**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# **IDENTITAS PENGUSUL**

Nama : **Aini Rachmania**

NRP : **5107 100 077**

Dosen Wali : **Prof. Ir. Riyanarto Sarno, SE, MSc, PhD**

1. **JUDUL TUGAS AKHIR**

***Klasifikasi Kategori dan Identifikasi Topik pada Artikel Berita Berbahasa Indonesia***

1. **LATAR BELAKANG**

Sejak tahun 1990-an, internet menjadi media vital penyebaran berita kepada khayalak ramai. Pada dasarnya, berita yang disebarkan melalui internet merupakan artikel berita tekstual yang dikemas dengan bahasa tulis yang lugas dan informatif. Selain itu, berita mempunyai perbedaan yang begitu mendasar dengan koleksi dokumen – dokumen lainnya, yaitu pada model aliran informasi. Aliran informasi sebuah berita merupakan sebuah *stream* yang dinamis dan terus diperbarui. Dokumen berita baru terus menerus datang, sehingga koleksi data selalu bertambah dan menjadi sebuah kumpulan data dengan jumlah yang sangat besar.

Dengan variasi data yang sangat besar, permasalahan mulai muncul ketika pembaca ingin meramba berita – berita yang berbeda namun mempunyai tema yang sama. Maka untuk memudahkan navigasi, berita – berita tersebut harus dipecah menjadi topik dan kategori yang berbeda – beda. Pada laman internet misalnya, biasanya dipisah menjadi halaman olahraga, bisnis, dan teknologi sendiri. Semakin besar arus dokumen berita yang masuk, makan semakin luas pula *range* topik dan kategori berita yang ada, misalnya: kategori “Olahraga” dibagi menjadi beberapa topik yang lebih spesifik seperti “balap motor” atau “sepak bola”. Kemudian dalam topik yang spesifik tersebut, terdapat beberapa berita yang mempunyai tema mengenai satu topik olahraga yang sama. Pemetaan yang terjadi antara topik dan kategori merupakan *one-to-many* yang berarti sebuah topik dapat berada pada banyak kategori, misalnya sebuah topik mengenai teknologi *hybrid* dapat saja berada pada kategori “Teknologi”, “Sains dan Ilmu Pengetahuan” atau “Lingkungan”.

Permasalahan lain yang muncul adalah saat berita menjadi suatu *raw material* yang krusial bagi profesi tertentu seperti presenter ataupun pembawa acara, terutama acara yang membutuhkan pengetahuan umum yang luas mengenai berita yang sedang hangat terjadi, seperti acara dialog atau *talk show*, sebagai sumber informasi yang aktual dan akurat, sementara dengan banyaknya jumlah berita yang ada tidak mungkin untuk membaca dan memahaminya satu persatu hingga menjadi sebuah simpulan singkat yang dapat mencakup seluruh informasi dan detail penting mengenai sebuah kejadian. Selain itu, terkadang dalam profesi tersebut, pembawa acara tidak mempunyai banyak waktu untuk mempersiapkan dirinya dengan mencari berita yang berhubungan dengan sebuah kejadian dan memahaminya. Dalam keadaan seperti itu, rangkuman singkat yang dapat mencakup seluruh detail penting mengenai sebuah topik maupun kejadian dapat menjadi sangat berguna bagi seorang pewawancara atau pembawa acara.

Seperti yang telah dijelaskan sebelumnya, berita mempunyai karakteristik yang berbeda dengan koleksi dokumen lainnya yaitu aliran dinamis berupa dokumen – dokumen baru yang mungkin saja memiliki informasi yang tidak pernah ada pada dokumen sebelumnya, tak lupa data training yang terdistribusi secara luas. Maka untuk melakukan klasifikasi topik dan kategori dibutuhkan sebuah algoritma khusus yang mampu menangani klasifikasi online, penemuan topik, dan klasifikasi menggunakan *sparse* data training.

Oleh karena itu, penulis mengusulkan sebuah aplikasi yang dapat menunjang penglasifikasian berita sesuai topik dan kategorinya, dan mengambil kesimpulan dari setiap berita tersebut. Aplikasi ini mempunyai fungsi utama menglasifikasikan berita – berita online yang ada menjadi hirarki dengan topik dan kategori yang lebih spesifik, lalu dari setiap berita yang terdapat dalam topik tersebut ditampilkan kesimpulan setiap berita. Dengan begitu, diharapkan pemahaman pembaca mengenai sebuah kejadian yang terliput dalam banyak berita akan menjadi lebih mudah.

1. **RUMUSAN MASALAH**

Rumusan masalah yang diangkat dalam Tugas Akhir ini dapat dipaparkan sebagai berikut :

1. Bagaimana membangun aplikasi yang mampu menglasifikasikan kategori berita online?
2. Bagaimana membangun sebuah aplikasi yang dapat menemukan topik dari sebuah berita yang ada?
3. Bagaimana menarik kesimpulan dari setiap dokumen berita?
4. **BATASAN MASALAH**

Permasalahan yang dibahas dalam Tugas Akhir ini memiliki beberapa batasan, diantaranya sebagai berikut:

* + - 1. Dokumen teks yang digunakan adalah berita bukan berupapublikasi ilmiah.
      2. Dokumen berita yang digunakan sebagai sumber data uji adalah dokumen teks berita berbahasa Indonesia
      3. Dokumen berita diambil dari situs berita *online* http://www.kompas.com dan http://www.detik.com.

1. **TUJUAN TUGAS AKHIR**

Tugas Akhir ini memiliki tujuan membuat sebuah aplikasi yang dapat menglasifikasikan berita ke kategori yang sesuai dan menemukan topik dari berita tersebut dan menampilkan ringkasannya.

1. **MANFAAT TUGAS AKHIR**

Manfaat perancangan Tugas Akhir ini adalah mempermudah pencarian berita online, navigasi berita melalui hirarki *drill down*  dengan kategori dan topik yang jelas, dan pemahaman berita lewat rangkuman yang singkat, namun informatif dan menyeluruh. Selain itu, dengan adanya aplikasi ini, diharapkan dapat membantu meningkatkan kerja presenter atau pembawa acara dalam mengritisi berita.

1. **RINGKASAN TUGAS AKHIR**

Pada Tugas Akhir ini penulis mengusulkan untuk membuat sebuah aplikasi yang menglasifikasikan kategori dan mengidentifikasi topik pada dokumen berita berbahasa Indonesia. Aplikasi ini berfungsi menglasifikasikan kategori berita online, dan menemukan topik, lalu menarik kesimpulan dari berita itu. Aplikasi ini berguna untuk mempermudah pembawa acara maupun presenter acara dialog untuk memahami sebuah kejadian yang dipublikasikan melalui banyak berita. Domain Tugas Akhir ini adalah dokumen berita online, dengan karakter yang berbeda dari domain – domain lainnya yaitu aliran informasi yang dinamis, sehingga mungkin saja sebuah berita baru memiliki informasi yang tidak ada di berita sebelumnya.

Algoritma yang digunakan dalam tugas akhir ini merupakan algoritma yang berbasis pada *keyword extraction* yang mampu menangani beberapa bahasa dan tidka memerlukan koleksi dokumen maupun statistik korpus. Algoritma klasifikasi kategori membangun sebuah model kategori untuk mendeskripsikan sebuah kategori. Setiap entri dalam keyword memuat *stemmed keyword*, versi terpendek dari *non-stemmed keyword*, dan sejumlah dokumen training di mana kata tersebut muncul. Keywords diekstraksi menggunakan algoritma yang di usulkan oleh Bracewell et al. Untuk setiap kategori, dilakukan *training* terhadap sebuah *classifier*. *Classifier* yang digunakan bukan merupakan classifier biner, melainkan merupakan *classifier* yang menghasilkan estimasi kemiripan yang dimiliki artikel dalam kategori tersebut. Kemiripan (*likelihood*) dari seluruh kategori digunakan untuk menentukan kategori mana yang diberikan kepada artikel tersebut.

Algoritma identifikasi topik dimulai dengan menglasifikasikan artikel yang ada sebagai topik yang sebelumnya sudah pernah ada menggunakan *one-pass clustering*. Namun, karena topik – topik baru muncul setiap hari, maka terdapat mekanisme identifikasi topik untuk menentukan apakah topik yang diberikan sebelumnya tersebut merupakan topik yang tepat. Klasifikasi topik juga berdasar pada algoritma keyword extraction yang digunakan untuk *multilingual single document keyword extraction*. Algoritma ini menghitung kemiripan antar topik yang sudah diketahui dan artikel yang ada menggunakan keyword. Kemudian algoritma ini menetapkan artikel yang paling mirip sebagai topik yang telah ditetapkan sesuai syarat yang ada. Identifikasi topik menentukan apakah topik yang telah ditetapkan sesuai syarat sudah sesuai atau harus membuat topik baru. Ini dilakukan dengan *dynamic thresholding.* Ikhtisar klasifikasi kategori dan identifikasi topik seperti pada [1] dapat dilihat dengan jelas pada gambar 1.



Gambar 1. Gambaran umum algoritma klasifikasi kategori dan identifikasi topik.

Data yang digunakan dalam aplikasi ini adalah dokumen berita berbahasa Indonesia yang diambil dari situs http://www.detik.com dan http*:*//www.kompas.comsebanyak 300 dokumen berita. Berita – berita tersebut akan diambil dari kurun waktu 1 bulan dimulai dari tanggal 1 Oktober 2010 hingga 1 November 2010.

Untuk eksperimen yang dilakukan terhadap klasifikasi kategori, digunakan sebuah *classifier* yang di-*training* dengan 1000 artikel. Kategori yang terdapat pada situs asli berita tersebut menjadi kategori yang akan ditetapkan terhadap artikel. Berikutnya, untuk identifikasi topik, beberapa uji akan dilakukan. Yang pertama adalah dengan membandingkannya dengan metode klasifikasi lain yang dilatih dengan data yang *sparse training data*. Kemudian, uji selanjutnya dilakukan kepada dua *corpora* yang berbeda. Evaluasi yang akan digunakan sebagai pembanding adalah Recall dan Precision. Diharapkan metode klasifikasi kategori menggunakan *keyword extraction* dan identifikasi topik menggunakan *likelihood calculation* ini akan menghasilkan recall dan precision yang tinggi.

1. **METODOLOGI**
2. **Studi Literatur**

Pada tahap ini dilakukan studi dan pemahaman mengenai beberapa permasalahan yang berkaitan dengan algoritma dari aplikasi ini yaitu: multilingual single document keyword extraction(untuk tahap keyword extraction pada klasifikasi kategori)*,* dan one-pass clustering (untuk tahap identifikasi topik). Selain itu, studi lebih lanjut juga akan dilakukan terhadap algoritma – algoritma pemecahan masalah yang telah ada sebelumnya seperti: Naive Bayesian Classifiers, Support Vector Machine, Boosting, dan Algoritma Em. Literatur yang digunakan meliputi buku referensi, publikasi ilmiah, dan dokumentasi dari internet.

1. **Perencanaan Perangkat Lunak**

Perangkat lunak ini akan dibagi menjadi dua aplikasi utama, yaitu aplikasi desktop dan aplikasi berbasis web. Secara umum, aplikasi ini dirancang melalui 3 fase besar yang dapat dilihat pada gambar 2.

Gambar 2. Rencana rancangan perangkat lunak

1. **Implementasi**

Aplikasi ini akan dibuat menggunakan bahasa pemrogaman Java untuk aplikasi desktop, php untuk *scripting* pada aplikasi web, dan MySQL untuk database.

1. **Uji Coba dan Evaluasi**

Uji coba akan dilakukan terhadap dua jenis kumpulan data, yaitu *training data*  dan *testing data*. *Training data* terdiri dari 1000 artikel berita, dan *testing data*  terdiri dari 500 data. Evaluasi hasil uji coba menggunakan precision dan recall sebagai pembanding terhadap *ground truth*, dalam hal ini *ground truth* adalah klasifikasi kategori dan identifikasi topik yang dianalisis secara manual oleh ahli.

1. **Penyusunan Buku Tugas Akhir**

Pada tahap ini dilakukan untuk pembuatan laporan dari semua dasar teori dan metode yang digunakan serta hasil – hasil yang diperoleh selama pengerjaan tugas akhir. Laporan tugas akhir ini akan dibagi menjadi beberapa bab sebagai berikut:

1. Bab I, Pendahuluan, berisi latar belakang, permasalahan, tujuan, batasan permasalahan, metodologi, dan sistematika penulisan.
2. Bab II, Landasan Teori, akan dibahas dasar ilmu yang mendukung pembahasan tugas akhir ini.
3. Bab III, Desain Aplikasi.
4. Bab IV, Implementasi dari aplikasi yang telah dibuat, akan dilakukan pembuatan aplikasi yang dibangun dengan komponen-komponen yang telah ada yang sesuai dengan permasalahan dan batasannya yang telah dijabarkan pada bab pertama.
5. Bab V, Uji coba dan analisa hasil, akan dilakukan uji coba berdasarkan parameter-parameter yang ditetapkan, dan kemudian dilakukan analisa terhadap hasil uji coba tersebut.
6. Bab VI, Penutup, berisi kesimpulan yang dapat diambil dari Tugas Akhir ini beserta saran untuk pengembangan selanjutnya.
7. **JADWAL PEMBUATAN TUGAS AKHIR**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Tahapan | Bulan | | | | | | | | | | | | | | | | | | | | |
| September | | | | Oktober | | | | November | | | | Desember | | | | Januari | | | | |
| 1. | Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |
| 2. | Perancangan Perangkat Lunak |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |
| 3. | Implementasi Perangkat Lunak |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |
| 4. | Uji Coba dan Evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |
| 5. | Penyusunan Buku Tugas Akhir |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |

1. **DAFTAR PUSTAKA**

[1] David B. Bracewell, Jiajun Yan, Fuji Ren, Shingo Kuroiwa, 2009. “*Category Classification and Topic Discovery of Japanese and English News Articles”*. Electronic Notes in Theoretical Computer Science. 51-65

**LEMBAR PENGESAHAN**

###### **Surabaya, Oktober 2010**

Menyetujui,

Dosen Pembimbing I

Dr. Agus Zainal Arifin, S.Kom., M.Kom.

NIP. 197208091995121001