UNIVERSITAS INDONESIA

ANALISA MEDIA SOSIAL *TWITTER* DENGAN PERHITUNGAN *GRAPH EDIT DISTANCE* UNTUK MENDETEKSI RUMOR PADA *TRENDING TOPIC* SIAK-NG

SKRIPSI

ADITYA ABIMANYU

0706275851

FAKULTAS TEKNIK UNIVERSITAS INDONESIA

PROGRAM STUDI TEKNIK KOMPUTER

DEPOK

JUNI 2012

UNIVERSITAS INDONESIA

ANALISA MEDIA SOSIAL TWITTER DENGAN PERHITUNGAN *GRAPH EDIT DISTANCE* UNTUK MENDETEKSI RUMOR PADA *TRENDING TOPIC* SIAK-NG

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik

ADITYA ABIMANYU

0706275851

FAKULTAS TEKNIK UNIVERSITAS INDONESIA

PROGRAM STUDI TEKNIK KOMPUTER

DEPOK

JUNI 2012

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri,

dan semua sumber baik yang dikutip maupun dirujuk

telah saya nyatakan dengan benar.

Nama : Aditya Abimanyu

NPM : 0706275851

Tanda Tangan :

Tanggal : 13 Juni 2012

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh	:
Nama	: Aditya Abimanyu
NPM	: 0706275851
Program Studi	: Teknik Komputer
Judul Skripsi	: Analisa Media Sosial Twitter dengan Perhitungan *Graph Edit Distance* Untuk Mendeteksi Rumor Pada *Trending Topic* SIAK-NG

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjanah Teknik pada Program Studi Teknik Komputer, Fakultas Teknik, Universitas Indonesia.

DEWAN PENGUJI

Pembimbing : Prof. Dr. Ir. Riri Fitri Sari M.Sc, MM ( )

Penguji : I Gde Dharma Nugraha S.T, MT ( )

Penguji : Yan Maraden S.T., M.Sc ( )

Ditetapkan di : Depok

Tanggal : 5 Juli 2012

UCAPAN TERIMA KASIH

Puji syukur saya panjatkan kehadirat Allah SWT, karena atas segala rahmat dan hidayat-Nya saya dapat menyelesaikan skripsi ini. Saya menyadari bahwa skripsi ini tidak akan terselesaikan tanpa bantuan dari berbagai pihak. Oleh karena itu, saya mengucapkan terima kasih kepada :

1. Ibu Prof. Dr. Ir. Riri Fitri Sari M.Sc, MM selaku pembimbing skripsi saya, terima kasih atas arahan serta koreksi skripsi saya ini dengan sabar, dan telah memberikan banyak waktu untuk mengarahkan saya hingga selesai.
2. Prof. Takako Hashimoto yang telah memperkenankan saya untuk menggunakan jurnalnya sebagai referensi dalam penulisan tugas akhir ini dan meluangkan waktunya untuk diskusi dengan video conference.
3. Orang tua dan keluarga saya yang telah memberikan bantuan dukungan material dan moral kepada saya.
4. Sahabat dan teman satu bimbingan (Mega Oktafiani Putri, Prasetya Widhi, Nur Muhammad Ridho, Slamet Budiyatno, Iyoga, Devi Zumarudin Syah) serta semua yang telah memberi semangat saya dalam menyelesaikan skripsi ini tepat waktu.
5. Dan seluruh Sivitas akademik Departemen Teknik Elektro yang tidak dapat saya sebutkan satu persatu.

Akhir kata, semoga Allah SWT berkenan membalas kebaikan semua pihak yang telah membantu. Semoga skripsi ini bermanfaat bagi perkembangan ilmu pengetahuan.

Depok, 13 Juni 2012

Aditya Abimanyu

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademika Universitas Indonesia, saya yang bertanda tangan di bawah ini :

Nama : Aditya Abimanyu

NPM : 0706275851

Program studi : Teknik Komputer

Departemen : Teknik Elektro

Fakultas : Teknik

Jenis karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Nonoksklusif (*Non-exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul :

Analisa Media Sosial Twitter dengan Perhitungan *Graph Edit Distance* Untuk Mendeteksi Rumor Pada *Trending Topic* SIAK-NG.

Dengan Hak Bebas Royalti Non Eksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta sebagai pemegang Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok

Pada tanggal : 13 Juni 2012

Yang menyatakan

Aditya Abimanyu

ABSTRAK

Nama : Aditya Abimanyu

Program Studi : Teknik Komputer

Judul : Analisa Media Sosial Twitter dengan Perhitungan Graph Edit Distance Untuk Mendeteksi Rumor Pada Trending Topic SIAK-NG

Pesatnya perkembangan teknologi disertai dengan tingkat penggunaannya membawa dampak positif di berbagai bidang kehidupan manusia, namun juga dapat membawa dampak negatif jika tidak didukung dengan tanggung jawab pengguna teknologi itu sendiri. Bidang telekomunikasi adalah salah satu bidang yang perkembangannya sangat dirasakan oleh manusia. Salah satu dari perkembangan telekomunikasi adalah lahirnya media sosial. Manusia menggunakan media sosial untuk berbagi informasi apapun kepada siapapun. Namun yang menjadi masalah kemudian adalah apakah informasi yang tersebar merupakan informasi yang nilai kebenarannya telah teruji atau hanya sebuah rumor. Rumor dapat saja mengakibatkan tersebarnya informasi yang salah di suatu golongan atau komunitas manusia. Adapun topik yang terkait pada tugas akhir ini adalah siak-ng yang menjadi trending topic di media sosial twitter. l. Mengidentifikasi rumor pada media sosial online sangat krusial nilainya karena mudahnya informasi yang disebar oleh sumber yang tidak jelas. Pada tugas akhir ini akan ditunjukkan salah satu cara pengidentifikasian rumor dengan menggunakan kalkulasi graph edit distance. *Graph edit distance* merupakan salah satu langkah yang paling cocok untuk menentukan persamaan antar grafik dan pengenalan pola jaringan kompleks. Untuk mencapai tujuan akhir, langkah-langkah yang dilakukan adalah pengambilan data, konversi data, pengolahan data, dan visualisasi. Dengan pengolahan data didapat Sembilan padanan kata antara *Parent Node dan Child Node* serta 3 kategori *edge label*. Pada akhirnya ditemukan bahwa rumor sistem siak-ng sedang mengalami load tinggi merupakan rumor yang nilai kebenarannya tinggi.

Kata kunci : *Graph edit distance*, Visualisasi, Rumor, Trending Topic, Relevansi.

ABSTRACT

Name : Aditya Abimanyu

Major : Teknik Komputer

Title : Implemetation of Graph Edit Distance for Rumor Detection on Twitter Trending Topic: Case Study UI Academic Information System

Rapid development of technology coupled with the utilizing bring positive impact in many areas of human life, but also have negative impacts if not supported with the responsibility of the users. Telecommunications is one area in which development is perceived by humans. One of the development of telecommunications is social media established.Humans use social media to share any information with anyone. However, the issue then is whether the spread of information is information whose truth value has been tested or just a rumor. Rumors will lead to the spread of false information in a group or people's community. The topics related to this thesis is the SIAK-NG become trending topic on social media Twitter. Identifying online rumors on social media is crucial value because of the information ease spread by unverified sources. At the end of this assignment will be demonstrated one way of identifying the rumor by using graph edit distance calculations. Graph edit distance is one of the most appropriate steps to determine the similarities between graphs and pattern recognition of complex networks.. To achieve the ultimate goal, the steps taken are data retrieval, data conversion, data processing, and visualization. By data processing obtain nine words comparison between Parent node and Child Node with three edge label category. Finally, the tweet that said the system has high range of load was the true rumor.

Keywords: Graph Edit Distance, Visualization, Rumor, Trending Topic, Relevance.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Dunia teknologi informasi dan telekomunikasi semakin canggih dan pesat dengan adanya perkembangan internet. Saat ini teknologi informasi dan telekomunikasi sudah tidak dapat dipisahkan dari kehidupan sehari-hari dan sudah menjadi kebutuhan untuk memenuhi dan mendukung berbagai macam kegiatan, baik individu maupun organisasi.

Dengan teknologi, setiap orang dapat mengakses dan mendapat informasi secara cepat, tanpa mengenal batas-batas wilayah dan batasan waktu. Ini menyebabkan informasi menjadi sesuatu yang berharga dan sangat dibutuhkan guna mengambil keputusan. Namun kecepatan informasi itu berkembang kadangkala dapat menimbulkan kekhawatiran, karena informasi yang berkembang tersebut tidak selalu positif. Banyak kasus dimana informasi yang berkembang hanyalah rumor semata yang belum jelas kebenarannya.

Informasi yang bersifat rumor bisa menimbulkan reaksi yang beraneka ragam, yang menjadi masalah adalah apabila rumor yang beredar menyebabkan keresahan di masyarakat yang kemudian menimbulkan reaksi yang berlebihan yang bisa menimbulkan kerugian. Contoh kasus yang bisa diambil adalah pada saat terjadi *tsunami* di Jepang, pada saat itu masyarakat Jepang sedang mengalami beban berat akibat bencana yang terjadi di negaranya kemudian tersebar rumor yang menyebutkan bahwa terjadi krisis bahan bakar. Hal tersebut menimbulkan kepanikan dalam masyarakat yang membuat situasi di daerah bencana semakin tidak kondusif. Masalah seperti ini dapat diminimalisasi seandainya ada suatu sistem yang dapat mengidentifikasi kebenaran dari rumor yang berkembang sehingga masyarakat tidak perlu mengalami kekhawatiran akibat informasi yang tidak benar.

1.2 Tujuan Penulisan

Tujuan dari skripsi ini adalah mengimplementasikan suatu *framework rumor analysis* yang memanfaatkan statistik dan probabilitas yang dipadukan dengan sistem basis data untuk kemudian dijadikan sebagai alat bantu dalam menganalisa kebenaran rumor yang berkembang di media social.

1.3 Pembatasan Masalah

Pada skripsi ini yang dibahas dibatasi pada implementasi *framework rumor analysis* dalam bentuk sistem basis data yang bersumber pada media sosial, dimana media sosial yang digunakan adalah *twitter*.

1.4 Identifikasi Masalah

Dalam penulisan skripsi ini, penulis mengidentifikasikan masalah berdasarkan tema dan judul yang telah ditetapkan. Berikut ini adalah identifikasi masalah dalam skripsi ini :

1. Bagaimana mengumpulkan rumor yang beredar pada media sosial *twitter*?
2. Bagaimana mengidentifikasi kebenaran rumor yang telah dikumpulkan berdasarkan bobot?
3. Bagaimana cara memvisualisasikan rumor yang terkumpul dalam bentuk grafik berdasarkan bobotnya?
4. Apakah *framework rumor analysis* ini berhasil mengidentifikasi kebenaran rumor yang berkembang pada media sosial *twitter*?

1.5 Metodologi Penulisan

Metode penulisan pada skripsi ini adalah:

a. Studi literatur, yaitu mencari buku-buku, jurnal dan sumber-sumber lainnya untuk dijadikan bahan referensi.
b. Pendefinisian masalah, yaitu mendefinisikan masalah apa saja yang akan dibahas dalam skripsi ini.
c. Perancangan sistem, yaitu merancang desain *framework rumor analysis* yang akan dibuat.

d. Implementasi dan Uji Coba, yaitu menerapkan rancangan yang telah dibuat dan kemudian diuji coba untuk diambil datanya sebagai bahan analisa.

1.6 Sistematika Penulisan

Skripsi ini terdiri dari empat bab, dimana masing-masing bab akan menjelaskan sebagai berikut:

a. Bab 1: Pendahuluan

Pada bab ini, akan dijelaskan mengenai Latar Belakang, Tujuan, Pembatasan Masalah, Metodologi Penulisan, dan Sistematika penulisan.

b. Bab 2: Konsep *Framework Rumor Analysis*

Pada bab ini, akan dijelaskan tentang landasan atau dasar teori yang digunakan dalam penulisan skripsi ini yang berhubungan dengan masalah yang dibahas yaitu mengenai *rumor analysis,* metode analisa yang digunakan dan hal-hal yang mendukungnya seperti *software* dan aplikasi pendukung.

c. Bab 3: Perancangan *Framework Rumor Analysis*

Pada bab ini, akan dijelaskan mengenai perancangan sitem analisa rumor yang berupa algoritma pengerjaan.

d. Bab 4: Implementasi dan Analisa Sistem Analisa Rumor

Pada bab ini, akan dibahas mengenai proses implementasi dari rancangan sistem analisa rumor yang telah dibahas pada bab 3. Proses implementasi ini terdiri dari proses instalasi dan konfigurasi sistem serta analisa hasil pembobotan.

e. Bab 5: Kesimpulan

Pada bab ini, akan dijelaskan mengenai kesimpulan yang dapat diambil dari pembahasan skripsi ini.

BAB 2
KONSEP SISTEM KERANGKA ANALISA RUMOR

2.1 Media Sosial

Media sosial telah menjadi bagian kehidupan dari pengguna internet di dunia. Dapat dikatakan sebagian besar waktu yang dihabiskan dalam penggunaan internet adalah mengunjungi situs-situs media sosial, seperti *Facebook*, *Twitter*, *Myspace*, *Blog*, dan lain-lain. Arti media sosial sendiri adalah sebuah media yang menggunakan internet sebagai medianya, dan memungkinkan penggunanya untuk berpartisipasi, berbagi, dan menciptakan isi meliputi *blog*, jejaring sosial, wiki, forum dan dunia virtual.

Gambar 2.1. Diagram motivasi jejaring sosial [1]

Sementara jejaring sosial merupakan situs dimana setiap orang bisa membuat *web page* pribadi, kemudian terhubung dengan teman-teman untuk berbagi informasi dan berkomunikasi. Jejaring sosial adalah struktur sosial yang terdiri dari elemen-elemen individual atau organisasi. Jejaring ini menunjukan jalan dimana mereka berhubungan karena kesamaan sosialitas, mulai dari mereka yang dikenal sehari-hari sampai dengan keluarga. Istilah ini diperkenalkan oleh profesor J.A. Barnes di tahun 1954. Jejaring sosial adalah suatu struktur sosial yang dibentuk dari simpul-simpul (yang umumnya adalah individu atau organisasi) yang diikat dengan satu atau lebih

tipe relasi spesifik sepert inilai, visi, ide, teman, keturunan, dan lain-lain [2].

Situs jejaring sosial pada mulanya merupakan situs yang berfokus pada hubungan antar teman satu sekolah, yaitu Classmates.com pada tahun 1995 dan Six Degrees.com pada tahun 1997 yang membuat ikatan tidak langsung. Dua model berbeda dari jejaring sosial yang lahir sekitar pada tahun 1999 adalah berbasiskan kepercayaan yang dikembangkan oleh Epinions.com, dan jejaring sosial yang berbasiskan pertemanan seperti yang dikembangkan oleh Uskup Jonathan yang kemudian dipakai pada beberapa situs UK regional di antara 1999 dan 2001. Inovasi meliputi tidak hanya memperlihatkan siapa berteman dengan siapa, tetapi memberikan pengguna kontrol yang lebih akan isi dan hubungan. Pada tahun 2005, suatu layanan jejaring social MySpace, dilaporkan lebih banyak diakses dibandingkan Google dengan Facebook, pesaing yang tumbuh dengan cepat. Jejaring sosial mulai menjadi bagian dari strategi internet bisnis sekitar tahun 2005 ketikaYahoo meluncurkan Yahoo! 360°. Pada bulan Juli 2005 News Corporation membeli MySpace, diikuti oleh ITV (UK) membeli Friends Reunited pada Desember 2005. Diperkirakan ada lebih dari 200 situs jejaring sosial menggunakan model jejaring sosial ini[3].

Jejaring sosial terbesar antara lain *Facebook, Myspace, Plurk,* dan *Twitter*. Jika media tradisional menggunakan media cetak dan media elektronik, maka media sosial menggunakan internet. Media sosial mengajak siapa saja yang tertarik untuk berpertisipasi dengan memberi kontribusi dan *feedback* secara terbuka, memberi komentar, serta membagi informasi dalam waktu yang cepat dan tak terbatas.

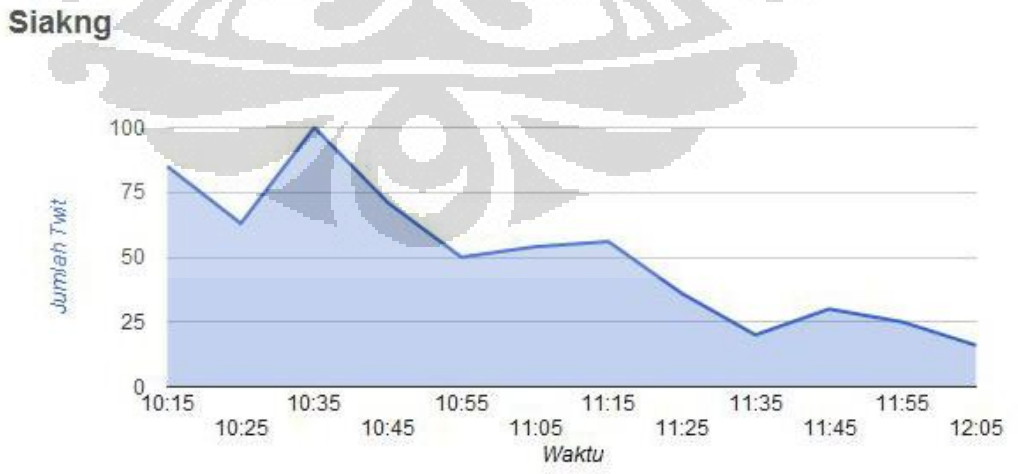
Pada tugas akhir ini akan dipergunakan media sosial *Twitter* sebagai sumber informasi. Twitter adalah suatu situs web layanan jaringan sosial dan mikroblog yang memberikan fasilitas bagi pengguna untuk mengirimkan "pembaharuan" berupa tulisan teks dengan panjang maksimum 140 karakter melalui SMS, pengirim pesan instan, surat elektronik, atau aplikasi seperti Twitterrific dan Twitbin. Twitter didirikan pada Maret 2006 oleh perusahaan rintisan Obvious Corp. Kesuksesan Twitter membuat banyak situs lain meniru konsepnya, kadang menawarkan layanan spesifik lokal suatu negara atau menggabungkan dengan layanan lainnya.

Twitter dapat menjadi sumber yang sangat bermanfaat untuk mengumpulkan data yang dipergunakan dalam menelusuri rumor-rumor yang berkembang. Keragaman dari pengguna *Twitter* membuat bahan informasi tersebut memiliki nilai lebih. Pada September 2010, didapatkan data 95 juta tweet per hari [4]. Ini menjadikan *Twitter* sebuah sumber informasi tepat untuk menganalisa kebenaran rumor di media sosial.

Disamping kelebihan-kelebihan Twitter di atas, terdapat pula kelemahan *Twitter*, yaitu tata bahasa yang kurang baik dikarenakan batas huruf (karakter) yang dipergunakan dalam sekali tweet hanya 140 karakter. Hal lainnya yaitu terdapat informasi-informasi yang bersifat sarkas dan humor.

Penggunaan twitter pada tugas akhir ini bertujuan untuk menggali informasi atau data yang berhubungan dengan SIAK-NG. SIAK-NG sendiri merupakan singkatan dari Sistem Informasi Akademis – New Generation milik Universitas Indonesia. SIAK-NG digunakan civitas academica untuk berbagai hal, seperti misalnya registrasi akademis, pengisian IRS, penyusunan rencana studi, dan penyusunan jadwal kuliah ataupun jadwal ujian. Pada masa pengisian IRS di awal semester baru biasanya terdapat kendala yang muncul akibat banyaknya mahasiswa yang mengakses SIAK-NG.

Berikut merupakan grafik tweet pada awal masa pengisian IRS[5].



Grafik 2.1. Jumlah tweet dengan *tag* SIAKNG.

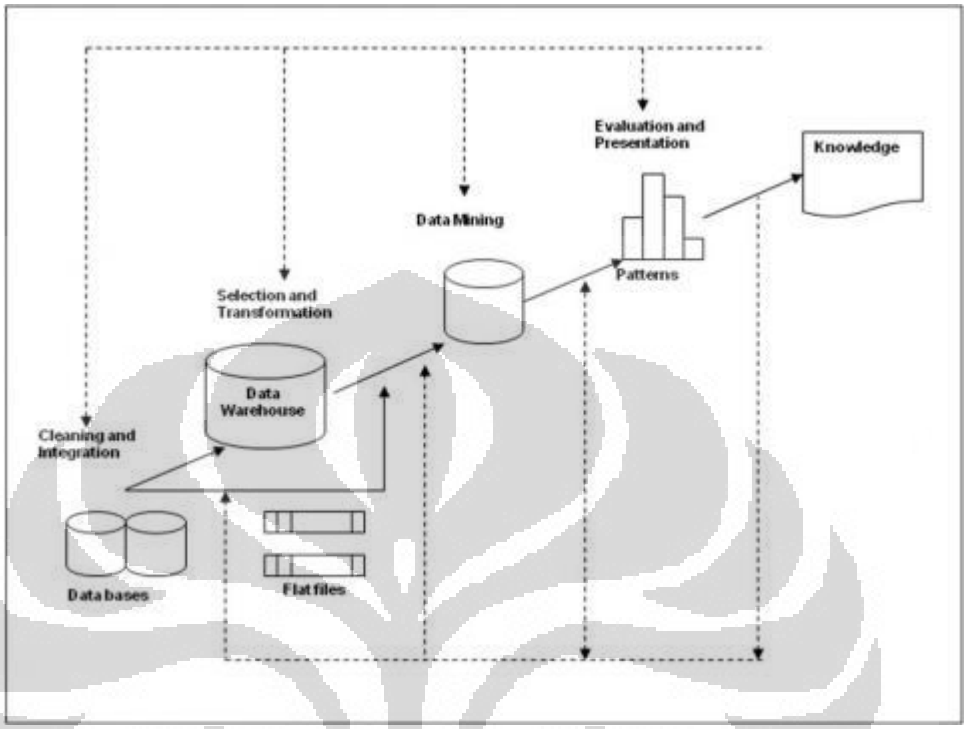
2.2 Teknik Penggalian Web

Perkembangan informasi berbasis *web* pada masa kini telah berkembang dengan sangat pesat dan memiliki jumlah data yang menyentuh kisaran jutaan *terabytes*. Ini artinya terdapat informasi terselubung (*hidden*) yang sangat besar pula dari berbagai kemungkinan tujuan. Untuk itu diperlukan cara yang efektif dalam menggali informasi dari web. Salah satu dari cara tersebut adalah *Data Mining* dan *Text Mining*. Baik *Data Mining* maupun *Text Mining* menggunakan metode analitis signifikan dan menghasilkan output visual dan grafik yang berkualitas baik.

2.2.1 Data Mining

Penggalian data didefinisikan sebagai ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database besar. Pola tersebut harus didukung dengan sifat-sifat : implisit, tidak popular, dan memiliki daya guna.

Sebagian sumber mengatakan istilah alternative dari Data Mining, seperti *Knowledge Discovery in Database* (KDD), analisis pola, arkeologi data, pemanenan informasi, dan intelegensia bisnis. Penggalian data diperlukan saat data yang tersedia terlalu banyak (misalnya data yang diperoleh darisistem basis data perusahaan, e-commerce, data saham, dan data bioinformatika), tapi tidak tahu pola apa yang bisa didapatkan.



Gambar 2.2. Konsep Data Mining [6]

Berikut merupakan tahapan proses Data Mining :

1. Pembersihan Data; yaitu menghapus data pengganggu (noise) dan mengisi data yang hilang.
2. Integrasi Data; yaitu menggabungkan berbagai sumber data.
3. Pemilihan Data; yaitu memilih data yang relevan.
4. Transformasi Data; yaitu mentransformasi data ke dalam format untuk diproses dalam penggalian data.
5. Penggalian Data; yaitu menerapkan metode cerdas untuk ekstraksi pola.
6. Evaluasi pola; yaitu mengenali pola-pola yang menarik saja.
7. Penyajian pola; yaitu memvisualisasi pola ke pengguna.

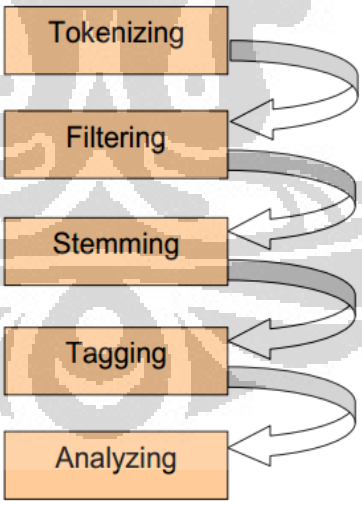
2.2.2 Text Mining

Text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi[7]. Pada dasarnya text mining hampir sama dengan data mining, hanya saja yang membedakan adalah sumber data yang digunakan. Sumber data tersebut berupa data text yang bersifat tidak terstrukur.

Masalah yang sering muncul dalam text mining sama halnya dengan data mining, yaitu jumlah data yang besar, dimensi tinggi dan data yang tidak diinginkan (noise). Sedangkan perbedaan text mining dengan data mining adalah data yang digunakan. Pada text mining, seperti yang telah dijelaskan di atas, menggunakan sumber data tidak terstruktur, sedangkan data mining menggunakan data terstruktur.

Tujuan dari penggunaan text mining sendiri adalah untuk mendapatkan informasi yang diinginkan pengguna dari dokumen-dokumen. Selain itu untuk dapat mengkategorikan dan mengelompokkan teks.

Tahapan pada text mining [8] :

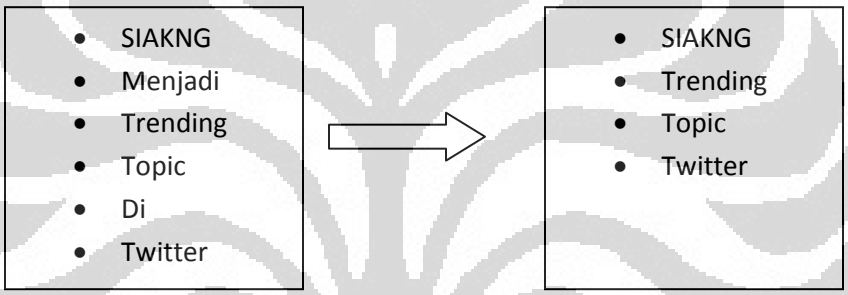


Gambar 2.3. Tahap Text Mining

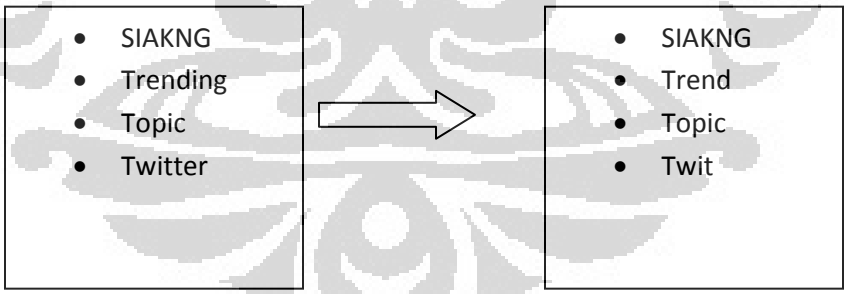
1. Tahap tokenizing / parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.

   Contoh : SIAK NG menjadi trending topic di twitter

2. Tahap filtering adalah tahap mengambil kata - kata penting dari hasil token.



3. Tahap stemming adalah tahap mencari *root* kata dari tiap kata hasil filtering.



4. Tahap tagging adalah tahap mencari bentuk awal dari tiap kata lampau atau kata hasil stemming.
5. Tahap analyzing merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata antar dokumen yang ada.

2.3 Ektraksi Web

Ekstraksi Web merupakan suatu cara penggalian informasi dari web. Istilah popular dari ekstraksi web adalah *web scraping*. *Web Scraping* berkaitan erat dengan *Web Indexing*, dimana indeks informasi pada sebuah web menggunakan *bot* dan merupakan teknik universal yang diadopsi oleh sebgaian besar *Search Engine*. Sebaliknya, *Web Scraping* lebih berfokus pada transformasi data terstruktur di web, biasanya dalam format HTML, menjadi data terstruktur yang dapat disimpan dan dianalisa dalam database lokal pusat atau spreadsheet. *Scraping Web* juga terkait dengan otomatisasi web, yang mensimulasikan penjelajahan manusia menggunakan perangkat lunak komputer. Penggunaan Scraping web meliputi perbandingan harga online, pemantauan data cuaca, penelitian, *web mashup* dan web integrasi data.

2.3.1 Teknik Dasar Web Ekstraksi

Pada dasarnya banyak teknik yang dapat dilakukan untuk mengekstraksi web. Namun secara umum ada tiga teknik dasar dalam pengekstraksian web, yaitu:

1. Pengumpulan Data (*retrieving data/web content harvesting)*

   Teknik pertama ini berkaitan dengan mengumpulkan informasi pada web yang relevant, seperti file HTML, gambar atau e-mail. Pada umumnya teks dokumen relative tidak tersruktur untuk itu dilakukan pembelajaran mengenai struktur umum yang membentuk suatu dokumen yang telah terdefinisi jenisnya terlebih dahulu dan memetakannya dengan model data atau dokumen sejenis dalam suatu tempat penyimpanan local. Proses ini membutuhkan tools untuk mencari dan mengarahkan Web, contohnya *crawler* dan *means* untuk berinteraksi dengan *web pages* yang dalam dan dinamis, dan juga alat untuk membaca, meng-index dan membandingkan konteks isi dari halaman web tersebut.

2. Pemfilteran Data (*extracting data/ web structure harvesting)*

   Teknik kedua ini berkaitan dengan proses pengidentifikasian data yang relevant untuk mendapatkan isi dari halaman web tersebut dan untuk kemudian diekstraksi ke sebuah format yang terstruktur. Tools penting yang mengizinkan akses ke dalam data untuk analisis lebih jauh yaitu *parsers, content spotters,* dan *adaptive wrappers*. Jika kita amati pada kenyataannya *web pages* dapat memberikan informasi lebih bukan hanya isi *web* biasa. Link yang terpasang pada satu dokumen biasanya emmiliki keterkaitan topic dengan isi atau informasi yang terdapat pada dokumen itu.

3. Pengintegrasian Data (*integrating data/web usage harvesting)*

   Teknik ketiga ini berkaitan dengan proses *cleaning, filtering, transforming, refining* dan *combining* data yang diekstraksi dari satu atau lebih web sources, dan terakhir *structuring*. Sehingga hasilnya akan sesuai dengan format output yang diinginkan. Aspek penting dari proses ini adalah mengorganisasi data yang telah diekstraksi tersebut ke dalam bentuk yang memungkinkan akses kesatuan untuk proses analisis dan pengumpulan data lebih lanjut. Proses pengumpulan data dapat dilakukan melalui rekaman *web server* mengenai *user interactions* untuk memahami kebiasaan *user* dan mengevaluasi efektifitas dari *Web structure*.

2.3.2 Mekanisme Pengekstraksian Web

Tidak dapat diragukan bahwa internet merupakan sumber data terbesar di dunia dan dengan melihat kenyataan bahwa data yang tersedia terus bertambah secara eksponensial, maka kita dapat menyimpulkan bahwa terdapat potensi yang sangat besar untuk mengumpulkan data melalui internet dibandingkan dari sumber lain.

Masalah timbul setelah mendapatkan halaman *web* yang dicari, *link-link* yang ada pada halaman *web* tersebut satu persatu harus di-*browsing* untuk mencari informasi yang diinginkan. Hal ini harus dilakukan secara *online*. Ditambah lagi proses penyimpanan harus dilakukan secara manual untuk tiap halaman *web*. Dengan proses web ekstraksi kita dapat mengumpulkan web dengan topic yang sama dalam

satu database. Beberapa hal yang perlu dipersiapkan dalam pembentukan web ekstraksi ini adalah:

- Infrastruktur yang terpusat :

Dibutuhkan infrastruktur yang terpusat yang dipelihara oleh institusi/consortium yang terkait. Karena sekali suatu gateway dari data yang terintegrasi mulai melakukan kegiatannya, ia akan berkembang sangat cepat dan akan segera membutuhkan *backup* financial untuk pemeliharaan dan ekspansi lebih jauh seiring meningkatnya *traffics, spaces* dan memori untuk menampung semua data.

a. Metode yang cerdas dan efisien untuk mengumpulkan data yang relevan:
    - Crawling, yaitu mengumpulkan data tekstual dari internet lalu menyimpannya di database. Dengan crawling, data-data secara umum dalam jumlah besar diunduh dari internet dan disimpan di database lokal sehingga dapat dimanfaatkan secara offline
    - Indexing, yaitu memindai kata per kata dari seluruh teks yang ada di database, kemudian membuat daftar kata pencarian yang disebut index.
    - Searching, yaitu mencari informasi yang diperlukan dari dalam database tersebut berdasarkan kata kunci yang terdaftar di index. Fungsi searching memungkinkan pencarian dari dalam tumpukan data-data tersebut, informasi tertentu yang diperlukan oleh penelitian ini untuk diteruskan ke bagian selanjutnya untuk dianalisa

b. Kecerdasan analisis data untuk menghasilkan index database yang akurat:

    Pengembangan teknologi berbasis text-mining untuk menyaring data yang terkumpul dan merekonstruksi index database ke struktur yang baik dan memiliki ruang lingkup yang luas untuk end user.

    Dengan tujuan untuk meningkatkan tingkat akurasi dan menghindari menyia-nyiakan sumber irrelevant web pages, dilakukan adopsi terhadap proses web ekstraksi/ *web harvesting* konvensional dengan lebih memanfaatkan setup parameter human-guidance. Berikut ini prosedur yang dapat dilakukan untuk bisa memperoleh target ekstraksi yang diinginkan:

1) Menentukan target URL yang spesifik, dengan informasi yang relevan dengan topik data terintegrasi yang sedang dibangun. URL tersebut harus diarahkan langsung ke halaman pertama dari jenis isi yang diinginkan.
2) Menghitung kedalaman = skala dari halaman yang diarahkan oleh URL tersebut di atas ke halaman akhir dimana data yang relevan dikumpulkan.
3) Menempatkan beberapa kriteria yang terkait dengan halaman antara halaman yang dituju dengan akhir halaman untuk membedakan hyperlink yang tersimpan yang langsung diarahkan ke judul konten terakhir atau page number dari list yang diinginkan. Kriteria *paging* ini terdiri dari beberapa parameter, termasuk *keyword*, parameter *separator*, parameter *count* dan sebagainya.
4) Menentukan *focus-point number* dari isi yang diinginkan pada *final page*. *Point number* tersebut ditujukan kepada "box" yang berisi konten yang relevan. User dapat menentukan sebuah parameter sebagai penunjuk untuk tujuan ini, seperti misalnya *table tag* dari bagian atas halaman sampai ke "box". Proses marking seperti ini cukup efektif untuk menunjukkan informasi yang relevan.
5) Meletakkan beberapa keywords untuk mengenali hperlink yang terdapat di dalam box. Baik link yang yang mengarah ke relevan image, full-text, maupun yang lainnya. Sama halnya dengan criteria paging, criteria halaman konten juga berisi parameter, termasuk *keyword*, parameter *separator*, parameter *count* dan sebagainya..
6) Memilih periode pengumpulan data atau ekstraksi web ulang sesuai dengan frekuensi update dari setiap jenis konten.

2.4 Konsep Framework

Konsep *framework* dapat dijelaskan sebagai fungsi atau perintah yang membantu *user* untuk menyelesaikan masalah dengan cara sederhana. Akibatnya waktu yang dibutuhkan menjadi lebih singkat. Definisi *framework* dapat juga diartikan sebagai koleksi atau kumpulan potongan-potongan program yang disusun atau diorganisasikan sedemikian rupa, sehingga dapat digunakan untuk membantu membuat aplikasi utuh tanpa harus membuat semua kodenya dari awal.

Sebagai ilustrasi, seorang *user* yang mengerjakan suatu proyek pembuatan sebuah halaman web dapat dengan mudah dan efisien menyelesaikan proyek tersebut dengan menggunakan bagian-bagian dari proyek yang pernah dikerjakannya diwaktu yang lalu. Biasanya dalam pembuatan website ada bagian-bagian yang selalu ada untuk digunakan, bagian tersebut dapat di*edit* dan digunakan kembali pada pengerjaan proyek yang lain.

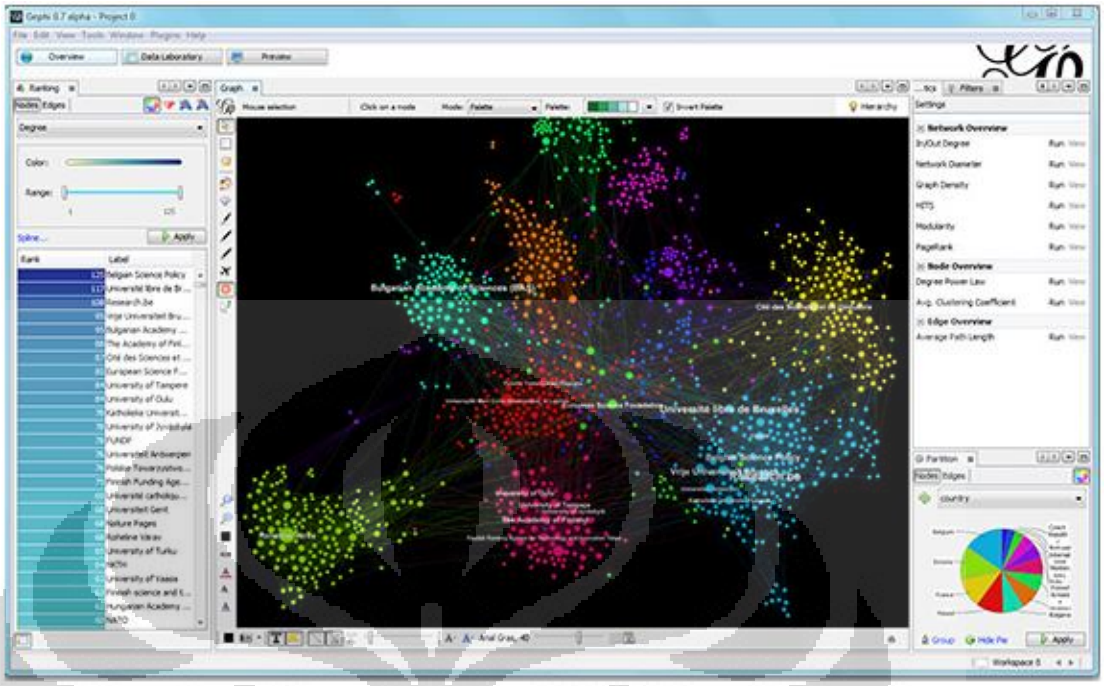
2.5 Perangkat Lunak Yang Digunakan

Dalam pengerjaan tugas akhir ini terdapat alat bantu dalam bentuk perangkat lunak yang digunakan.

2.5.1 Gephi Graph

Gephi merupakan perangkat lunak open source untuk visualisasi dan analisa jaringan. Perangkat lunak ini membantu analis data untuk mengungkapkan pola dan tren intuitif, *highlight outlier* dan menceritakannya dengan data tersebut. Gephi menggunakan mesin render 3D untuk menampilkan grafik berukuran besar dalam real-time dan untuk mempercepat eksplorasi [9].

Grafik dapat dibuat dan dimuat menggunakan format file grafik yang umum digunakan, dan dapat pula dieksplorasi secara interaktif. Algoritma tata letak Gephi secara otomatis memberi bentuk pada grafik untuk memberi bentuk pada grafik untuk membantu eksplorasi, dan pengguna dapat merubah warna sesuai kehendak dan tata letak parameter untuk meningkatkan penampilan.



Gambar 2.4. Gephi Graph

Gephi menggabungkan fungsi *built-in* dan arsitektur fleksibel untuk eksplorasi, analisa, manipulasi, *export*, penyaringan, *spatialize* semua jenis jaringan. Gephi berbasis pada paradigma visualisasi dan manipulasi, yang memungkinkan *user* untuk menemukan jaringan dan sifat data.

2.5.2 Twitter Application Programming Language (API)

API merupakan sebuah gerbang dimana perangkat lunak eksternal dapat mengakses program utamanya, yaitu Twitter yang digunakan pada tugas akhir ini. Karena Twitter merupakan aplikasi berbasis internet, maka dibutuhkan suatu aplikasi lain yang mendukung koneksi antara suatu website atau aplikasi lain dengan Twitter. Aplikasi yang dimaksud adalah API.

API sangat bermanfaat untuk pengguna Twitter. Salah satu kegunaan dari API sendiri adalah dapat menggali tweet-tweet tentang suatu query, pada tugas akhir ini query yang diinginkan adalah siakng. Disamping manfaat itu juga terdapat batasan penggunaan yang hanya 100 API per jamnya bagi akun personal.

2.5.3 Website topsy.com

Topsy.com merupakan situs untuk mengetahui berbagai macam hal yang berkaitan dengan media sosial. Situs ini memiliki API yang telah populer digunakan berbagai *search engine*, agregator konten, penerbit online dan *marketer*. Perangkat lunak ini dirancang untuk penyebaran data konsumer dengan jumlah besar.

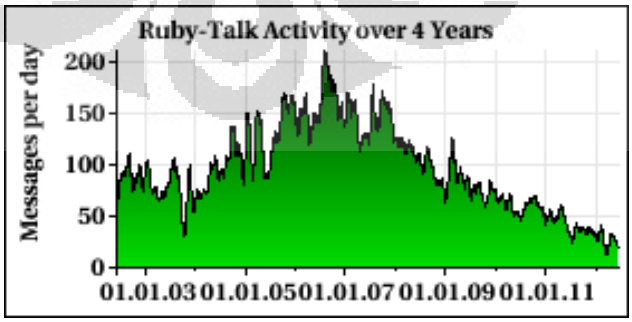
Topsy digunakan dalam *crawling* data *tweet* pada tugas akhir ini. Data yang didapat *username*, nama asli pengguna, isi tweet, waktu tweet, berapa kali di-*retweet*, siapa saja yang me-*retweet*. Adapun *query* yang digunakan adalah 'siakng'.

2.5.4 Bahasa Pemrograman Ruby

Ruby merupakan bahasa pemrograman yang seimbang. Maksud dari seimbang itu sendiri adalah bahasa yang seimbang antara pemrograman fungsional dan imperative. Ruby pertama kali dikenalkan pada tahun 1995 oleh sang pencipta, yaitu Yukihiro Matsumoto. Matsumoto menggabungkan berbagai bahasa yang digemarinya seperti Perl, Smalltalk, Eiffel, Ada dan Lisp.

Ruby dianggap sebagai bahasa yang fleksibel, karena bagian-bagian dari Ruby bisa diubah-ubah dengan bebas. Bagian-bagian yang esensi di Ruby bisa dihapus maupun didefinisikan ulang. Bagian-bagian yang sudah ada bisa ditambahkan. Ruby mencoba untuk tidak membatasi programmer.

Kini banyak programmer yang menggunakan Ruby dan mengembangkannya. Hal ini dikarenakan Ruby merupakan *open source* yang tidak hanya gratis, tetapi juga bebas untuk menggunakan, memodifikasi dan mendistribusikannya.



Gambar 2.5. Aktifitas diskusi Ruby

Grafik di atas menunjukkan jumlah *email* yang masuk di milis Ruby-Talk. Ruby-Talk sendiri merupakan sebuah milis diskusi bagi pengguna Ruby. [12]

2.5.5 GVedit

GVedit merupakan sebuah Graphviz tool untuk membuat, melihat, mengedit dan memproses file DOT. GVedit mengizinkan penggunanya untuk mengatur atribut suatu grafik menggunakan kotak dialog dan menyimpannya untuk kebutuhan mendatang. GVedit merupakan perangkat lunak pendukung untuk tahap visualisasi dari sistem.

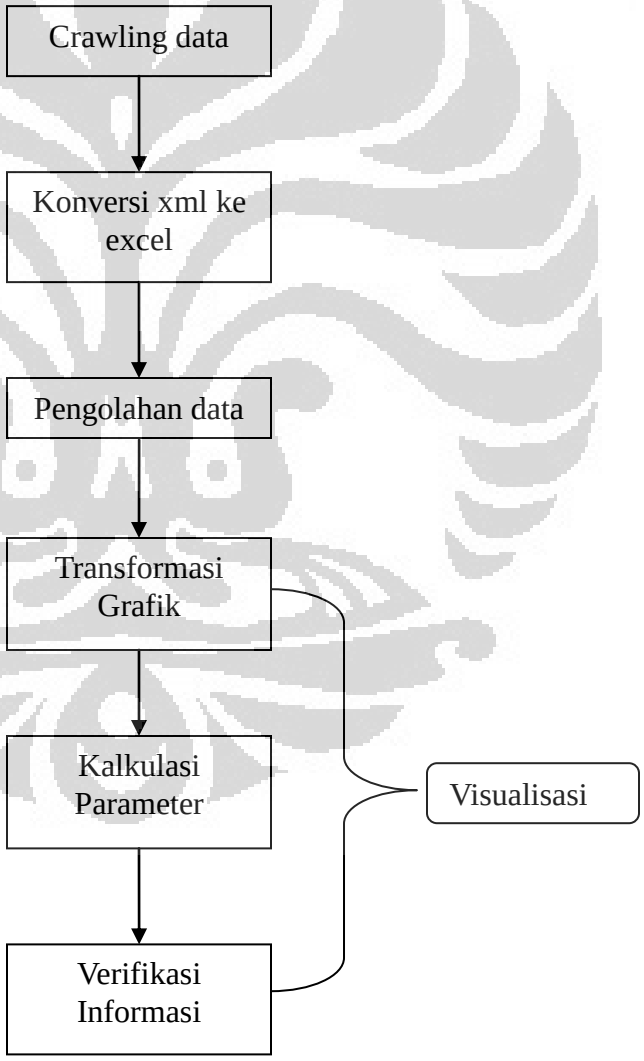
Dengan menggunakan GVedit beberapa grafik dapat dibuka disaat bersamaan. Akan tetapi hanya ada satu grafik yang aktif, grafik yang aktif ini yang akan mengalami perubahan apabila user melakukan pengaturan menggunakan GVedit menu. Pengaturan setiap grafik terpisah satu sama lain. Pada tampilan jendela pengaturan yang tampak merupakan pengaturan dari grafik yang aktif.

Untuk membuat suatu grafik menjadi aktif, klik pada jendela grafik tersebut. Latar belakang dari grafik yang aktif berwarna putih sedangkan grafik yang lain berwarna abu-abu. [13]

BAB 3
PERANCANGAN KERANGKA ANALISA RUMOR SIAK-NG DI TWITTER

3.1 Algoritma Sistem

Sistem ini dirancang dengan tujuan untuk mengumpulkan dan menganalisa rumor atau *tweet* yang berkembang di *Twitter*, tentunya *tweet* dengan tag "siakng" ataupun dengan *hashtag* #siakng, dan kemudian menyajikannya dalam bentuk grafik. Grafik tersebut disajikan dalam bentuk multiform serta memberikan penyajian visual yang mudah dipahami. Implementasi sistem dapat dilihat pada algoritma di bawah.



Gambar 3.1. Algoritma Sistem Analisa Rumor SIAK NG di Twitter.

3.2 Deskripsi Algoritma Sistem

Asumsinya ketika terdapat beberapa tweet yang mengandung aspek kontroversi namun memiliki topik yang hampir sama, maka harus ada cara bagaimana menentukan mana rumor yang benar dan mana rumor yang melenceng dari fakta. Terdapat 2 tujuan pokok dari tugas akhir ini, pertama mengektraksi *tweet* yang memiliki aspek kontroversial dari rumor yang berkembang dan persebaran informasi yang salah, dan kedua adalah mengidentifikasi pengguna yang percaya akan rumor yang salah tersebut serta pengguna yang menolak atau masih mempertanyakan kebenaran rumor tersebut.

Berikut ini terdapat dua tweet yang memiliki *corpus* sama yaitu siakng. Tweet pertama merupakan contoh tweet non-rumor atau fakta, dan berikutnya merupakan tweet yang berisi informasi yang salah mengenai siakng.

"Tulisan Kamu: Balada **#SIAKNG**: Trending Topic Indonesia" (non-rumor)

"Sistem **SIAKNG** saat ini tidak dapat diakses. Liburan diperpanjang sampai bulan Agustus." (rumor)

Untuk dapat memperoleh hasil seperti contoh di atas maka diimplementasikanlah algoritma sistem yang disebutkan pada subbab 3.1. Berikut ini pemaparan lebih lanjut dari tiap tahapan algoritma sistem analisa rumor siak ng di *twitter*.

3.2.1 Pengambilan Data

Tahap pertama dari algoritma sistem ini yaitu pengambilan data yang dilakukan dengan cara *crawling* data di Twitter dengan metode *Web Content Mining. Crawling* bertujuan untuk menggali data berupa *tweet* yang bersumber pada media sosial Twitter. Data-data berupa *tweet*, waktu, pengguna, dan keterangan lain yang mendukung didapat dengan bantuan layanan dari situs *topsy.com* dan Twitter API. Twitter dipilih karena merupakan sumber yang sangat baik untuk menganalisa

informasi yang salah. Penulis menggunakan Twitter untuk menggali informasi mengenai SIAK-NG pada periode 31 Januari 2012 hingga kini. Penggalian informasi mengenai SIAK-NG menggunakan *query* 'siakng' yang diposting pada 31 Januari 2012 hingga 29 Maret 2012. Ada 796 *tweet* yang berhasil didapatkan pada situs '*topsy.com'*.

3.2.2 Konversi Data

Tahap kedua dalam algoritma sistem ini adalah konversi data. Konversi yang dilakukan adalah mengubah data yang berbentuk TXT menjadi data *CSV*. Data hasil *crawling* berbentuk TXT sehingga untuk mempermudah pengolahan data maka data tersebut harus dikonversi kedalam bentuk *CSV*.

3.2.3 Pengolahan Data

Tahap selanjutnya adalah pengolahan data. Awal pengolahan data dilakukan dengan text mining, yaitu pemisahan kata-kata dari tweet hasil *crawling*. Tahap ini menggunakan *output* dari tahap ke dua sebagai inputannya dimana satu *tweet* dianggap sebagai satu dokumen. Pada tahap ini dilakukan ekstraksi kata kunci seperti subjek, predikat, objek dan keterangan menggunakan analisa morfologi.

Nilai dari masing-masing ekstraksi kata kunci dikalkulasi dengan metode *Residual IDF (RIDF), Latent Semantic Analysis (LSA), dan Term Frequency-Inverse Document Frequency (TF-IDF)*. LSA merupakan metode untuk menggali dan merepresentasikan konteks yang digunakan sebagai sebuah arti kata dengan memanfaatkan komputasi statistik untuk *corpus* yang besar. TF-IDF merupakan metode perhitungan bobot dari setiap term pada suatu dokumen. Sedangkan RIDF adalah kombinasi antara IDF dan distribusi Poisson. Menurut hemat penulis, RIDF sesuai untuk mengekstraksi kata kunci yang menunjukkan isi dokumen [11].

Sistem ini dibuat menggunakan bahasa pemrograman Ruby. Program ini akan di *run* menggunakan CMD dengan inputan berupa file CSV yang telah diperoleh pada tahap sebelumnya kemudian menghasilkan data berupa file *.dot*.

3.2.4 Visualisasi

Visualisasi bertujuan untuk memperlihatkan struktur konsep grafik yang dibuat pada pengolahan data. Visualisasi terdiri atas beberapa tahap yang pertama adalah transformasi grafik yang menggunakan input dari hasil olahan data yang berupa matriks antara *message id* dan kata kunci dengan hasil tertinggi. Grafik yang akan dibuat adalah grafik yang menunjukkan struktur informasi rumor. Terdapat simpul *parent*, simpul *child*, bobot, dan waktu *posting*. Grafik yang akan dibuat berdasar pada bobot kata kunci *tweet*. Bobot dihitung dengan melihat korelasi antara simpul *parent* dan simpul *child*. Tahap selanjutnya kalkulasi parameter, bertujuan mendeteksi informasi rumor menggunakan *Graph topology-based distance* [10] untuk mengukur besar perubahan yang terjadi pada waktu tertentu. Pada tahap akhir, verifikasi informasi, diketahui mana rumor yang merupakan fakta dan mana yang mengandung informasi yang salah. Inputan dari tahap visualisasi ini adalah file .dot yang diperoleh dari tahap pengolahan data.

3.3 Tahap Pengujian dan Analisa

Tahap terakhir adalah tahap pengujian. Tujuan dari tahap ini adalah untuk mengetahui apakah sistem yang telah dibuat telah bekerja sesuai dengan tujuan awalnya. Pada tahap ini hasil keluaran dari sistem rumor analisis yang diperoleh akan diuji dan dianalisa. Untuk mendapatkan tujuan akhir berupa *tweet* yang memiliki tingkat kebenaran yang tinggi, maka diperlukan pengujian terhadap data-data berupa *tweet* yang telah digali. Pengujian yang dilakukan meliputi pemeriksaan isi *tweet*, penghitungan nilai RIDF dan LSI, visualisasi grafik untuk mendapatkan nilai modularity dan relevansi.

BAB 4
IMPLEMENTASI DAN ANALISA SISTEM KERANGKA ANALISIS RUMOR

Pada bab sebelumnya telah dilakukan perancangan dari sistem yang akan dibuat yaitu sistem kerangka analisa rumor SIAK-NG di Twitter. Setelah tahap perancangan maka pada bab ini masuk ke tahap implementasi dan analisa sistem.

4.1 Tahap Implementasi

Pada bagian ini akan dijelaskan tahapan implementasi berupa instalasi dan konfigurasi dari tiap komponen sistem dari tahap instalasi software sampai dengan pengambilan data.

Step 1: Instalasi Ruby

Ruby merupakan bahasa pemrograman yang digunakan untuk mengimplementasi sistem analisa rumor. Proses instalasi Ruby cukup mudah seperti proses instalasi program-program lain pada umumnya. Installer Ruby diunduh melalui website http://www.ruby-lang.org/id/. Ruby merupakan bahasa pemrograman open source sehingga tidak perlu membayar untuk mengunduh dan meng-*install*-nya.

Step 2: Instalasi GVedit

GVedit merupakan program yang dibutuhkan untuk mengolah file .dot yang diperoleh dari hasil pengolahan data. GVedit merupakan program pendukung untuk tahap visualisasi. Proses instalasinya pun sama seperti proses instalasi program pada umumnya, cukup mengikuti perintah yang tampil di layar sampai dengan selesai.

Step 3: Modifikasi Script

Script yang digunakan dalam implementasi sistem ini merupakan script yang dibuat oleh Prof. Takako Hashimoto dari *Chiba University of Commerce* yang merupakan penulis dari jurnal *Rumor Analysis Framework in Social Media*.

Modifikasi dilakukan pada bagian *library* program dan pada bagian perhitungannya. Library dari program ini di modifikasi sehingga terdiri atas empat buat file:

1. Katadenganimbuhan.txt

   Library ini memuat informasi mengenai hubungan kata-kata menggunakan imbuhan dan kata-kata dasarnya serta hubungan singkatan dengan kepanjangannya. Tujuan dari *library* ini adalah untuk mengambil kata dasar.
2. Tandabaca.txt

   Library ini memuat informasi mengenai tanda baca dalam bahasa indonesia.
3. Katasambung.txt

   Library ini memuat kata sambung, kata tunjuk , dan kata-kata lain yang tidak memiliki kedudukan pendukung pada kalimat.
4. Stopwordindo.txt

   Library ini memuat *stopword* untuk Bahasa Indonesia. *Stopword* merupakan kata yang memiliki frekuensi tinggi namun tidak memiliki makna seperti: kata sapaan, angka, istilah asing, koma, *emoticon* dan lain lain.

Keempat file ini berguna pada proses data crawling yang akan membantu mengkonversi kalimat yang tidak baku menjadi kalimat yang baku, sehingga mudah diolah pada tahap selanjutnya yaitu tahap pengolahan bahasa.

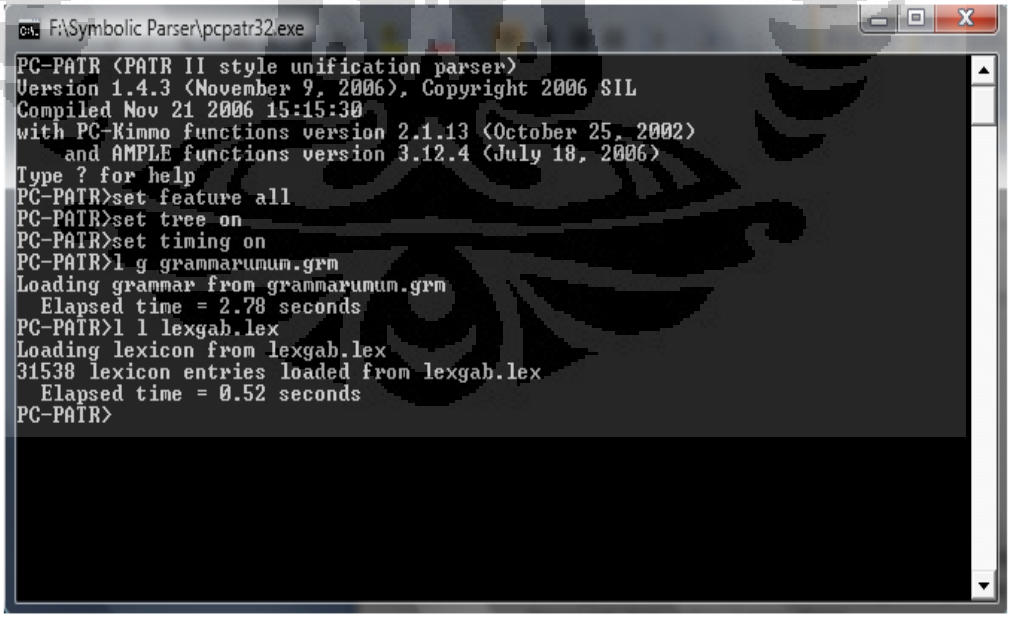
Step 4: Crawling Data dan Konversi Data

Proses selanjutnya pada tahap implementasi ini ada melakukan pengambilan data yang dilakukan dengan cara *crawling* data di Twitter dengan metode *Web Content Mining. Crawling* bertujuan untuk menggali data berupa *tweet* yang bersumber pada media sosial Twitter. Data-data berupa *tweet*, waktu, pengguna, dan keterangan lain yang mendukung didapat dengan bantuan layanan dari situs *topsy.com*. Data yang telah diperoleh tersebut kemudian dikonversi ke dalam bentuk CSV. Data harus dikonversi menjadi CSV dikarenakan program dengan bahasa pemrograman ruby yang digunakan hanya bisa menerima inputan file CSV.

Hasil penggalian (*crawling*) *tweet* dengan *query* siakng yang dilakukan dengan bersumber pada situs topsy.com. Didapatkan 796 *tweet* dengan *query* siakng dari tanggal 30 Januari 2012 hingga 29 Maret 2012. Hasil *crawling* dapat dilihat pada Gambar 4.1.

Gambar 4.1 Tampilan hasil pencarian *tweet* pada *website Topsy.com*

Selanjutnya akan dilakukan pemeriksaan isi *tweet* yang dapat dikenali oleh *library* aplikasi *Symbolic Parser.*



Gambar 4.2 Symbolic Parser

Penggunaan *Symbolic Parser* dikolaborasikan dengan *notepad++*, pemeriksaan *tweet* dilakukan dengan *Symbolic Parser* lalu *notepad++* berguna untuk menghilangkan kata-kata tidak penting ataupun merubah suatu kata yang tidak baku menjadi kata baku. Gambar 4.2 menunjukkan tampilan *Symbolic Parser*.

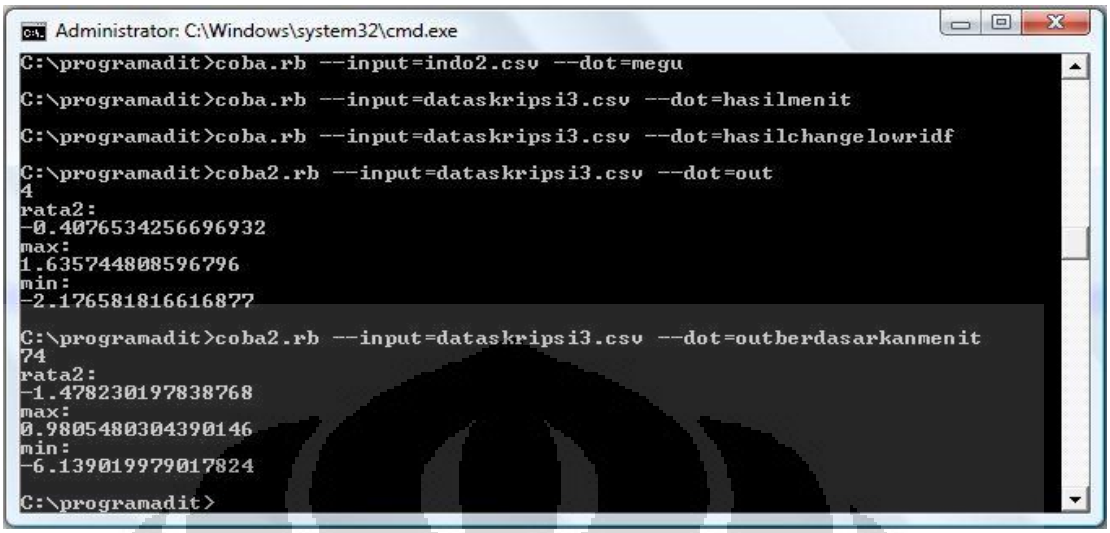
Step 5: Pengolahan Data

Data yang telah diperoleh selanjutnya diproses dengan bantuan CMD. Untuk me-running program ruby yang telah dibuat diperlukan bantuan CMD dengan menggunakan command:

```
C:\program>coba.rb --input=dataskripsi3.csv --dot=out
```

Gambar 4.3 Command yang digunakan pada Pengolahan Data

Command pada Gambar 4.3 selanjutnya divariasikan berdasarkan output yang diinginkan. Pengolahan data ini outputnya divariasikan menjadi 4 macam yaitu output berdasarkan kategori, output berdasarkan menit, output berdasarkan jam dan output berdasarkan pesan. Setiap command di atas masing-masing akan menghasilkan 4 buah file DOT yang membagi outputan berdasarkan tanggal. File DOT yang dihasilkan dengan menggunakan program GVedit data tersebut kemudian divisualisasi menjadi grafik. Selain file DOT, program coba.rb juga akan mengeluarkan outputan berupa nilai bobot rata-rata, nilai bobot max, dan nilai bobot minimal dari data yang diolah, seperti yang tampak pada Gambar 4.4.



Gambar 4.4 Tampilan CMD setelah command pengolahan data dijalankan

Output divariasikan dengan cara memodifikasi bagian dari program *coba.rb*. Nilai "c" pada blok program coba.rb pada Gambar 4.4, seperti yang telah disebutkan sebelumnya akan diubah-ubah berdasarkan jenis output yang diinginkan. Gambar 4.5 menunjukkan blok program untuk memvariasikan output.

```
    t = Date.strptime(date, "%Y/%m/%d")

      d_dt[id]= t.strftime("%Y%m%d")
#c= cat2.sub(/ > .*$/,'') #dokumen berdasarkan kategori
    #c = time #dokumen berdasarkan menit
#time = time[0..-4]#dokumen berdasarkan jam
#c = time
#puts c
#c = rt
c = id #dokumen perpesan
```

Gambar 4.5 Blok program *coba.rb* untuk memvariasikan output

Step 6 : Pengolahan Bahasa

Penentuan *threshold* = 1 dan kata kunci yang memiliki nilai RIDF tinggi memudahkan pencarian *tweet* dengan nilai kebenaran tinggi. Hasil dari parsing data yang telah dilakukan selanjutnya akan menjadi input dari pengolahan data.

Pengolahan data yang dilakukan bertujuan mencari *Term Frequency, Document Frequency, Invers Document Frequency, Residual Invers Document Frequency, dan TF-IDF*.

Tabel 4.1. Pengolahan data RIDF

Term	df	D	tf	Idf	Idf	ridF	Tf*idf
Siakng	44	64	44	0,162727297	-0,29858	-0,13585	7,160001
Siak	8	64	8	0,903089987	-0,05429	0,848803	7,22472
Server	5	64	5	1,10720997	-0,03393	1,073281	5,53605
Down	5	64	5	1,10720997	-0,03393	1,073281	5,53605
Admin	2	64	5	1,505149978	-0,03393	1,471221	7,52575
IRS	8	64	8	0,903089987	-0,05429	0,848803	7,22472
UI	9	64	9	0,851937465	-0,06107	0,790865	7,667437
Kampus	2	64	2	1,505149978	-0,01357	1,491578	3,0103
akademis	1	64	1	1,806179974	-0,00679	1,799394	1,80618
Sistem	8	64	8	0,903089987	-0,05429	0,848803	7,22472
mahasiswa	2	64	2	1,505149978	-0,01357	1,491578	3,0103
Load	2	64	2	1,505149978	-0,01357	1,491578	3,0103
trending	11	64	11	0,764787289	-0,07464	0,690143	8,41266
Topic	11	64	11	0,764787289	-0,07464	0,690143	8,41266
semester	7	64	7	0,961081934	-0,0475	0,913581	6,727574

Pada Tabel 4.1 mengenai pengolahan data di atas, terdapat 8 kolom sebagai berikut

a. Term

   Isi kolom term merupakan *keyword* yang terdapat pada dokumen *tweet*.

b. *Document Frequency* (df)

   Df berisi jumlah dokumen yang terdapat suatu keyword yang diinginkan.

c. Dokumen (D)

   Jumlah dari dokumen yang berisi 64 dokumen *tweet*.

d. *Term Frequency* (tf)

   Hampir sama dengan Document Frequency, namun nilainya berdasar pada jumlah kata atau keyword yang terdapat pada seluruh dokumen.

e. *Inverse Document Frequency* (idf)

IDF biasanya digunakan pada Information Retrieval. Perhitungan Idf menggunakan rumus - l o g 2 d f / D , dengan D adalah jumlah dokumen dan df adalah Document Frequency.

f. Residual IDF

Kombinasi antara IDF dan distribusi Poisson.

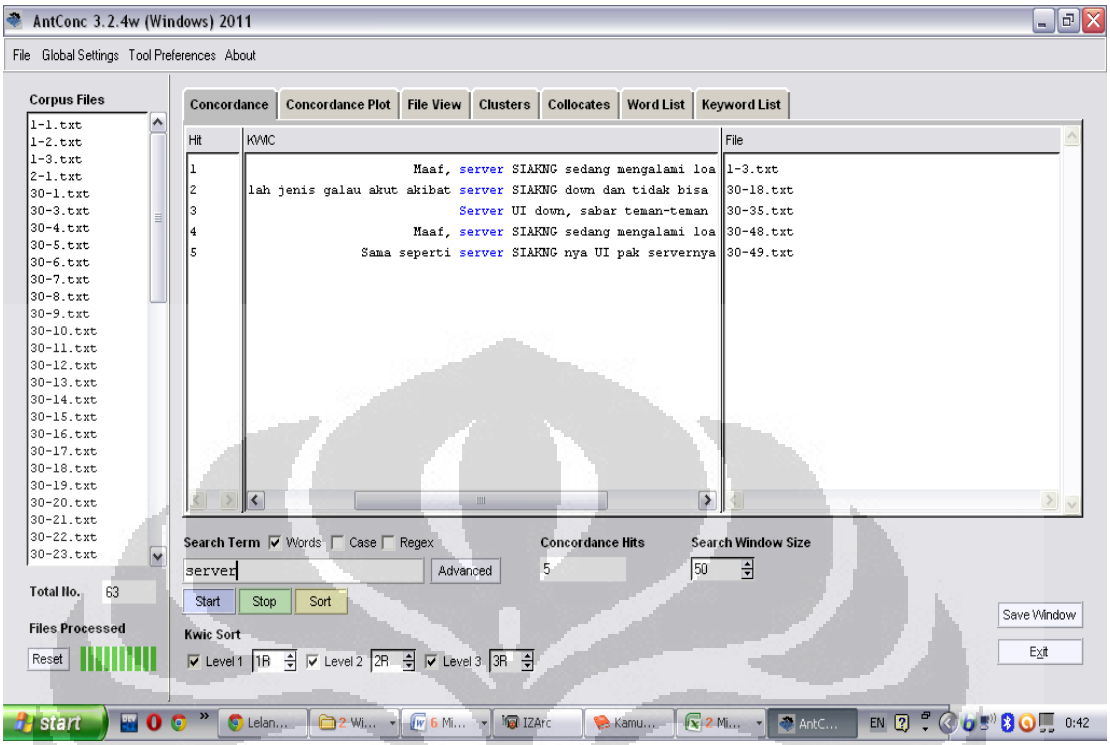
g. TF-IDF

Penghitungan bobot dari setiap term pada suatu dokumen.

Tabel 4.2. Kata Kunci untuk Pengolahan Data RIDF

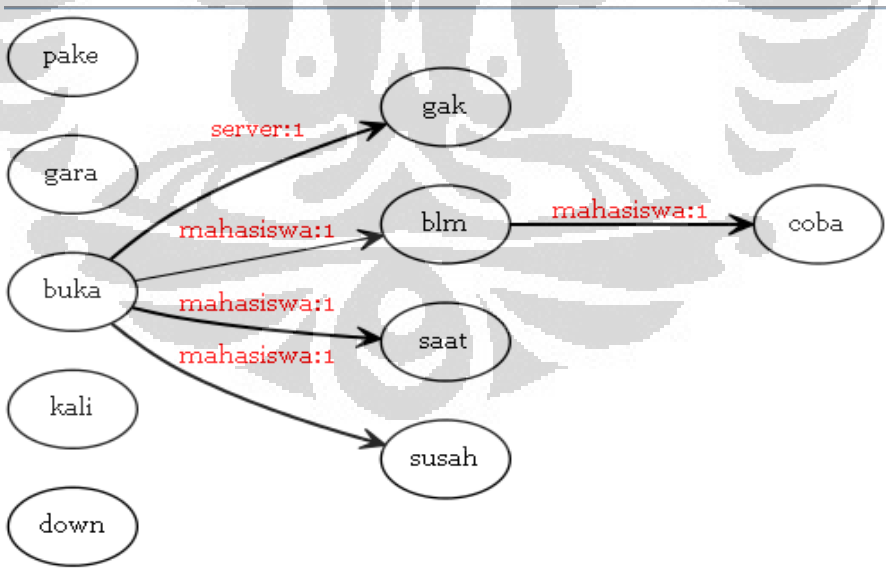
tweet ID	**kata kunci (RIDF>1)**
1-3	server, load
30-18	server, down
30-35	server, down
30-48	server, load
30-49	Server
30-41	down, kampus
31-2	Down
30-19	Kampus
30-6	Mahasiswa
30-2	Mahasiswa
30-30	Akademis

Tabel 4.2 melampirkan keseluruhan kata kunci yang memiliki nilai RIDF di atas 1. Tweet ID merupakan nama dari dokumen yang di dalamnya terdapat kata kunci yang nilai RIDF melebihi angka 1. Dalam pengerjaannya, digunakan sebuah program yang bernama *AntConc* untuk pencarian kata kunci dari *corpus file* yang telah ditetapkan. Tampilan program *AntConc* dapat dilihat pada Gambar 4.6.



Gambar 4.6 AntConc

Step 7: Visualisasi Data



Gambar 4.7 Bentuk visualisasi data dari hasil pengolahan data CSV

Gambar 4.7 menunjukkan ilustrasi dari sebuah contoh mengenai visualisasi grafik, terdapat subgraph yang mendiskusikan tentang server SIAK-NG yang mengalami kerusakan atau sering juga disebut *server down*. Hal ini mengindikasikan bahwa *tweet* yang menyebutkan limit server SIAK-NG tidak sebanding dengan jumlah mahasiswa yang mengakses SIAK-NG pada waktu yang bersamaan tersebut benar adanya. Selain itu ada tema lain yang sedang dibicarakan, seperti SIAK-NG menjadi *trending topic* di Indonesia. Suatu hal yang jarang terjadi, kecuali pada masa registrasi akademik di Universitas Indonesia

Step 8 : Transformasi Grafik

Saat pengujian dilakukan, data *tweet* dibatasi antara kurun waktu 30 Januari 2012 hingga 2 Februari 2012. Pada langkah ini penulis membuat grafik untuk menunjukkan struktur informasi rumor. Untuk mendukung grafik tersebut, dibutuhkan konsep yang membuat hubungan relevansi hipernim dari kata kunci yang muncul dalam satu set dokumen. Untuk membentuk sebuah concept graph, suatu set dokumen akan diambil sesuai kata kunci tertentu lalu kemudian kata-kata yang terkait tersebut diekstraksi. Maka relasi hipernim dari kata-kata yang terkait diperoleh dengan menggunakan frekuensi *co-occurrence*. Hirokawa [14] menggunakan hubungan *upper-lower* antara kata-kata dalam dokumen sebagai *concept graph*. Set dari keseluruhan dokumen target direpresentasikan sebagai *"U"*. Diberikan sebuah subset dari *"U"* sebagai *"X",* dan kata kunci *"u"* dan *"v", "df(u, X)"* mewakili jumlah dokumen dalam *X* yang berisi kata kunci *"u"* dan *" df(u * v, X) "* mewakili jumlah dokumen yang berisi *"u"* dan *"v"* dalam *X*. Relevansi antara "v" dan "u" didefinisikan dengan Persamaan 4.1:

$$r\,(v,u) = \frac{df\,(u*v,X)}{df\,(v,X)} \qquad (4.1)$$

Step 9 : Penentuan Hasil

Langkah yang terakhir adalah bagaimana menemukan informasi yang muncul seperti informasi rumor. Dengan menganalisa variasi waktu serangkaian *concept graph*, informasi yang dibutuhkan dapat dideteksi lebih tepat. Untuk mendeteksi informasi rumor, digunakan grafik topologi berbasis jarak untuk mengukur perubahan dalam topologi concept graph dari waktu ke waktu. Sebuah *concept graph* dinotasikan oleh G = (V , E , α , β), dimana V merupakan kumpulan simpul yang terbatas, dan E ≤ V x V merupakan kumpulan edge. Setiap simpul dilabelkan oleh sebuah *labeling function* $\alpha : V \rightarrow L_v$, dimana Lv adalah kumpulan dari *node label*, dan tiap edge dilabelkan oleh sebuah *labeling function* $\beta : E \rightarrow L_E$ dimana L_E adalah kumpulan dari *edge label*. *Concept graph* pada tugas akhir ini dianggap sebagai grafik dengan *node label* yang unik. Oleh karena itu grafik *edit distance* De antara dua grafik $G_1 = (V_1, E_1, \alpha_1, \beta_1)$ dan $G_2 = (V_2, E_2, \alpha_2, \beta_{21})$ diformulasikan dalam Persamaan 4.2 :

$$D_e\,(G_1, G_{12}) = |V_1| + |V_1| - 2|\alpha(V_1) \cap \alpha(V_2)| + |E_1| + |E_2| - 2|\beta(E_1) \cap \beta(E_2)| \quad (4.2)$$

Didefinisikan α (V) sebagai {α(v) Є L_v | v Є V} dan β (E) sebagai {β(e) Є L_E | e Є E} berturut-turut. Grafik *edit distance* mengukur perubahan absolut dari struktur grafik dari waktu ke waktu. Sehingga grafik *edit distance* sangat berguna untuk mendeteksi perubahan struktur global.

4.2 Hasil dan Analisa Hasil

Analisa yang dilakukan berdasar pada olah data *tweet* dengan tag SIAK-NG di media sosial twitter. Didapatkan 796 *tweet* dengan *query* siakng dari tanggal 30 Januari 2012 hingga 29 Maret 2012. Dari jumlah tersebut selanjutnya dilakukan eliminasi data berupa *retweet* ataupun *tweet* yang tidak sesuai dengan topik.

Pengukuran dilakukan dengan variasi dokumen berdasarkan kategori, jam, menit dan pesan. Variasi tersebut dilakukan dengan pembobotan dengan *Residual*

Inverse Document Frequency (RIDF) atau pun *Term Frequency - Inverse Document Frequency* (TF-IDF). Variasi dokumen mempunyai definisi sebagai berikut :

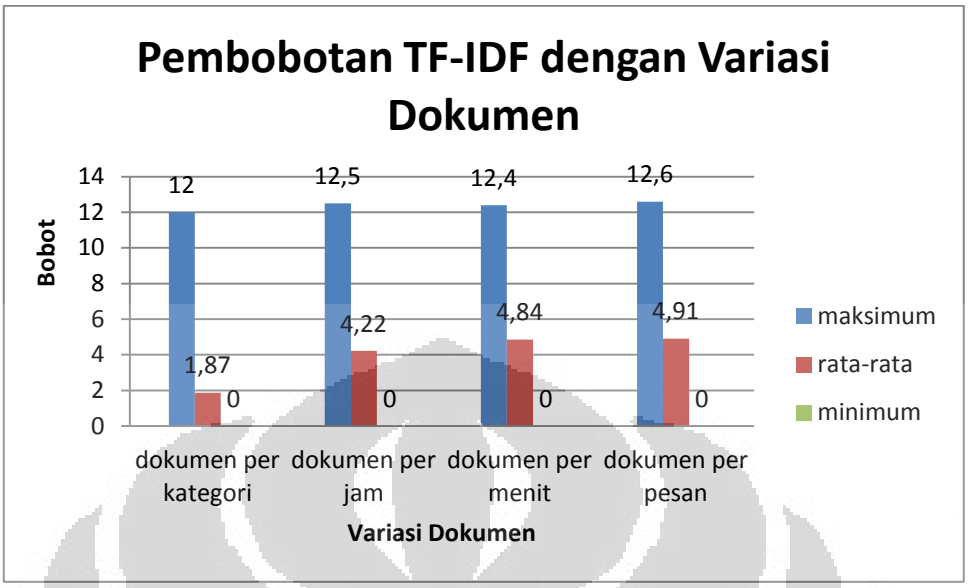
1. Dokumen per kategori: mendefinisikan dokumen berdasarkan kategori yang telah didefinisikan sebelumnya. Pemilihan kata kunci untuk dijadikan sebuah kategori mengacu pada Tabel 4.2.
2. dokumen per jam: mendefinisikan bahwa semua pesan yang dipublikasikan pada jam sama merupakan satu dokumen.
3. dokumen per menit: mendefinisikan bahwa semua pesan yang dipublikasikan pada menit yang sama merupakan satu dokumen.
4. Dokumen per pesan : mendefinisikan bahwa sebuah dokumen berisi pesan-pesan yang dipilih secara acak *(random).*

Pengukuran selanjutnya berdasar pembobotan *Residual Inverse Document Frequency* (RIDF) atau pun *Term Frequency - Inverse Document Frequency* (TF-IDF). Hal ini dilakukan untuk perbandingan dan mendapatkan hasil yang akurat dalam penentuan grafik konsep mana yang akan digunakan dalam mendeteksi rumor.

Tabel 4.3 dan Gambar 4.8 menunjukan hasil pengukuran dengan pembobotan *Term Frequency - Inverse Document Frequency* (TF-IDF).

Tabel 4.3 Hasil Pengukuran TF-IDF

TF-IDF	Dokumen per kategori	Dokumen per jam	Dokumen per menit	Dokumen per pesan
Bobot maksimum	12	12,5	12,4	12,6
Bobot rata-rata	1,87	4,22	4,84	4,91
Bobot minimum	0	0	0	0



Gambar 4.8 Pembobotan TF-IDF dengan variasi dokumen

Metode pembobotan TF-IDF merupakan metode pembobotan dalam bentuk metode integrasi antara *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* (TF) dapat diartikan frekuensi kemunculan sebuah kata atau *term* dalam sebuah dokumen. Oleh karena itu nilai TF bervariasi dari satu dokumen dengan dokumen lain tergantung kepada tingkat kepentingan sebuah kata dari dokumen. Sedangkan *Inverse Document Frequency* (IDF) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Semakin sedikit dokumen yang mengandung term yang dimaksud, maka nilai idf semakin besar. Jika setiap dokumen dalam koleksi mengandung term yang bersangkutan, maka nilai dari idf dari term tersebut adalah nol. Hal ini menunjukkan bahwa sebuah term yang muncul pada setiap dokumen dalam koleksi tidak berguna untuk membedakan dokumen berdasarkan topik tertentu.

Formula yang digunakan dalam menghitung bobot berdasarkan metode ini yaitu:

$$w(t, d) = tf(t, d) * \log N/nt \qquad (4.3)$$

Bobot suatu term *t* dalam suatu dokumen *d* dilambangkan dengan *w(t,d).* Frekuensi kemunculan term t dalam dokumen d dilambangkan dengan *tf(t,d),* Sedangkan banyaknya dokumen yang digunakan dalam uji coba dilambangkan dengan *N* sementara *nt* adalah banyaknya dokumen yang mengandung term *t.*

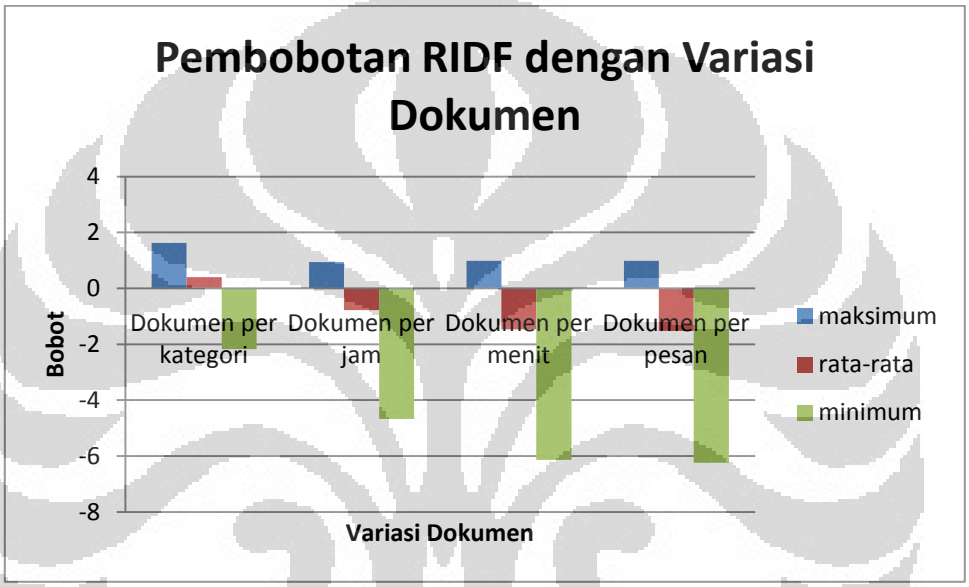
Perhitungan bobot dari suatu term dalam sebuah dokumen dengan menggunakan TF-IDF menunjukkan bahwa deskripsi terbaik dari suatu dokumen tersebut adalah term yang banyak muncul dan sangat sedikit muncul pada dokumen yang lain. Demikian juga sebuah term yang muncul dalam jumlah yang sedang dalam proporsi yang cukup dalam dokumen di koleksi yang diberikan juga akan menjadi deskripsi yang baik mengenai dokumen tersebut. Bobot terendah akan diberikan pada term yang muncul sangat jarang pada beberapa dokumen *(low-frequency document)* dan term yang muncul pada hampir atau seluruh dokumen *(high-frequency document).*

Pada Gambar 4.8 didapatkan hasil dengan nilai maksimum yang paling tinggi pada dokumen per pesan. Namun secara keseluruhan nilai yang dihasilkan baik itu nilai maksimum, minimum ataupun rata-rata tidak signifikan berbeda jika dibandingkan antar klasifikasi dokumen. Bobot maksimum pada tiap variasi dokumen mempunyai nilai pada kisaran 12. Nilai yang terbilang cukup besar ini karena formula untuk mendapatkan nilai TF-IDF hanya mengalikan (*multiply*) nilai TF dan IDF. Bobot rata-rata pun hanya dokumen per kategori yang berbeda jelas nilai bobotnya diantara variasi dokumen yang lain. Hasil nol yang didapat pada bobot minimum dikarenakan banyaknya atau hampir keseluruhan dokumen dari data set mengandung sebuah kata kunci atau dapat juga dikatakan nilai idf bernilai nol. Hal ini mengakibatkan nilai tf yang dikalikan dengan nilai idf juga akan bernilai nol.

Selanjutnya hasil pengukuran dengan pembobotan *Residual Inverse Document Frequency* (RIDF) dapat dilihat pada Tabel 4.4 dan Gambar 4.9.

Tabel 4.4 Hasil Pengukuran RIDF

RIDF	Dokumen per kategori	Dokumen per jam	Dokumen per menit	Dokumen per pesan
Bobot maksimum	1,63	0,94	0,98	0,98
Bobot rata-rata	0,4	-0,78	-1,47	-1,52
Bobot minimum	-2,17	-4,67	-6,14	-6,24



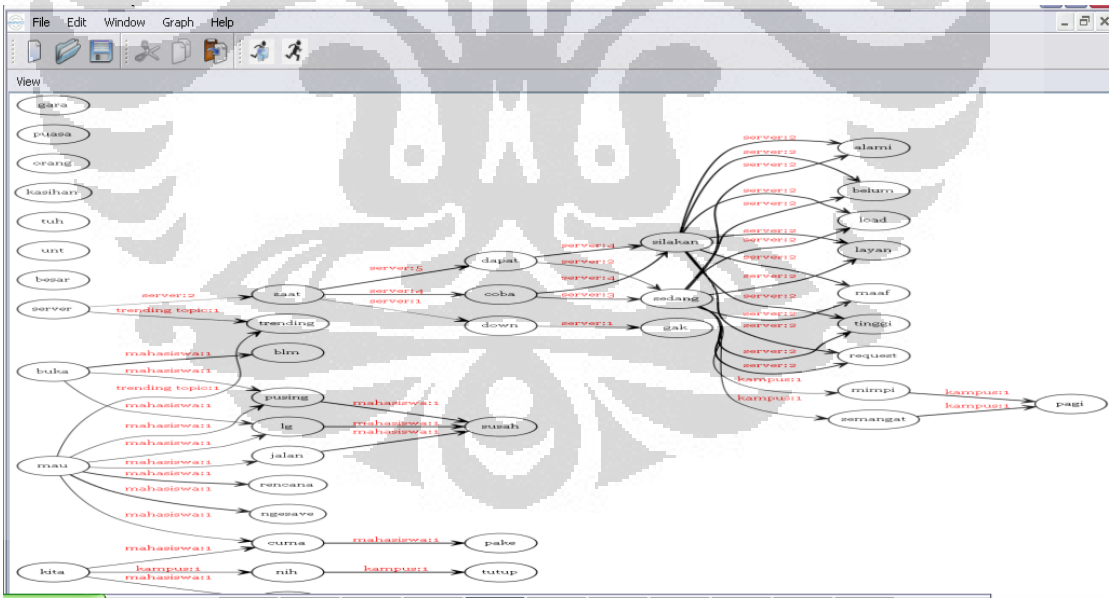
Gambar 4.9 Pembobotan RIDF dengan variasi dokumen

Residual Inverse Document Frequency (RIDF) merupakan varian dari IDF dan memiliki efektifitas yang tinggi dalam bidang *Information Retrieval (IR).* Kelebihan yang dimiliki RIDF dibandingkan IDF adalah term yang jarang muncul pada dokumen tidak akan dijadikan relevansi. Pada Tabel 4.4 dan Gambar 4.9 didapatkan nilai yang kurang baik pada dokumen yang dikelompokkan berdasarkan pesan acak. Hal tersebut diakibatkan pesan yang dikumpulkan menjadi sebuah dokumen diambil secara acak sehingga kecil kemungkinan memiliki keterkaitan kata. Sedangkan pengelompokkan dokumen berdasarkan jam ataupun menit memberikan hasil yang lebih baik dibanding dokumen berdasarkan pesan Hasil yang sangat baik didapat dari pengukuran dokumen yang dikelompokkan berdasar pada kategori. Nilai bobot yang

bernilai negatif dikarenakan penentuan threshold di program ‘coba.rb’ ditetapkan bernilai -100. Dengan demikian kata kunci yang bernilai negatif juga ikut diloloskan dalam penyaringan kata kunci. Hal ini mempengaruhi nilai minimum dan rataan pada perhitungan bobot menggunakan metode RIDF. Pada penentuan variasi dokumen yang akan dijadikan acuan sebagai proses visualisasi *graph* mana yang akan dipilih dalam langkah penentuan *node*. Hal ini benar secara logika karena pengelompokkan *tweet* sesuai kategori menandakan bahwa antara satu *tweet* dengan *tweet* lainnya memiliki term yang berkaitan ataupun memiliki kata yang sama. Dengan begitu dapat disimpulkan pendefinisian dokumen terbaik adalah dokumen berdasarkan kategori dengan pembobotan RIDF.

4.2.1 Penentuan *Parent Node* dan *Child Node*

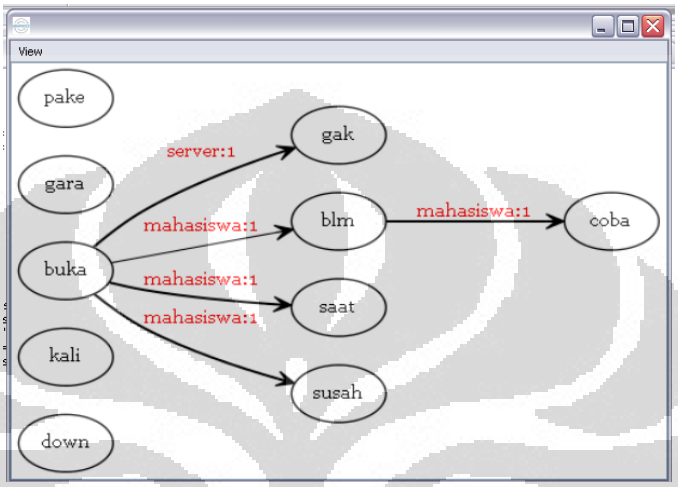
Langkah selanjutnya adalah penentuan kata-kata yang ditetapkan sebagai *parent node* dan *child node.* Gambar 4.10 dan Gambar 4.11 merupakan hasil grafik konsep pengukuran variasi dokumen berdasarkan kategori dengan pembobotan RIDF:



Gambar 4.10 Keluaran proses visualisasi *graph* tanggal 30 Januari 2012

Dari Gambar 4.10 didapatkan delapan padanan kata *parent node* dan *child node*. Padanan ini berdasar pada keterikatan suatu kata dengan kata lain. Hal ini

berkaitan dengan nilai derajat yang berarti semakin tinggi nilai derajat suatu kata maka keterikatan atau relevansi dari kata tersebut semakin besar pula. Suatu kata yang ditetapkan sebagai parent node berarti memiliki nilai derajat yang tinggi pula. Delapan padanan kata yang dimaksud tercantum pada Tabel 4.5.



Gambar 4.11 Keluaran proses visualisasi *graph* tanggal 31 Januari 2012

Dari empat keluaran proses visualisasi *graph*, diketahui bahwa grafik pada tanggal 1 dan 2 Februari 2012 tidak terdapat simpul yang menghubungkan antara satu kata dengan kata lainnya. Hal berbeda ditunjukkan oleh grafik konsep pada tanggal 30 dan 31 Januari 2012 yang terdapat simpul penghubung antar kata. Sebab dari hal ini karena pada tanggal 30 dan 31 Januari 2012 masih banyak *tweet* yang membicarakan masalah SIAK-NG. Namun seiring berjalannya waktu *tweet* dengan topik SIAK-NG semakin berkurang karena mahasiswa telah dapat mengakses situs akademis milik Universitas Indonesia. Penurunan jumlah yang mengakses SIAK-NG memiliki dampak yang baik karena sistem hanya melayani request yang lebih minim disbanding hari pertama masa registrasi akademis.

Dengan Gambar 4.10 dan 4.11 maka dapat ditetapkan kata-kata yang menjadi *parent node* dan *child node*, yang dapat dilihat pada Tabel 4.5.

Tabel 4.5 Penentuan *node* data grafik konsep.

Tanggal posting	Child Node	Parent Node	Edge Label
31-Jan-12	coba	buka	mahasiswa
30-Jan-12	susah	mau	mahasiswa
30-Jan-12	susah	buka	mahasiswa
30-Jan-12	Pagi	server	kampus
30-Jan-12	request	server	server
30-Jan-12	Load	server	server
30-Jan-12	tinggi	server	server
30-Jan-12	Layan	server	server
30-Jan-12	Maaf	server	server

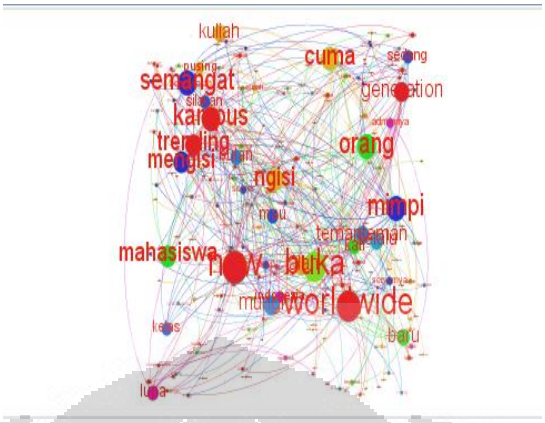
Dari Tabel 4.5 terdapat sembilan padanan kata *parent node* dan *child node* dari dua tanggal *posting. Edge label* merupakan kategori yang mempunyai korelasi antara 2 *node.* Tabel 4.5 di atas nantinya juga akan menjadi parameter pembanding saat hasil deteksi rumor telah diketahui.

4.2.2 Deteksi Rumor

Terdapat dua tahap yang diakukan dalam deteksi rumor, pertama adalah menganalisa grafik dan yang kedua adalah menganalisa dengan metode *edit distance*.

4.2.2.1 Analisa Grafik

Dalam analisa grafik yang perlu diperhatikan adalah nilai kepusatan. Nilai kepusatan merupakan satu parameter yang menunjukkan penting tidaknya sebuah simpul (node) pada data set yang sedang dianalisa. Nilai kepusatan yang digunakan dalam percobaan ini adalah nilai derajat (menunjukkan nilai popularitas simpul), dan keantaran kepusatan (menunjukkan nilai sumber).



Gambar 4.12. *Graph data set* Trending Topic SIAK-NG

Pada Gambar 4.12 besar simpul ditentukan oleh nilai derajat. Semakin tinggi nilai derajat maka semakin besar ukuran simpul. Nilai derajat diukur dari banyaknya ikatan yang dimiliki simpul. Karena itu simpul yang memiliki nilai derajat tinggi memiliki nilai popularitas tinggi dan sering muncul dalam setiap pesan dalam jangka waktu penelitian. Jika dilihat pada grafik di atas dapat dikatakan bahwa kata-kata seperti mahasiswa, *worldwide*, *trending*, kampus, mengisi, orang, dan sebagainya memiliki niai derajat yang tinggi.

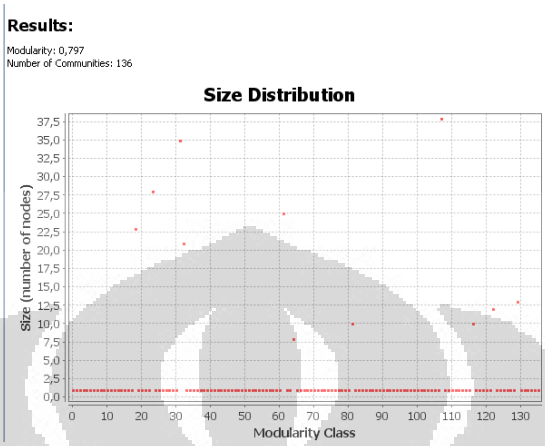
Lalu dengan menganalisa nilai keantaran kepusatan dapat diketahui kata kunci yang merupakan sumber dari persebaran tweet pada Twitter. Dengan menganalisa grafik secara kesuluruhan dapat diketahui kata kunci seperti semangat, *worldwide*, *now,* buka, kampus, mimpi merupakan kata kunci yang memiliki niai keantaran kepusatan tinggi.

Jika ditelaah lebih jauh, kata kunci yang memiliki nilai derajat yang tinggi ternyata memiliki nilai keantaran kepusatan yang tinggi pula. Hal ini menunjukkan semakin penting suatu kata kunci dari suatu data set maka semakin tinggi pengaruhnya terhadap grafik.

4.2.2.2 Analisa dengan Metode *Edit Distance*

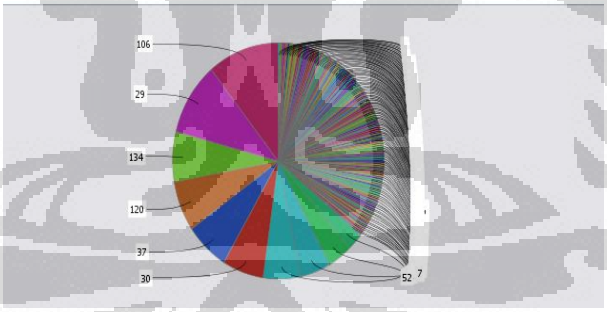
Selanjutnya adalah analisa dengan metode *edit distance*. Metode *edit distance* adalah suatu cara menganalisa grafik dengan membandingkan dengan sumber yang terpercaya. Sumber yang dijadikan acuan dalam menentukan hasil pada tugas akhir

ini adalah akun @siakng_ui. Sebagai langkah awal dilakukan analisa sebuah grafik modularitas



Gambar 4.13. Modularitas terhadap jumlah simpul

Dari Gambar 4.13 didapatkan sebelas kelas modularitas yang nilai jumlah simpulnya tinggi. Kelas modularitas merupakan label yang otomatis melekat pada kata kunci yang memiliki relevansi. Kelas modularitas yang dimaksud diantaranya



Gambar 4.14. Persentase kelas modularitas

Dari grafik persentase kelas modularitas pada Gambar 4.14 di atas ditunjukkan kelas modularitas mana saja yang banyak muncul pada data set kasus *trending topic* SIAK-NG. Pada tugas akhir ini hanya lima dari sebelas kelas modularitas yang digunakan untuk selanjutnya dibandingkan dengan Tabel 4.5. Secara rinci kelas modularitas yang sering muncul disebutkan pada Tabel 4.6.

Tabel 4.6. Daftar Kata Kunci tiap Kelas Modularitas

Kelas modularitas	Kata kunci
126	allah, benar, tahu, pke, trending, university, nambah, perjuangan, now, topic, diijabah, ikhtiar, tetaplah, sulit, biasa, keras, sponsorin, pantes, generation, null, nambah, class, biasa, pantes, and, world, worldwide, mahasiswa, kampus, inilah, akademis, univindonesia.
29	maghrib, jalan, siakan, hati, bersama, kuliah, lintas, highend, rayap, masuk, lalu, rencana, ngesave, gimana, susah, coba, laptop, tadi, daftar, kata, padat, buka, akses, gagal, komputer,
134	penuh, diawal, awal, bem, fasilkom, tahun, disimpan, nasibnya, masalah, isu, akhir, manis, kasihan, selalu, bermasalah, habis, mengisi, dibuang, dikaji, dilaporkan, harap,
120	sedang, maba, berburu, mengalami, load, tinggi, cepatan, tutup, lokal, servernya, request, mau, maaf, silakan, server, down, mencoba, melayani
37	tiket, layak, kaya, malam, dikira, mahasiswa, baru, ngerti, secara, fakultas, login, orang, saingannya, aman, bersamaan

Jika dianalisa dengan membandingkan kata kunci dari tiap kelas modularitas dengan padanan kata kunci dari *child node* dan *parent node* maka didapatkan bahwa kelas modularitas 120 merupakan kelas modularitas dengan tingkat kebenaran yang tinggi. Adapun tweet yang berisi kata kunci dari kelas modularitas 120 adalah "Maba pada tutup SIAKNG deh, MK lo kan udah di paket. Santai aja sih. Nih kita pada berburu kelas belanja! Udah cepetan tutup!" dan *tweet* "Maaf, server SIAK-NG sedang mengalami load tinggi dan belum dapat melayani request anda saat ini. Silakan mencoba beberapa saat lagi.". Dengan membandingkan dengan informasi

(tweet) dari akun SIAK-NG (@siakng_ui) maka dapat disimpulkan bahwa tweet yang memiliki nilai kebenaran tertinggi atau dapat juga dikatakan *tweet* non rumor adalah "Maaf, server SIAK-NG sedang mengalami load tinggi dan belum dapat melayani request anda saat ini. Silakan mencoba beberapa saat lagi.".

BAB 5
KESIMPULAN

Dari perancangan dan pengujian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut :

1. Dalam upaya untuk menganalisa rumor pada media sosial dibutuhkan cara merepresentasikan informasi rumor dalam bentuk grafik
2. Sistem dalam tugas akhir ini terdiri dari 3 buah subsistem yaitu penggalian informasi (crawling), pengolahan data, dan visualisasi.
3. Dari hasil pengujian diketahui bahwa pembobotan dengan RIDF variasi dokumen berdasarkan kategori mempunyai nilai yang paling baik dengan nilai maksimum 1,63, nilai rata-rata 0,4 dan nilai minimum -2,17.
4. Dari grafik konsep didapatkan Sembilan padanan kata antara *Parent Node* dan *Child Node* serta tiga kategori *edge label*.
5. Dari hasil penelitian didapatkan lima kelas modularitas tertinggi yang masing-masingnya berisi kata kunci untuk kemudian dicocokkan dengan sumber informasi terpercaya. Kelas modularitas 120 yang mengatakan bahwa "Maaf, server SIAK-NG sedang mengalami load tinggi dan belum dapat melayani request anda saat ini. Silakan mencoba beberapa saat lagi.". memiliki nilai kebenaran paling tinggi.
6. Pertimbangan lebih lanjut diperlukan dalam menganalisa sebuah rumor. Data set dan topik penelitian harus diperluas sehingga manfaat dari deteksi rumor dapat lebih meyakinkan.

DAFTAR ACUAN

[1] *Jejaring Sosial.* Diakses pada tanggal 3 Maret 2012, dari Drise. http://drise-online.com/tafakoor/44-jejaring-sosial.html

[2] *Profession Ethics.* Diakses pada tanggal 3Maret 2012, dari Oppapers. http://www.oppapers.com/essays/Profession-Ethics/666484

[3] Feldman, Ronen., Sanger, James., " The Text Mining Handbook",Cambridge University Press, 2007.

[4] Qaazvinian, Vahed., Rosengren, Emily., Radev, Dragomir., Mei, Qiaozhu., " Rumor has it : Identifying Misinformation in Microblogs", University of Michigan, 2011.

[5] *Awal Semester, SIAK-NG Jadi Tren Percakapan Linimasa.* Diakses pada tanggal 5 Maret 2012, dari Saling Silang. http://salingsilang.com/baca/awal-semester-siak-ng-jadi-tren-percakapan-linimasa

[6] *Top 10 Websites To Search Old Tweets.* Diakses Pada Tanggal 5 Maret 2012, dari Freenuts. http://freenuts.com/top-10-websites-to-search-old-tweets/

[7] Feldman, Ronen., Sanger, James., " The Text Mining Handbook",Cambridge University Press, 2007.

[8] Mooney, Raymond., " CS 391L: Machine Learning Text Categorization", University of Texas, 2006.

[9] *The Open Graph Viz Platform.* Diakses Pada Tanggal 3 Maret 2012, dari Gephi. http://gephi.org/

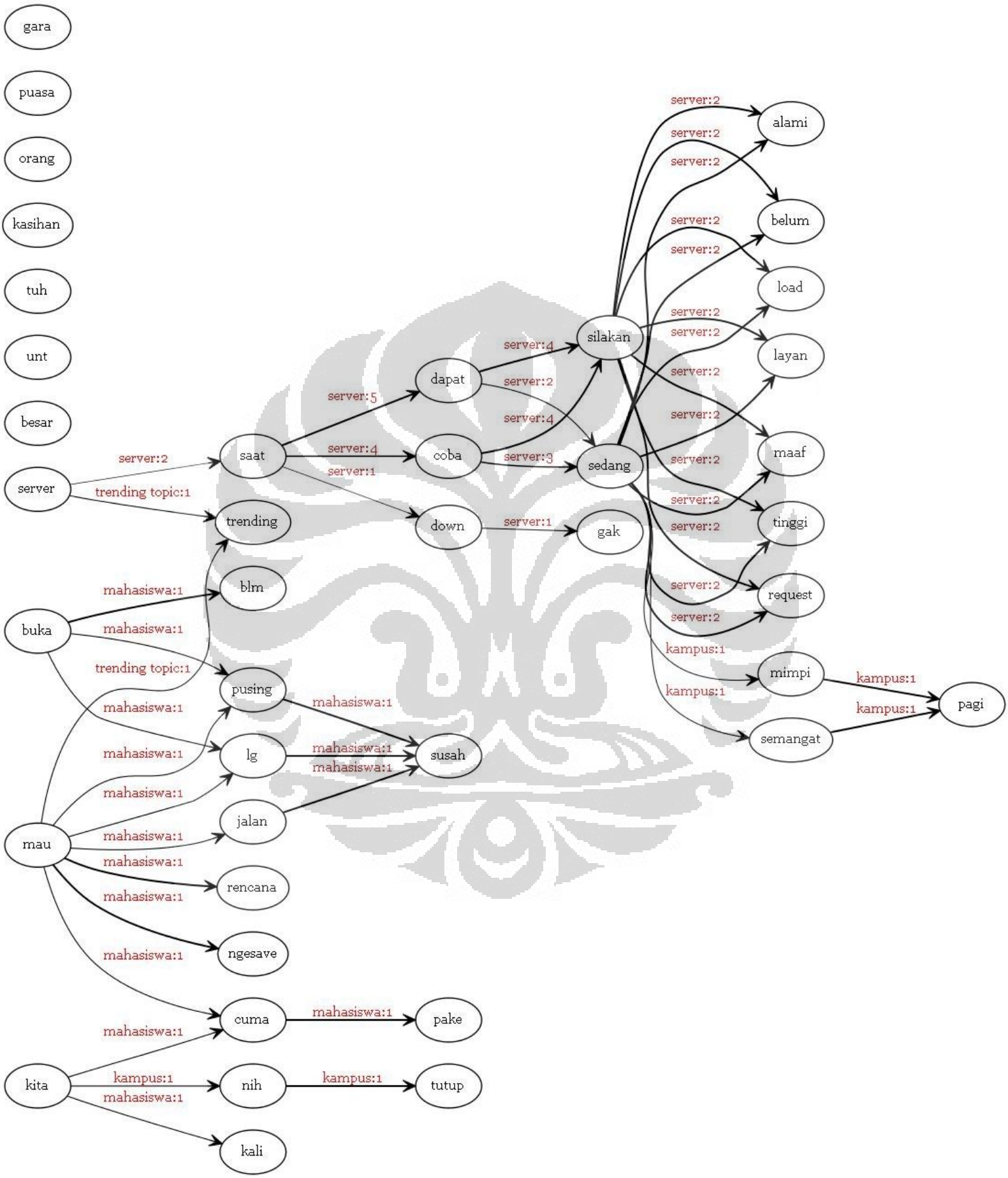
[10] Bunke, H., On a relation between graph edit distance and maximum common subgraph, Pattern Recognition Letters, Volume 18, Issue 8, Agustus 1997, pp. 698-694, 1997.

[11] Hashimoto, Takako., Kuboyama, Tetsuji., dan Shirota, Yukari. 2011.*Rumor Analysis Frame in Social Media.* Indonesia: 2011 IEEE Region 10 Conference, 21-24 November 2011.

[12] Tentang Ruby. Diakses Pada Tanggal 20 May 2012, dari Ruby-lang. http://www.ruby-lang.org/id/about/

[13] *GVedit Help.* Diakses Pada Tanggal 20 May 2012, dari Home Final. http://home.fnal.gov/~stoughto/build/graphviz-2.22.2/windows/cmd/gvedit/GVedit.html

[14] Iino, Y., Hirokawa, S., "Time Series Analysis System of R&D Team Using Patent Information, Lecture Notes in Computer Science, 2009, Volume 5712/2009, pp464-471, 2009


gara
puasa
orang
kasihan
tuh
unt
besar
server
buka
mau
kita
saat
coba
dapat
down
gak
silakan
sedang
server:2
trending topic:1
server:5
server:4
server:1
server:3
server:4
server:2
server:1
trending
blm
pusing
lg
jalan
rencana
ngesave
cuma
pake
nih
tutup
kali
susah
mahasiswa:1
kampus:1
alami
belum
load
layan
maaf
tinggi
request
mimpi
semangat
pagi