BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi sudah semakin maju. Beberapa aplikasi text mining awal menggunakan penyajian sederhana yang disebut dengan ‘bag-of-words' ketika mengenalkan struktur ke suatu kumpulan dokumen berbasis teks untuk mengklasifikasikan nya menjadi dua atau lebih kelas yang sudah ditentukan atau untuk meng-cluster nya menjadi pengelompokan-pengelompokan alami. Dalam model ‘bag-of-words’ tersebut, teks, misalnya suatu kalimat, paragraph, atau dokumen penuh, disajikan sebagai kumpulan kata, dengan mengabaikan tata bahasa atau urutan kata-kata yang akan muncul. Secara alami, kita (manusia) tidak menggunakan kata-kata tanpa suatu urutan atau struktur. Kita menggunakan kata-kata dalam kalimat, yang memiliki semantik dan dan struktur sintaksis. Teknik-teknik otomatis (seperti teks mining) harus mencari cara untuk melebihi kemampuan interpretasi ‘bag-of-words’ dan menyatukan makin lama makin banyak struktur semantik dalam operasinya. Trend saat ini dalam text mining mengarah untuk selalu memasukkan banyak fitur-fitur canggih yang bisa diperoleh dengan menggunakan pemrosesan bahasa alami (natural language processing).

Text mining (juga disebut dengan text data mining, atau knowledge discovery in textual database) adalah proses semi-otomatis dalam mengekstrak berbagai pola data (informasi dan

database yang bermanfaat) dari sumber data tak-terstruktur. Perlu diingat bahwa data mining adalah suatu proses untuk mengidentifikasi pola-pola yang valid, baru, berpotensi bermanfaat, dan akhirnya bisa dipahami yang ada di dalam data yang disimpan dalam database terstruktur, dimana data dikelola secara terstruktur berdasarkan atribut atau variable-variabel categorical, ordinal, atau continuous. Text mining sama dengan data mining dalam arti dia punya maksud yang sama dan menggunakan proses yang sama, tetapi dalam text mining input terhadap proses adalah file-file data tak-terstruktur (atau kurang terstruktur) seperti dokumen word, file-file pdf, kutipan-kutipan text, file-file XML, dan seterusnya. Pada dasarnya, text mining bisa dipikir sebagai suatu proses (dengan dua langkah utama) yang mulai dengan memaksakan struktur ke berbagai sumber data berbasis teks yang diikuti dengan mengektrak informasi dan knowledge yang relevan dari data berbasis teks yang sudah terstruktur tersebut dengan menggunakan berbagai tool dan teknik data mining(beritati, 2015).

Tagging, atau *bookmark social*, mengacu pada tindakan menghubungkan kata kunci yang relevan atau frase dengan entitas (misalnya dokumen, gambar, atau video). Dengan perkembangan baru-baru ini *Web* 2.0 aplikasi seperti *Del.icio.us* 1 dan *Flickr* dukungan yang bookmark sosial pada halaman *web* dan gambar masing-masing, layanan pemberian *tag* menjadi populer di kalangan pengguna dan telah menarik banyak perhatian dari kedua akademisi dan industri. situs

web ini memungkinkan pengguna untuk menentukan kata kunci atau *tag* untuk sumber daya, yang pada gilirannya memfasilitasi pengorganisasian dan berbagi sumber daya tersebut dengan pengguna lain. Karena jumlah data *tagged* berpotensi tersedia hampir bebas dan tidak terbatas, telah muncul dalam menyelidiki penggunaan data *mining* dan metode pembelajaran mesin untuk otomatis *tag* rekomendasi atau teks dan data digital di *web* (Chirita, et al., 2007).

Dengan rincian data tersebut, penulis berharap dengan adanya aplikasi ini dapat membantu proses NLP *(Natural Language Processing)* untuk teks berbahasa Indonesia dengan menyediakan data ekstraksi berupa data klasifikasi dan data kalimat.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas maka dapat dirumuskan masalah sebagai berikut :

1. Bagaimana mengekstrak data berita secara kolaboratif untuk membuat basis data kosakata dalam bahasa Indonesia.
2. Bagaimana mengekstrak data berita secara kolaboratif untuk membuat basis data treebank (kalimat yang telah diuraikan berdasarkan struktur semantik) tentang struktur kalimat SPOK (Subyek Predikat Obyek dan Keterangan).

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah :

1. Kosakata terdiri dari 5 kategori yaitu : kategori orang, kategori organisasi, kategori perusahaan, kategori geografis dan kategori topik.
2. Struktur kalimat hanya berbentuk SPOK.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah :

Membangun pembangunan perangkat lunak untuk membuat basis data taksonomi berita yang berguna untuk mengekstrak data berita untuk membuat basis data kosakata dalam bahasa Indonesia dan untuk membuat basis data Treebank tentang struktur kalimat SPOK.

1.5 Metode Penelitian

a. Metode Studi Pustaka

   Metode penelitian kepustakaan digunakan untuk mencari literatur atau sumber pustaka yang berkaitan dengan perangkat lunak atau aplikasi yang akan dikembangkan dan untuk membantu mempertegas teori-teori yang ada, serta memperoleh data yang sesungguhnya. Literatur dapat berupa jurnal dan atau buku yang berkaitan dengan perangkat lunak atau aplikasi yang akan dikembangkan dalam hal ini adalah tentang aplikasi ini.

b. Metode Pembangunan Perangkat Lunak

   1. Analisis, yaitu menganalisa kebutuhan dari aplikasi yang akan dibangun. Hasil analisis

berupa Spesifikasi Kebutuhan Perangkat Lunak (SKPL).

2. Perancangan, yaitu untuk mendapatkan deskripsi arsitektural perangkat lunak, antarmuka, data, dan procedural. Hasil perancangan berupa Deskripsi Perancangan Perangkat Lunak (DPPL).
3. Pembuatan program, yaitu proses penerjemahan dari desain yang telah dibuat ke bahasa pemrograman.
4. Pengujian, yaitu proses pengujian fungsionalitas perangkat lunak. Tahap pengujian mempunyai 2 macam cara. Yang pertama pengujian terhadap pembuat, pengujian ini dituliskan dalam dokumen Perancangan, Deskripsi, dan Hasil Uji Perangkat Lunak (PDHUPL). Lalu yang kedua pengujian ke tingkat pengguna dan didokumentasikan dalam bentuk kuisioner.

c. Metode Pelaporan

Metode pelaporan digunakan untuk mengetahui proses bisnis pada perangkat lunak atau aplikasi yang berkaitan. Analisis ini diperlukan untuk memahami cara penerapan proses bisnis tersebut dan mengetahui kelemahan dan kelebihan perangkat lunak atau aplikasi yang berkaitan agar dapat dikembangkan lebih baik pada aplikasi ini.

1.6 Sistematika Penulisan

Laporan ini ditulis dengan sistematika sebagai berikut :

BAB 1 : Pendahuluan

Bab ini berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan, metodologi penelitian, dan sistematika penulisan laporan.

BAB 2 : Tinjauan Pustaka

Bab ini berisi penjelasan mengenai penelitian yang pernah dilakukan sebelumnya yang berkaitan dengan topik yang dibahas, dan penjelasan mengenai perbandingan antara penelitian yang dilakukan sebelumnya dengan penelitian yang akan dilakukan.

BAB 3 : Landasan Teori

Berisi penjelasan mengenai dasar teori yang berkaitan dengan permasalan yang dibahas.

BAB 4 : Analisis dan Peneracangan Perangkat Lunak

Bab ini berisi penjelasan analisis permasalahan yang akan diatasi serta membahas mengenai perancangan perangkat lunak yang dibuat.

BAB 5 : Implementasi dan Pengujian Perangkat Lunak

Bab ini berisi penjelasan mengenai implementasi perangkat lunak yang dibuat dan gambaran umum sistem.

BAB 6 : Kesimpulan dan Saran

Bab ini berisi kesimpulan dari pembahasan secara keseluruhan beserta saran-saran uang bermanfaat untuk pengembangan lebih lanjut.

DAFTAR PUSTAKA

LAMPIRAN