Bab 1 Pendahuluan

1.1 Latar Belakang

Perkembangan Twitter sebagai *platform* Media Sosial (MedSos) telah mengalami pertumbuhan pesat dalam keterlibatan pengguna dan pembuatan konten di berbagai demografi. Sebagai *platform* MedSos, Twitter menciptakan lingkungan dimana pengguna dapat mengungkapkan pendapat secara bebas, terlibat dalam diskusi, dan berbagai informasi tentang beragam topik. Meskipun keterbukaan ini dapat dianggap sebagai aspek positif, hal ini juga telah menimbulkan masalah yang mengkhawatirkan. *Hate Speech* yang ditandai oleh ekspresi yang merendahkan, mendiskriminasi, atau menyerang individu atau kelompok berdasarkan atribut seperti suku, agama, ras, antargolongan (SARA), atau faktor lain adalah masalah yang sering terjadi di Twitter. Kejadian *Hate Speech* di *platform* Twitter sering melibatkan penggunaan bahasa merendahkan dan penyebaran sentimen berbahaya.

Implikasi dari *Hate Speech* sangat signifikan karena dapat menyebabkan trauma, penyebaran kebencian, dan dalam beberapa kasus bahkan memprovokasi tindakan diksriminatif atau kekerasan. Selain dampaknya pada individu, *Hate Speech* juga merupakan tantangan serius bagi reputasi *platform* yang berpotensi mencemarkan citranya dan merusak pengalaman positif pengguna. Masalah yang umum terkait *Hate Speech* di Twitter mencakup penyebaran pesan berisi kebencian, kata-kata kasar berdasarkan SARA, dan pelecehan seringkali menargetkan beragam pengguna *platform* tersebut. Tantangan ini membutuhkan pendekatan yang kuat untuk mendeteksi *tweet* yang bersifat *Hate Speech* pada *platform* Twitter. Teknik yang sering digunakan dalam mendeteksi atau mengklasifikasi kalimat yang mengandung *Hate Speech* yaitu Machine Learning (ML).

Oleh karena itu, penelitian ini menggunakan teknik *Machine Learning* untuk mengklasifikasikan postingan atau *Tweet* di twitter menjadi dua kelas yaitu *Hate Speech* dan *Non-Hate Speech*. Algoritma ML populer untuk mengklasifikasikan *Hate Speech* meliputi *Naïve Bayes Classifier* (NBC), *Support Vector Machine* (SVM), dan berbagai model jaringan saraf tiruan (*Neural Networks*) seperti *Bidirectional Encoder Representations from Transformers* (BERT).

Algoritma NBC menggunakan prinsip teorema Bayes untuk menghitung probabilitas bahwa suatu teks tertentu termasuk dalam kategori klasifikasi yang sesuai dengan sentimennya. SVM adalah algoritma pembelajaran mesin yang biasa diterapkan dalam analisis sentimen. SVM digunakan untuk membuat model prediktif menggunakan teknik pembelajaran yang diawasi dan mengklasifikasikan dokumen teks, seperti ulasan produk atau pesan medsos, sebagai positif, negatif, atau netral. SVM bekerja dengan menemukan batas keputusan optimal yang memisahkan data positif dan negatif di ruang fitur.

Algoritma ini digunakan dalam konteks analisis sentimen atau deteksi teks berbahaya dan bertujuan untuk mengidentifikasi ekspresi atau kata-kata yang merendahkan atau menyerang individu atau kelompok berdasarkan atribut seperti SARA. Penggunaan algoritma ini membantu dalam mengidentifikasi *Hate Speech* dengan *accuracy* yang lebih baik, membantu *platform* dan layanan *online* dalam mitigasi masalah *Hate Speech*. Berdasarkan hal tersebut, penelitian ini menggunakan algoritma NBC dan SVM untuk mendeteksi *Hate Speech* berdasarkan twitt atau postingan pada *platform* Twitter.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dipaparkan maka dapat diidentifikasi masalah sebagai berikut:

1. Twitter adalah *platform* MedSos yang menganut kebebasan berpendapat sehingga sangat berpotensi terjadi penyebaran *Hate Speech*.

2. Masalahnya adalah memberikan sentimen Hate Speech atau Non-Hate Speech pada Tweet di Twitter sehingga diperlukan analisis sentimen untuk memahami perbedaan antara Tweet yang masuk dalam kategori *Hate Speech* atau *Non-Hate Speech*.
3. Masalahnya adalah mengidentifikasi Tweet di Twitter yang termasuk dalam kategori Hate Speech atau Non-Hate Speech sehingga diperlukan algoritma ML seperti NBC, SVM, *Logistic Regression* (LR), *Random Forest* (RF), dan *K-Nearest Neighbors* (KNN).

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Data yang digunakam dalam penelitian ini yaitu *Tweet* pada *platform* Twitter yang diakuisisi menggunakan *Twitterscraper* dan Twitter API dalam rentang waktu antara 21 November hingga 13 Maret 2023.
2. Penelitian ini menganalisis postingan yang mengandung *Hate Speech* dan *Non-Hate Speech*.
3. Penelitian ini menggunakan *Machine Learning* yaitu Algoritma NBC dan SVM.

1.4 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah:

1. Bagaimana mengklasifikasi *Hate Speech* menggunakan algoritma NBC?
2. Bagaimana mengklasifikasi *Hate Speech* menggunakan algoritma SVM?
3. Bagaimana hasil perbandingan performa antara algoritma NBC dan SVM berdasarkan teknik *Stratified K-Fold*?

1.5 Tujuan Penelitian

Penelitian ini memiliki tujuan sebagai berikut:

1. Mengklasifikasi *Hate Speech* menggunakan algoritma NBC.
2. Mengklasifikasi *Hate Speech* menggunakan algoritma SVM.
3. Menganalisis perbandingan performa antara algoritma NBC dan SVM berdasarkan teknik *Stratified K-Fold*.

1.6 Manfaat Penelitian

Penelitian yang dilakukan diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat bagi ilmuwan, yaitu memberikan referensi kepada peneliti lain dengan topik penelitian serupa ataupun pengembangan dari penelitian ini.
2. Manfaat secara teoritis, yaitu memberikan masukan pada bidang ilmu pengetahuan dan teknologi terhadap materi mengenai analisis sentimen dengan algoritma NBC dan SVM.