BAB 1
PENDAHULUAN

1.1. Latar Belakang

Selama dekade terakhir, *blog* menjadi sangat populer di internet. Menurut WordPress, salah satu penyedia layanan *blog publishing* menyatakan bahwa penggunanya rata-rata membuat 69,8 juta postingan baru dan 42 juta komentar baru setiap bulannya [1]. Sayangnya, dengan jumlah trafik yang besar terdapat celah pengelolaan yang kurang baik sehingga *blog* dapat menjadi sasaran untuk *spammers*. Sekarang ini, sebenarnya pemilik *blog* sudah menggunakan beberapa teknik untuk mengurangi komentar *spam*. Beberapa pemilik *blog* memilih melakukan monitoring dan mengelola komentar secara *manual*. Teknik lain yang digunakan pemilik *blog* untuk membedakan komentar yang dilakukan secara otomatis oleh *bot* dengan komentar asli yang dilakukan oleh *user* adalah dengan menggunakan CAPTCHA [2]. CAPTCHA biasanya berbentuk gambar yang berisi huruf dan angka yang mana sulit untuk dikenali secara otomatis oleh *bot*. Akan tetapi, riset telah membuktikan bahwa metode ini sangat mudah untuk dirusak [3].

Di tahun 2005 Mishne et al. [4] menggunakan pendekatan pemodelan bahasa untuk mendeteksi komentar spam dengan metode *Kullback-Leibler divergence* mendapatkan tingkat akurasi 83%. Di tahun 2011 Bhattarai et al. [5] menggunakan analisis konten untuk mengidentifikasi *spam* dengan fitur *words duplications, stopwords ratio etc.*, dengan hasil terbaik menggunakan metode *Support Vector Machine* (SVM). Dari pendekatan tersebut didapatkan tingkat akurasi tertinggi 86%. Di tahun 2012 Ashwin et al. [6] menggunakan analisis komentar dan hubungan antara postingan *blog* dengan komentar dengan menggunakan beberapa metode klasifikasi didapatkan tingkat akurasi tertinggi menggunakan metode *decision tree* dengan akurasi 92%.

Pada penelitian lain di tahun 2013 Pausta dkk. [7] membandingkan SVM dengan *Rocchio* untuk melakukan penelusuran katalog perpustakaan hasilnya *Rocchio* memiliki waktu pemrosesan lebih kecil 57,2% dan tingkat presisi 37,8% lebih besar dari SVM. *Rocchio Classification* yang diambil dari konsep *Rocchio*

Relevance Feedback memiliki konsep desain hanya untuk mengklasifikasi dua kelas yaitu relevan dan tidak relevan [8]. Berdasarkan konsep tersebut pada penelitian ini akan sangat cocok dikarenakan pada penelitian ini akan mengklasifikasikan komentar ke dalam kategori *spam* atau bukan *spam*. Oleh karena itu, pada penelitian ini akan dilakukan implementasi *Rocchio Classification* dalam mengidentifikasi komentar spam dengan harapan mendapatkan tingkat presisi yang lebih baik.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka identifikasi masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan metode *Rocchio Classification* untuk mengkategorikan komentar *spam* berdasarkan analisis komentar dan relevansi komentar terhadap konten.
2. Bagaimana tingkat akurasi yang dihasilkan dari proses *Rocchio Classification* dalam mengkategorikan komentar *spam* berdasarkan analisis komentar dan relevansi komentar terhadap konten.

1.3. Maksud dan Tujuan

Maksud dari penelitian ini adalah mengimplementasikan algoritma *Rocchio Classification* untuk mengkategorikan komentar *spam*. Sedangkan tujuan yang ingin dicapai dari penelitian ini adalah:

1. Mengkategorikan komentar *spam* dengan metode *Rocchio Classification*.
2. Menguji tingkat akurasi *Rocchio Classification* dalam mengkategorikan komentar *spam*.

1.4. Batasan Masalah

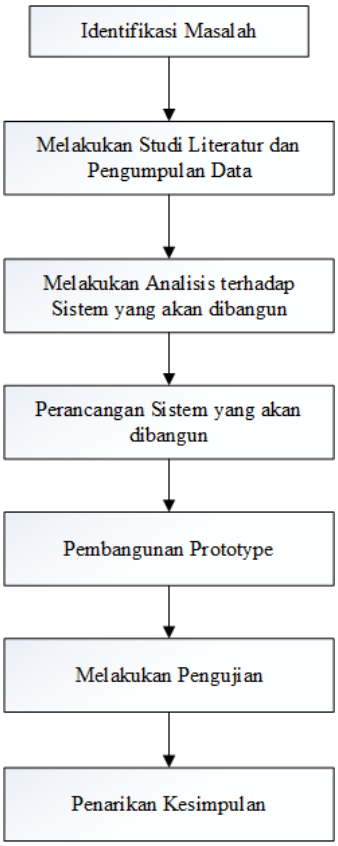
Agar pembahasan menjadi lebih jelas dan terarah, maka permasalahan dibatasi terhadap masalah-masalah berikut:

1. Data masukan menggunakan format json.
2. Sumber data set yang digunakan dari penelitian Mishne et al [4].
3. Bahasa yang digunakan adalah bahasa Inggris.

4. Tidak ada penanganan untuk kata tidak baku dan *typo*.

1.5. Metodologi Penelitian

Pada penelitian ini metode penelitian yang digunakan yaitu metode penelitian eksperimental. Metode eksperimental adalah metode yang mempunyai tujuan untuk menjelaskan hubungan sebab-akibat antara satu variabel dengan lainnya. Alur penelitian yang akan dilakukan pada penelitian ini dapat dilihat pada gambar 1.1 sebagai berikut:



Gambar 1.1 Alur Penelitian

1.5.1. Metode Pengumpulan Data

Metode pengumpulan data dengan menggunakan data pada penelitian yang dilakukan oleh Mishne et al [4] disisi lain yang digunakan dalam penelitian ini adalah sebagai berikut:

Studi pustaka dilakukan dengan cara mempelajari, meneliti dan menelaah berbagai literatur dari perpustakaan yang bersumber dari buku-buku, jurnal ilmiah, situs-situs internet, dan bacaan-bacaan yang ada kaitannya dengan topik penelitian.

1.5.2. Metode Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak yang digunakan pada penelitian ini adalah model *Prototype*. Berikut tahapan-tahapan yang dilakukan dalam penelitian ini:

1. Analisis

   Analisis masalah dilakukan untuk memahami masalah yang timbul dan mencari solusi untuk memecahkan masalah dalam menghasilkan klasifikasi komentar *spam*.

2. Kebutuhan Data

   Pada tahap ini peneliti akan mengumpulkan data komentar untuk data masukan sistem.

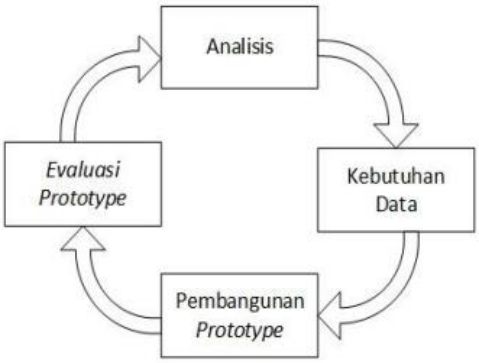
3. Pembangunan *Prototype*

   Pada tahap ini akan diimplementasikan dari proses analisis dan kebutuhan sistem yang sudah didapatkan dan peneliti mencoba mengimplementasikan metode *Rocchio Classification* ke dalam logika-logika program.

4. Evaluasi *Prototype*

   Program akan diuji, dimana uji coba dilakukan untuk mengetahui kekurangan pada program. Jika masih ada kekurangan, maka *prototype* direvisi dengan tahapan-tahapan yang sebelumnya telah dilakukan.

Tahapan *prototype* yang dilakukan pada penelitian ini akan dijelaskan pada gambar 1.2.



Gambar 1.2 Model *Prototype*

1.6. Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum mengenai penelitian yang dijalankan. Sistematika penulisan penelitian sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang latar belakang permasalahan dari komentar spam, bagaimana menyelesaikan permasalahan komentar spam, penelitian yang telah ada beserta masalahnya dan solusi yang ditawarkan, bagaimana merumuskan ini permasalahan yang dihadapi, menentukan maksud dan tujuan penelitian, batasan masalah, metodologi penelitian serta sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA

Bab ini membahas berbagai konsep dasar dan teori-teori yang berkaitan dengan topik penelitian yang dilakukan dan hal-hal yang berguna dalam proses analisis permasalahan serta tinjauan terhadap penelitian, seperti penjelasan mengenai konsep algoritma *Rocchio classification*, Perl, jenis-jenis spam di blog, klasifikasi, konfusi matriks dll.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini memamparkan penjelasan mengenai analisis data masukan, *preprocessing*, analisis pelatihan berupa fitur-fitur yang digunakan dan analisis pengujian.

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas mengenai pengujian terhadap *data set* untuk mendapatkan tingkat akurasi metode *Rocchio Classification*.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran yang diperoleh dari hasil penulisan penelitian berupa tingkat keakurasian metode *Rocchio Classification* dalam mengklasifikasi komentar spam di *blog* dan berupa saran-saran yang perlu dilakukan untuk penelitian lebih lanjut.