1. PENDAHULUAN

1.1 Latar Belakang

Mendapatkan representasi argumen semantik dari sebuah kalimat adalah hal yang penting dalam bidang keilmuan *text mining* maupun n*atural language processing,* contoh penerapannya adalah *information extraction* dan *question answering*. Proses untuk menentukan apa, siapa melakukan apa kepada siapa, kapan, dimana, mengapa dan bagaimana, dan lain sebagainya dari sebuah teks menjadi tantangan sendiri bagi para peneliti [**14**]**.** Proses tersebut memerlukan identifikasi kata dalam kalimat yang mewakili argumen semantik dan menetapkan pelabelan untuk tiap kata. Proses mengidentifikasi argumen dari suatu predikat dalam kalimat hingga menentukan *role* atau peran semantiknya inilah yang dikenal dengan istilah *semantic role labelling*. *Semantic role labelling* memiliki 2 tahapan besar dalam prosesnya, pertama adalah identifikasi argumen yang selanjutnya diikuti dengan klasifikasi argumen. Ketersediaan korpus semantik yang telah dijelaskan pada [**6**] seperti PropBank, juga mendorong penelitian tentang bagaimana memproduksi representasi semantik dari kalimat berbahasa inggris.

Dalam *semantic role labelling* pemilihan fitur menjadi hal yang berpengaruh pada kinerjanya dengan kata lain juga berpengaruh pada *recall* dan *precision* juga akurasi yang dihasilkan [**14**,**15**,**19**]. Penelitian yang sudah dilakukan membuktikan bahwa penambahan *additional feature* tertentu selain *baseline feature* pada *semantic role labelling* berpengaruh pada akurasi, *precision*, *recall*, dan *F-measure* yang dihasilkan pada tahapan klasifikasi argumen [**14**,**19**,**5**]. Sayangnya penelitian sebelumnya hanya menyajikan kombinasi fitur secara dangkal, tanpa mencoba kombinasi secara keseluruhan terhadap fitur-fitur tambahan lainnya seperti pada [**15**,**5**,**19**], analisis fitur yang dilakukan kebanyakan hanya dengan penambahan sejumlah 1 fitur tambahan terhadap *Baseline Feature*. Lalu bagaimana jika menggunakan kombinasi tertentu dari beberapa fitur, apakah nantinya akan menaikkan atau menurunkan performa klasifikasi. Ada beberapa fitur yang menghasilkan kinerja yang relatif tinggi berdasarkan penelitian [**14**]. Beberapa dari fitur tersebut adalah *Noun head of PP* dan *First word in constituent*. Sedangkan berdasarkan penelitian yang dilakukan oleh [**19**] akurasi dan *F-measure* tertinggi dihasilkan oleh fitur *Syntactic Frame,* selain itu ada juga fitur yang memperhatikan letak urutan argumen dan unsur seperti *Argument Order* dan *Constituent Order*. Mengkombinasikan kelima fitur tersebut dengan *Baseline Feature* merupakan hal yang menarik, melihat kelima fitur tersebut memiliki karakteristik yang berbeda satu sama lain.

Pemilihan *classifier* dalam suatu pekerjaan klasifikasi juga merupakan suatu hal yang perlu diperhatikan. Salah satu *classifier* yang telah menunjukkan performa yang cukup baik pada *text classification task* adalah classifier *Support Vector Machine* (SVM) [**14**]. Metode ini juga termasuk ke dalam peringkat 10 teratas algoritma dalam data mining [**18**]. Ini menjadi sebuah alasan yang kuat untuk menggunakan *classifier* tersebut dalam pekerjaan klasifikasi argumen semantik. Jumlah data yang digunakan juga memiliki hubungan yang erat dengan performa dari suatu pekerjaan klasifikasi [**16**]. Lalu bagaimana dengan klasifikasi argumen semantik pada *task semantic role labelling*, apakah penambahan dan pengurangan

jumlah dataset juga akan berpengaruh secara signifikan pada hasil klasfikasi yang dilakukan.

Tugas akhir ini mencoba untuk menganalisis dan mengkombinasikan beberapa fitur yang ada yang terbukti telah menunjukkan hasil performa yang optimal pada penelitian yang sudah ada. Selain itu juga dilakukan berbagai macam skenario klasifikasi dengan ukuran dataset yang berbeda untuk mengetahui apakah jumlah data akan mempengaruhi hasil klasifikasi secara signifikan atau tidak. Fitur yang digunakan pada tugas akhir ini adalah *BaselineFeature* ditambah *Additional Feature* seperti *Noun Head of PP, First Word in Constituent Syntactic Frame, Argument Order, dan Constituent Order* yang nantinya digunakan dalam klasifikasi argumen semantik menggunakan *classifier* SVM. Kelima fitur itu dipilih karena pada penelitian sebelumnya menghasilkan angka akurasi dan *f-measure* yang relatif tinggi dibandingkan dengan fitur yang lainnya. Kemudian diukur dari kombinasi yang dihasilkan, kombinasi mana yang optimal berdasarkan kelima kombinasi tersebut.

1.2 Perumusan Masalah

Adapun rumusan permasalahan yang di lakukan dalam tugas akhir ini adalah sebagai berikut:

1. Bagaimana pengaruh ukuran data yang digunakan dalam pekerjaan klasifikasi argumen semantik?
2. Bagaimana pengaruh yang terjadi pada setiap kombinasi fitur tambahan yang digunakan untuk klasifikasi argumen semantik?
3. Bagaimana hasil dari klasifikasi berdasarkan implementasi menggunakan *BaselineFeature* ditambah *Additional Feature* seperti *Noun Head of PP, First Word in Constituent Syntactic Frame, Argument Order, dan Constituent Order*?

1.3 Tujuan

Adapun tujuan yang ingin dicapai dalam pengerjaan tugas akhir ini adalah sebagai berikut:

1. Menyajikan hasil ujicoba pengaruh ukuran data yang digunakan dalam pekerjaan klasifikasi argumen semantik.
2. Mengkombinasikan *BaselineFeature* dengan *Noun Head of PP, First Word in Constituent, Syntactic Frame, Constituent Order, dan Argument Order* dalam tahapan *fitur extraction*.
3. Menganalisis hasil kinerja klasifikasi yang dihasilkan berdasarkan fitur-fitur yang telah dikombinasikan.

1.4 Batasan Masalah

Tugas akhir ini memiliki beberapa batasan masalah untuk membatasi ruang penelitian yang ada. Batasan masalah dari tugas akhir ini adalah sebagai berikut:

1. Fitur yang akan digunakan dalam proses ekstraksi fitur adalah fitur-fitur yang sudah pernah dikembangkan dalam penelitian sebelum-sebelumnya.
2. Kombinasi yang dilakukan adalah kombinasi antara *Baseline Feature* dengan 5 *Additional Feature* sehingga menghasilkan $1+C_1^5+C_2^5+C_3^5+C_4^5+C_5^5 = 32$ kombinasi fitur.
3. *Dataset* yang digunakan dalam tugas akhir ini adalah *dataset* berupa korpus yang berasal dari PropBank.
4. Penelitian ini hanya melakukan klasifikasi argumen semantik berdasarkan kombinasi fitur yang diterapkan, sehingga diasumsikan proses identifikasi argumen telah dilakukan.
5. Klasifikasi dilakukan menggunakan Weka dengan *classifier libsvm*.

1.5 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah dalam tugas akhir ini adalah sebagai berikut:

1. Studi Literatur
   Tahap ini merupakan tahapan dilakukannya pengumpulan teori dan data yang berkaitan dengan metode yang digunakan dalam pengerjaan tugas akhir ini sehingga mendapatkan gambaran yang jelas tentang implementasi nantinya. Tahap studi literatur dilakukan dari awal hingga akhir pengerjaan Tugas Akhir.
2. Pengumpulan Data yang Dibutuhkan
   Pada tahap ini dilakukan pengumpulan *dataset* berupa data teks, yang nantinya akan diujikan dalam implementasi klasifikasi argumen semantik. *Dataset* diperoleh dari PropBank, sebuah korpus semantik yang telah tersedia untuk bahasa Inggris.
3. Perancangan dan Implementasi Sistem
   Merancang dan mengimplementasikan *tools* yang digunakan untuk melakukan *semantic role labelling* dan klasifikasi argumen semantik berdasarkan metode *Multiclass Support Vector Machine (SVM)* kernel linear.
4. Pengujian dan Analisis
   Dalam tahap ini dilakukan pengujian dan analisis hasil dari kombinasi fitur *semantic role labelling* yang digunakan, dalam hal ini yang dianalisis adalah *precision* dan *recall* yang dihasilkan dari metode yang telah diimplementasikan.
5. Penyusunan Laporan Tugas Akhir
   Penyusunan laporan merupakan tahap untuk medokumentasikan apa yang telah dikerjakan pada tugas akhir ini, lengkap dengan hasil pengujian dan analisis, serta menyajikan kesimpulan dan saran yang diperoleh dari hasil tugas akhir ini.

1.6 Sistematika Penulisan

Adapun sistematika penulisan Tugas Akhir ini adalah sebagai berikut:

1. Bab 1 Pendahuluan

   Pada Bab 1 diuraikan isi dan gambaran umum Tugas Akhir secara menyeluruh yang meliputi latar belakang, perumusan masalah, batasan masalah dan tujuan dari penyelesaian masalah yang diterapkan serta sistematika penuliasan.

2. Bab 2 Landasan Teori

   Bab 2 memaparkan dasar-dasar teori yang berkaitan dengan semua aspek yang digunakan dalam penelitian Tugas Akhir ini seperti dasar *text mining*, korpus PropBank, penjelasan mengenai fitur-fitur yang digunakan, hingga algoritma *classifier* yang diimplementasikan.

3. Bab 3 Perancangan Sistem

   Dipaparkan gambaran umum sistem yang akan dibangun, mulai dari diagram alur pengerjaan Tugas Akhir, rancangan sistem berupa perancangan data, *preprocessing*, *feature extraction*, hingga klasifikasi, juga dilengkapi dengan deskripsi kebutuhan sistem.

4. Bab 4 Pengujian dan Analisis

   Berisikan pembahasan tentang hasil pengujian yang diperoleh berdasarkan implementasi dari bab sebelumnya. Hasil pengujian yang telah dilakukan selanjutnya dianalisis, kemudian dari hasil analisis yang diperoleh tersebut selanjutnya digunakan sebagai dasar untuk pengambilan kesimpulan.

5. Bab 5 Kesimpulan dan Saran

   Berisikan hasil kesimpulan yang diperoleh dari hasil analisis yang telah dijelaskan pada bab sebelumnya, juga memaparkan saran berkaitan dengan pengembangan penelitian selanjutnya berdasarkan penelitian yang sudah di lakukan pada tugas akhir ini.