

1. Pendahuluan

1.1 Latar Belakang Masalah

Surat kabar merupakan salah satu wadah aspirasi masyarakat dan media persebaran informasi sehingga surat kabar dianggap kaya akan informasi. Informasi dalam surat kabar dapat berasal dari opini dan dari fakta yang dijabarkan. Dalam kalimat opini banyak informasi yang dapat diekstraksi seperti topik opini, isi opini, dan lain-lain.

Namun proses pemanfaatan informasi pada surat kabar tidak mudah dilakukan, salah satu hambatannya adalah banyaknya anaphora pada kalimat di surat kabar[18]. Anaphora adalah keadaan dimana sebuah entitas menunjuk pada entitas lain yang mendahuluinya didalam satu data. Karena banyaknya kasus anaphora banyak penelitian dibuat untuk menangani kasus anaphora pada bahasa Inggris dan lainnya, yaitu penelitian mengenai *anaphora resolution*. *Anaphora resolution* adalah penelitian untuk menentukan entitas yang ditunjuk oleh entitas lain yang merupakan anaphora. Namun untuk bahasa Indonesia penelitian tentang anaphora belum pernah dilakukan sehingga pada tugas akhir ini akan dibuat sebuah sistem yang mampu mengenali kalimat anaphora khususnya *definite noun phrase anaphora* sebagai langkah awal untuk melakukan penelitian tentang *anaphora resolution*. Sistem yang dibangun akan menggunakan mesin *learning Support Vector Machine (SVM)* dengan kernel *Radial Basis Function (RBF)* dengan menggunakan *feature set* untuk membentuk *feature vector*-nya. *Feature set* adalah kumpulan aturan yang berupa ciri-ciri kalimat *definite noun phrase anaphora*.

SVM dipilih karena merupakan metode klasifikasi yang teliti dan menghasilkan akurasi yang cukup baik untuk pemrosesan anaphora pada surat kabar berbahasa Inggris yaitu 75% untuk keseluruhan *anaphora*[19]. Sedangkan kernel RBF digunakan untuk memetakan data input pada input space ke dimensi ruang yang lebih tinggi sehingga dapat dipisahkan secara linier. Kernel RBF juga memiliki performansi yang lebih baik dengan parameter yang lebih sederhana daripada kernel *polynomial* dan *sigmoid*[8].

1.2 Rumusan Masalah

Berdasarkan pada latar belakang yang telah dipaparkan diatas maka disusunlah beberapa rumusan masalah sebagai berikut:

1. Bagaimana menyusun *feature set* untuk identifikasi *definite noun phrase anaphora* pada kalimat berbahasa Indonesia.
2. Bagaimana menghasilkan sistem yang mampu mengenali kalimat anaphora dalam Bahasa Indonesia.
3. Bagaimana pengaruh parameter dan persebaran data pada kemampuan prediksi sistem berdasarkan perhitungan presisi.

1.3 Batasan Masalah

Adapun batasan dari masalah yang ditangan pada tugas akhir ini antara lain adalah:

1. Data yang ditangani pada tugas akhir ini adalah data berita yang berasal dari surat kabar Pikiran Rakyat, Republika, dan surat kabar Merdeka dengan topik bahasan Politik
2. Anafora yang ditangani adalah jenis *definite noun phrase anaphora*.
3. Dokumen yang dianalisis adalah dokumen dengan Bahasa Indonesia yang berupa kalimat berita.
4. Noun frase atau frase nominal yang ditangani dalam sistem ini adalah frase nominal baku atau tidak menangani frase nominal yang mengalami perluasan.
5. Mesin learning SVM menggunakan library yaitu LibSVM.
6. Menggunakan POS Tagging yang sudah ada.

1.4 Tujuan

Adapun tujuan yang ingin dicapai pada pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Menyusun *feature set* yang mengidentifikasi *definite noun phrase anaphora* pada kalimat opini berbahasa Indonesia.
2. Mengetahui kombinasi parameter yang paling baik untuk mendapat prediksi yang paling akurat berdasarkan perhitungan.
3. Mengetahui pengaruh persebaran data terhadap nilai akurasi.

1.5 Hipotesis

Feature-based RBF SVM akan digunakan sebagai mesin learning pada tugas akhir ini. Dimana *feature set* yang telah didefinisikan pada sistem mampu membentuk *feature vector* dari data *training* yang nantinya digunakan dalam proses *training* dan *testing*. *Feature vector* yang digunakan pada proses *training* nantinya akan menghasilkan pemodelan yang bersama dengan *feature vector* data *testing* merupakan input untuk proses *testing*. *Feature set* yang didefinisikan diharapkan dapat membentuk suatu *feature vector* yang baik dan menghasilkan pemodelan yang tepat dengan parameter yang ditentukan dengan cara *cross-validation* sehingga SVM dapat melakukan prediksi pelabelan yang akurat sehingga akurasi yang dihasilkan dapat optimal berdasarkan perhitungan presisi.

1.6 Metodologi Penyelesaian

Untuk menyelesaikan masalah yang telah didiskripsikan sebelumnya diperlukan beberapa tahapan antara lain:

1. Studi literatur
Mempelajari literatur yang berkaitan dengan frasa nominal. Mempelajari pola-polanya dan serta memperdalam pemahaman tentang *Support Vector Machine* dengan kernel RBF.
2. Pengumpulan data
Data yang akan digunakan dalam percobaan merupakan data yang berasal dari surat kabar Pikiran Rakyat *online*, surat kabar Republika *online* dan surat kabar Merdeka *online*. Data diambil secara manual.
3. Perancangan sistem dan analisis kebutuhan
Pada tahap ini akan dibuat rancangan sistem yang mampu melakukan identifikasi terhadap kalimat yang mengandung *definite noun phrase anaphora* dan yang tidak.
4. Implementasi dan Analisis
Pada tahap ini akan dilakukan implementasi sistem terhadap kasus serta melakukan kajian terhadap hasil percobaan dan menyusun kesimpulan.
5. Penyusunan laporan
Penyusunan laporan Tugas Akhir yang merupakan dokumentasi dari kegiatan percobaan, hasil percobaan, analisis yang dilakukan serta kesimpulan akhir yang didapat.