# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Kebutuhan akan informasi mendorong perkembangan penelitian dan teknologi yang dapat menjawab kebutuhan tersebut. Informasi yang dibutuhkan mengalami perkembangan dari informasi yang bersifat umum ke informasi yang bersifat lebih khusus dan spesifik. Perolehan informasi yang tepat dan cepat akan membantu dalam kemajuan dari suatu organisasi untuk dapat melakukan perubahan yang menjawab permasalahan yang dihadapi. Penelitian dalam analisis sentimen didorong oleh suatu pemikiran bahwa informasi berupa sentimen dari suatu data merupakan hal yang penting dan dibutuhkan. Sentimen berhubungan dengan penilaian terhadap suatu konteks atau wacana. Sentimen positif menyatakan pemberian nilai yang baik pada konteks dalam teks dan sentimen negatif menyatakan kebalikannya.

Analisa sentimen merupakan bagian dari *opinion mining* [1]. Bidang ini melakukan studi mengenai opini orang-orang, sentimen, evaluasi, tingkah laku dan emosi terhadap suatu entitas seperti produk, layanan, organisasi, individu, permasalahan, topik, acara dan atribut-atributnya[2].

Prose pengklasifikasian sentimen dari *tweet* pun mempunyai beberapa tantangan. Pertama, bahasa yang dipakai di dalam *tweet* sering tak mempunyai struktur formal dalam kalimat mereka, seperti pemakaian singkatan, perubahan dari huruf ke angka, kurangnya tanda baca, dan lain-lain. Kedua, kalimat di *tweet* mempunyai domain (sosial,politik, ekonomi, teknologi, dll) yang independen sehingga orang dapat bicara tentang apa saja dengan domain yang berbeda dan membuatnya sulit untuk mengklasifikasikan sentimen.

Struktur kata yang terdapat di dalam *tweet* sendiri merupakan aspek yang penting dalam menentukan sentimen. Kalimat yang terdapat di dalam *tweet*

menimbulkan banyak perspektif opini yang berbeda sehingga dapat menimbulkan ambiguitas terhadap pembacanya. Seperti kata "bisa", "bisa" dapat menjadi kata kerja apabila menjelaskan kata sifat seperti "bisa marah". Dapat juga menjadi kata benda ketika disandangkan dengan kata benda juga seperti "bisa ular".

*Part-Of-Speech* (POS) merupakan sintaks atau *tag* sebuah kata[20]. Dengan adanya pengidentifikasian sintaks sebuah kata terhadap kata-kata yang terdapat dalam tweet, akan lebih mudah dalam mengenali peran kata-kata tersebut. Dalam analisa sentimen sendiri, terdapat 4 tag yang berperan penting dalam menentukan sebuah sentimen yaitu kata benda(NN), kata kerja(VB), kata keterangan(RB), dan kata sifat(JJ)[24].

Untuk mengklasifikasi suatu sentimen yang tedapat pada tweet, penulis menggunakan pendekatan Support Vector Machine (SVM). Dan sebelum masuk proses klasifikasi yang akan dilakukan oleh SVM, terdapat tahap *feature selection*[19]. yang akan menggunakan hasil dari *Part-of-Speech Filtering* yang akan diproses oleh Stanford POS *Tagger* dengan model untuk bahasa indonesia yang telah dilatih dari korpus yang dipakai pada penelitian[22].

Penelitan yang pernah dilakukan sebelumnya terkait dengan analisa sentimen twitter berbahasa Indonesia dengan pendekatan SVM [23] dengan tingkat hasil pengujian yaitu 74,80% terhadap topik "penyedia layanan telepon seluler". Namun, dengan melakukan anotasi *Part of Speech Tag* pada tahap *feature selection*, rata-rata akurasi yang dihasilkan meningkat menjadi 95.89%[20] terhadap topik "operator seluler".

## 1.2 Perumusan Masalah

Dengan latar belakang yang telah dijelaskan, permasalahan dalam penelitian ini adalah bagaimana memanfaatkan Part Of Speech Filtering pada tahap *Feature Selection* dengan metode *Support Vector Machine*.

## 1.3 Batasan Masalah

Berdasarkan rumusan masalah yang telah dijelaskan diatas, terdapat beberapa pembatasan masalah pada beberapa pokok bahasan, yaitu:

1. Analisa sentimen dengan metode SVM menggunakan tools yang sudah tersedia dan mengklasifikasikan ke dalam tiga kelas sentimen, positif, negatif, dan netral. Pekerjaan yang dilakukan mencakup mengunduh *tweet*, *tweet preprocessing*, klasifikasi *tweet*, pengujian data, serta pemanfaatan *tools* atau *library* yang tersedia untuk analisis sentimen maupun *tweet preprocessing*.
2. Fokus dari penelitian ini yaitu menerapkan metode *Filtering* pada *Feature Selection* sebelum masuk ke tahap klasifikasi yang dilakukan oleh SVM.
3. Analisis sentimen dalam penelitian ini menggunakan data *tweet* dari Twitter yang di unduh selama satu hari dalam masa putaran kedua pemilihan gubernur DKI Jakarta.
4. Penelitian ini tidak berfokus pada tahap pembangunan model POS Tagger
5. Fitur yang di filter pada tahap POS *Filtering* hanya menyaring fitur yang memiliki tagset kata benda, kata kerja, kata sifat, dan kata keterangan.

## 1.4 Tujuan Penelitian

. Tujuan utama dari penelitian ini adalah untuk mengetahui kinerja dari penerapan metode Part Of Speech Filtering dalam tahap feature selection pada metode Support Vector Machine dalam melakukan analisis sentimen dokumen berbahasa Indonesia. Dengan mencoba mengaplikasikan kembali percobaan yang dilakukan dalam [19] Selain itu, penelitian ini juga bertujuan untuk mengetahui akurasi dari klasifikasi yang dihasilkan oleh Support Vector Machine dengan Part Of Speech Filtering dan tanpa Part of Speech Filtering

## 1.5 Metodologi Penelitian

Metodologi yang digunakan dalam pengerjaan tugas akhir ini adalah metode eksperimental dan dilaksanakan dalam tahapan-tahapan berikut:

1. Studi literatur.

   Melakukan berbagai macam pengumpulan bahan referensi, seperti jurnal penelitian, tesis, buku-buku teori dan sumber lain termasuk informasi yang diperoleh dari internet. Mempelajari literature dan teori pendukung penelitian mengenai klasifikasi, khususnya *Support Vector Machine* dan *Feature Selection*.

2. Analisis Sistem

   Melakukan analisa terhadap syarat yang dibutuhkan dalam pengembangan sistem.

3. Perancangan.

   Pada tahap ini dilakukan perancangan metode kombinasi antara feature selection menggunakan POS *Filtering* dan SVM yang akan di implementasikan untuk menganalisa sentimen. Data-data yang diambil dalam penelitian ini diambil dari Twitter API yang sudah disediakan, selanjutnya data tersebut disimpan dalam file dokumen. Perancangan sistem meliputi training dan testing. Training sendiri terdapat beberapa tahap meliputi melabeli tweet secara manual, *preprocessing*, dan proses training sendiri. Pertama-tama, secara manual, data training akan dikelompokan menjadi dalam 3 kelompok, positif, negatif, dan netral. Lalu dilanjutkan ke tahap *preprocessing*. Tahap *preprocessing* akan melakukan beberapa hal, yaitu :

   a. *Clear invalid UTF 8*
   b. *Casefolding*
   c. *Remove* URL
   d. *Remove* @username

e. *Remove* #hashtag
f. *Remove* Punctuation
g. Part Of Speech Tagging

Setelah melakukan tahap *preprocessing* maka data tweet yang sudah diberi label serta sintaks masing-masing kata dihitung bobotnya. Kemudian, bobot tersebut akan disimpan dalam berkas. Selanjutnya akan dilakukan proses pelatihan dengan memanfaat *library* Scikit-Learn. Hasil dari pelatihan ini adalah SVM model yang disimpan dalam berkas.

Pada proses testing akan menggunakan metode SVM yang memanfaatkan *library* Scikit-Learn. Dokumen yang berisikan tweet akan dianalisa sentimennya melalui SVM model yang telah dibuat. Selanjutnya, dilakukan preprocessing terhadap data tweet yang didapat. Lalu, data tweet akan di indentifikasi bobotnya dengan melakukan pembobotan *Cosine Similarity* terhadap tweet tersebut. Langkah terakhir adalah ditentukan labelnya melalui klasifikasi yang dilakukan oleh model SVM.

4. Implementasi

   Pada tahap ini akan dilakukan pembuatan kode program sampai proses klasifikasi sentimen dokumen.

5. Pengujian.

   Pada tahap ini dilakukan perhitungan akurasi dari implementasi yang telah dilakukan. Metode yang digunakan yaitu K-Fold Cross Validation. Dalam penelitian ini, tahap pengujian menggunakan 10-Fold sebagai parameter untuke melakukan Cross Validation. Pengujian yang dilakukan adalah perbandingan akurasi klasifikasi yang dihasilkan SVM. Tahap yang dibandingkan adalah dengan menggunakan POS Filtering dan tanpa melakukan POS Filtering. Adapun evaluasi terhadap jumlah fitur yang dihasilkan ketika melalui proses POS Filtering dengan jumlah fitur yang dihasilkan tanpa melalui proses POS Filtering.

## 1.6 Sistematika Penulisan

Agar dapat tercapai penulisan yang sistematis mengenai pokok permasalahan sebagai hasil penelitian, maka akan lebih baik apabila diberikan gambaran sistematika penulisan secara ringkas mengenai susunan skripsi ini maupun tentang apa yang dikandung dalam skripsi, sehingga akan mempermudah dalam pemahaman dan pembahasannya. Adapun sistematika penulisan yang digunakan dalam penelitian ini adalah sebagai berikut:

BAB I PENDAHULUAN

Dalam bab satu ini menguraikan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metodologi penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Landasan teori merupakan tinjauan pustaka. Memuat penelitian-penelitian terdahulu yang berkaitan dengan penelitian yang dilakukan penulis. Beberapa penelitian terkait *Opininon Mining* atau *Sentiment Analysis* dibahas secara singkat dan dibandingkan dengan penelitian penulis. Lalu, bab ini membahas teori-teori yang berkaitan dengan penelitian.

BAB III ANALISIS DAN RANCANGAN SISTEM

Bab ini membahas analisis kebutuhan data, analisis model beserta rancangan sistemnya.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Bab ini membahas implementasi opinion mining dengan metode SVM dan POS Filtering pada tahap feature selection. Bab ini juga membahas percobaan yang dilakukan pada proses pelatihan dan penentuan kelas beserta uraian mengenai hasil dan perbandingannya.

BAB VI KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan-kesimpulan dari hasil penelitian dan saran-saran yang berguna untuk penelitian selanjutnya.