya, setelah data terkumpul data dibagikan kepada lima orang partisipan untuk melakukan pelabelan data berupa angka 0 (*non-cyberbullying*) atau 1(*cyberbullying*). Setelah melakukan pelabelan data, dilakukan tahap *polling* terhadap label, label yang digunakan pada setiap data adalah hasil label terbanyak dari kelima partisipan.