

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Dengan berkembangnya Internet, semakin banyak orang yang menggunakan media sosial. Dengan banyaknya orang yang menggunakan media sosial, semakin banyak orang yang menuliskan opini ataupun berbagi cerita kehidupan mereka disana. Salah satu media sosial yang digunakan tersebut adalah Blog. Sejak kemunculannya di 1996, blog sudah berkembang sangat pesat [1]. Pengertian dari blog itu sendiri adalah suatu teks yang ditulis oleh seorang individu, sehingga satu blog dengan blog lainnya memiliki gaya penulisan yang berbeda satu sama lain. Karena sifat blog yang dapat digunakan oleh siapapun, tidak berbayar dan tanpa ada aturan penulisan yang diterapkan tulisan di dalam sebuah blog biasanya bersifat informal yang menyebabkan banyaknya tulisan di dalam blog memiliki ejaan yang salah, kata-kata *slang* dan penggunaan *emoticon* seperti “:D”, “:)”, dan lain-lain didalamnya [5]. Karenanya, beberapa tahun belakangan ini terdapat sejumlah besar riset yang berpusat pada tulisan-tulisan yang ada di blog dalam komunitas *Natural Language Processing* [3].

Salah satu riset yang adalah memprediksi jenis kelamin penulis blog. Menentukan jenis kelamin dari penulis blog adalah hal yang cukup sulit. Dalam prosesnya, dianalisa hubungan antara jenis kelamin penulis dengan berbagai kata yang terdapat pada tulisan penulis untuk melihat kata-kata yang dipakai oleh sang penulis [5]. Namun karena bentuk tulisan dari penulis yang informal, tahap *preprocessing* adalah menjadi tahap yang paling sulit [5]. Walaupun begitu, hasil yang didapat dari memprediksi jenis kelamin nantinya sangat berguna ketika kita ingin mengetahui jenis kelamin seseorang hanya dari tulisan yang mereka *post* pada Internet. Dengan klasifikasi jenis kelamin secara otomatis ini, kita yang bila hanya membaca suatu tulisan tidak dapat mengetahui penulis laki-laki atau perempuan, dengan sistem ini diharapkan dapat memprediksi secara langsung penulis blog tersebut laki-laki atau perempuan tanpa mereka mencantumkan jenis kelamin mereka di *profile* blog.

Berdasarkan hal diatas tersebut, *document classification* (dalam hal ini teks yang ada di blog) untuk menentukan jenis kelamin penulisnya dilakukan. Dalam penelitian ini, klasifikasi jenis kelamin dilakukan dengan menggunakan metode *Naïve Bayes*. Dengan terwujudnya sistem klasifikasi ini dapat membantu orang-orang untuk mengetahui jenis kelamin penulis blog yang mereka baca hanya dari tulisan yang ada.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan diatas, maka dapat dirumuskan beberapa masalah yang akan dicari solusinya, yaitu:

1. Bagaimana penerapan klasifikasi jenis kelamin penulis blog menggunakan metode *Chi-Square* dan algoritma *Naïve Bayes*?
2. Berapa peforma yang dihasilkan dari penelitian klasifikasi jenis kelamin penulis blog?

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Sumber data yang dipakai berasal dari situs-situs *website* pribadi. Data tersebut dibagi menjadi 2 *class*, yaitu *class Male* dan *class Female*.
2. Bentuk data teks yang informal dan menggunakan bahasa Inggris.

1.3 Tujuan

Tujuan dari penelitian Tugas Akhir ini adalah sebagai berikut:

1. Mengetahui performa dari penerapan klasifikasi dokumen menggunakan *Chi-Square* dan *Naïve Bayes*.
2. Mengimplementasikan kombinasi *Chi-Square* untuk *feature selection* dan *Naïve Bayes* untuk klasifikasi jenis kelamin penulis blog.

1.4 Metodologi Penyelesaian Masalah

Metodologi pada penelitian ini adalah sebagai berikut:

1. Studi Literatur

Penelitian ini dimulai dengan proses pengumpulan referensi mengenai klasifikasi dokumen terhadap penentuan jenis kelamin penulis blog.

2. Pengumpulan *Dataset*

Pada tahap ini, penulis melakukan pengumpulan *dataset* dari situs pribadi ataupun komunitas penulis blog. Data yang dikumpulkan sebanyak 480 *record*, yang terdiri dari dua *class* yaitu *male* dan *female*.

3. Analisis & Perancangan Sistem

Pada tahap ini, penulis melakukan klasifikasi teks pada blog. Sistem dibangun menggunakan *Chi-Square* sebagai *feature selection* dan *Naïve Bayes* untuk proses klasifikasi.

4. Implementasi Sistem

Pada tahap ini, penulis mengimplementasikan sistem yang dibangun dengan menggunakan *Chi - Square* dan *Naïve Bayes*.

5. Pengujian, analisis, validasi dan evaluasi

Hasil pengujian kemudian dianalisis serta dievaluasi untuk menyimpulkan hasil penelitian. Pada tahap ini, validasi yang digunakan adalah *k-fold cross validation*, dan evaluasi yang digunakan adalah *confusion matrix*.

6. Penyusunan laporan tugas akhir

Tahap ini merupakan tahap dimana, penulis melakukan pembuatan laporan akhir sebagai bentuk dari dokumentasi terkait penelitian ini.

1.5 Sistematika Penulisan

Sistematika penulisan dalam Buku Tugas Akhir ini terdiri dari lima bab, dengan rincian sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menjelaskan tentang latar belakang masalah, perumusan masalah, batasan masalah, tujuan, metodologi dan sistematika penulisan.

BAB 2 KAJIAN PUSTAKA

Bab ini menjelaskan tentang dasar teori yang mendukung dalam penelitian ini. Teori tersebut diambil dari beberapa literatur terkait.

BAB 3 PERANCANGAN SISTEM

Bab ini menjelaskan tentang gambaran umum sistem, tahapan tiap proses penelitian, dan spesifikasi lingkup operasional yang digunakan pada penelitian.

BAB 4 PENGUJIAN DAN ANALISIS

Bab ini berisi tentang hasil pengujian dan analisis dari hasil penelitian yang telah dilakukan.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan dan saran yang berkaitan dengan hasil penelitian.