ISSN : 2085-3688; e-ISSN : 2460-0997

Penggunaan *Web Crawler* Untuk Menghimpun *Tweets* dengan Metode *Pre-Processing Text Mining*

Bayu Rima Aditya
Program Studi Manajemen Informatika, Fakultas Ilmu Terapan, Universitas Telkom
Jl Telekomunikasi No 1, Terusan Buah Batu Bandung
bayu@tass.telkomuniversity.ac.id

Abstrak – Saat ini jumlah data di media sosial sudah terbilang sangat besar, namun jumlah data tersebut masih belum banyak dimanfaatkan atau diolah untuk menjadi sesuatu yang bernilai guna, salah satunya adalah *tweets* pada media sosial twitter. Paper ini menguraikan hasil penggunaan *engine web crawel* menggunakan metode *pre-processing text mining*. Penggunaan *engine web crawel* itu sendiri bertujuan untuk menghimpun *tweets* melalui API twitter sebagai data teks tidak terstruktur yang kemudian direpresentasikan kembali kedalam bentuk web. Sedangkan penggunaan metode *pre-processing* bertujuan untuk menyaring *tweets* melalui tiga tahap, yaitu *cleansing*, *case folding*, dan *parsing*. Aplikasi yang dirancang pada penelitian ini menggunakan metode pengembangan perangkat lunak yaitu model waterfall dan diimplementasikan dengan bahasa pemrograman PHP. Sedangkan untuk pengujiannya menggunakan *black box testing* untuk memeriksa apakah hasil perancangan sudah dapat berjalan sesuai dengan harapan atau belum. Hasil dari penelitian ini adalah berupa aplikasi yang dapat mengubah *tweets* yang telah dihimpun menjadi data yang siap diolah lebih lanjut sesuai dengan kebutuhan *user* berdasarkan kata kunci dan tanggal pencarian. Hal ini dilakukan karena dari beberapa penelitian terkait terlihat bahwa data pada media sosial khususnya twitter saat ini menjadi tujuan perusahaan atau instansi untuk memahami opini masyarakat.

Kata kunci – API twitter; *cleansing, case folding, parsing, waterfall, black box testing.*

Abstract - Amount of data the exponential growth we have seen in social-media, but it is not widely used to be something of value, one of which is tweets on twitter. This paper describes the result of using engine web crawler with pre-processing text mining method. The use of engine web crawler aims to take tweets via API Twitter as unstructured text data and then represented into web form. The use of pre-processing method aims to filter out tweets on three stages: cleansing, case folding and parsing. Applications designed in this research using the waterfall model and implemented with PHP. Testing method in this research using the black box testing to check whether the design result can already be run in accordance between expectations or not. Result from this research are in the form of applications that can help prepare tweets into data that can be processed according to needs based on keywords and the search date desire by the user. This is done because of several related research shown that data on social media, especially Twitter is currently the destination company or agency to understand public opinion.

Keywords - API twitter; cleansing, case folding, parsing, waterfall, black box testing

I. PENDAHULUAN

Analisis terhadap media sosial adalah alat yang ampuh untuk memahami sikap, preferensi dan opini masyarakat. Bagi suatu perusahaan, analisis media sosial dapat membantu perusahaan untuk membuat keputusan mengenai kebutuhan, sikap, pendapat atau trend tentang pelanggan atau calon pelanggan potensial. Tidak dapat dipungkiri, saat ini jumlah data di media sosial sudah terbilang sangat besar, namun jumlah data tersebut masih belum banyak dimanfaatkan atau diolah untuk menjadi sesuatu yang bernilai guna, salah satunya adalah pada salah satu media sosial terbesar di dunia yaitu twitter. Twitter menyediakan *Application Program Interface* (API) yang memungkinkan kita mendapatkan data mereka. Data yang dapat dimanfaatkan dari media sosial twitter itu sendiri adalah berupa status atau yang lebih dikenal dengan tweets.

Untuk mengubah *tweets* menjadi sebuah data yang memilki makna, maka diperlukan paling tidak dua proses awal, yaitu proses pengambilan *tweets* dari API Twitter dan proses penyaringan *tweets* agar dapat diolah menjadi data yang memiliki makna. Proses pengambilan *tweets* dari API twitter dapat dilakukan dengan menggunakan *engine web crawel* yang dapat merepresentasikan kembali data kedalam bentuk web. Untuk proses penyaringan *tweets* dapat menggunakan metode *pre-processing text mining* yang terdiri dari tiga tahap, yaitu *cleansing, case folding*, dan *parsing*. Untuk itu, dibutuhkan suatu aplikasi yang dapat membantu menghimpun data tweets dan membantu mempersiapkan data untuk diolah menjadi informasi.

Paper ini menguraikan hasil penggunaan *web crawel* dalam menghimpun *tweets* dengan metode *pre-processing text mining*. Pembuatan *web crawel* bertujuan untuk mengambil *tweets* melalui API twitter sebagai data teks yang tidak terstruktur. Sedangkan penggunaan metode *pre-processing* bertujuan untuk mempersiapkan data teks menjadi data yang dapat diolah lebih lanjut atau sesuai kebutuhan *user*. Untuk pengujian dari aplikasi pada penelitian ini menggunakan *black box testing* untuk memeriksa apakah aplikasi sudah dapat berjalan sesuai dengan harapan atau belum.

Kontribusi yang diberikan penelitian ini adalah membantu mempersiapkan *tweets* menjadi data yang dapat diolah lebih lanjut sesuai dengan kebutuhan. Hal ini dilakukan karena dari beberapa penelitian terkait terlihat bahwa data pada media sosial khususnya twitter saat ini menjadi tujuan perusahaan atau instansi untuk memahami opini masyarakat.

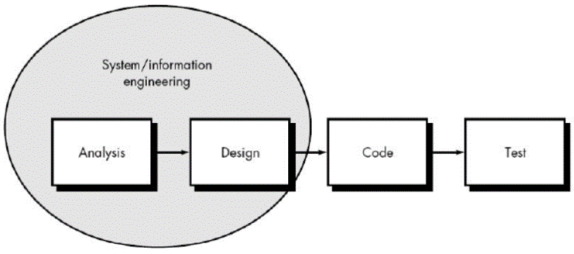
Pembahasan dalam paper ini akan mengikut alur sebagai berikut. Di bagian II diuraikan metode penelitian termasuk di dalamnya uraian mengenai penelitian-penelitian lain yang sudah dilakukan terkait *web crawler* dan twitter. Selanjutnya dibagian III, diuraikan hasil pengujian dan pembahasan. Kesimpulan dari penelitian ini akan diuraikan di bagian IV.

II. METODOLOGI PENELITIAN

A. *Metode Penelitian*

Metode-metode yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Studi pustaka, yaitu dilakukan dengan cara membaca berbagai referensi buku, literature, dan bacaan lainnya yang berhubungan dengan permasalahan yang sedang diteliti.
2. Metodologi pengembangan sistem yang digunakan pada penelitian ini adalah linear sequential model [7] seperti terlihat pada Gambar 1.



Gambar 1. Model Sekuensial Linier [6]

Model ini menyarankan pendekatan pengembangan secara sekuen dan sistematik untuk pengembangan sistem [7]. Adapun perangkat keras dan perangkat lunak yang dibutuhkan oleh dalam pengembangan sistem ini adalah seperti terlihat pada Tabel 1 dan Tabel 2.

Tabel 1. Kebutuhan Perangkat Keras

No	Perangkat Keras	Spesifikasi
1	Processor	Intel Core i3
2	RAM	2 GB
3	Harddisk	500 GB

Tabel 2. Kebutuhan Perangkat Lunak

No	Perangkat Lunak	Spesifikasi
1	Web *browser* dan *debugging tool*	Google Chrome
2	*Script Editor*	Macromedia Dreamweaver CS5

3. Metode yang digunakan dalam proses penghimpunan *tweets* pada penelitian ini adalah *preprocessing text mining*. Tahapan yang dilakukan dari proses *pre-processing text mining* itu sendiri adalah sebagai berikut [5].
   a. *Cleansing*, yaitu proses membersihkan tweets dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, hashtag (#), username (@username), url (http://situs.com), dan email (nama@situs.com).
   b. *Case folding*, yaitu mengubah semua huruf menjadi lowercase atau huruf kecil.
   c. *Parsing*, yaitu proses memecah tweets menjadi sebuah kata. Hal ini sesuai dengan fitur digunakan yaitu *unigram*.

B. *Penelitian Terkait*

1. Dewi Rosmala dan Rizkqi Rivani Syafei dalam penelitiannya berjudul "implementasi *web crawler* pada *social media monitoring*" menyimpulkan hasil penelitiannya bahwa, implementasi *web crawler* pada aplikasi *social media monitoring* dapat memudahkan user memantau issue-issue yang sedang terjadi terhadap sebuah brand image dari sebuah produk. Dengan adanya aplikasi ini diharapkan mampu membantu pengguna juga untuk menjada brand image. Dan jika dikembangkan aplikasi ini tidak hanya dapat memantau saja, melainkan diharapkan dapat menjadi sumber *key learning* bagaimana menciptakan strategi promosi yang sukses [8].
2. Yohanes Sigit, dkk dalam penelitiannya berjudul Analisis dan Perancangan Alat Bantu Monitor Brand Universitas Atma Jaya Yogyakarta di Situs Jejaring Sosial Twitter menyimpulkan bahwa untuk menganalisis *brand monitoring* universitas atma jaya Yogyakarta dapat dilakukan dengan

menggunakan sebuah perangkat lunak sebagai alat bantu untuk mengcapture data-data dari situs jejaring *social* twitter. Dan jika dikembangkan lebih lanjut, alat bantu ini dapat bermanfaat untuk mengetahui bagaimana pendapat masyaratak (pengguna twitter) mengenai brand Universitas Atma Jaya Yogyakarta untuk melakukan pengembangan Universitas menjadi lebih baik [9].

3. Pipit Pitria dalam penelitiannya berjudul Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naive Bayes menyimpulkan hasil penelitiannya bahwa pemanfaatan analisis sentimen biasanya digunakan untuk mengevaluasi sebuah produk yang sentimen-sentimennya didapat dari *feedback* sebuah produk itu sendiri melalui social media khususnya twitter. Studi kasus yang digunakan dalam penelitian ini sebagai uji coba implementasi adalah akun twitter resmi Samsung Indonesia yang aktif dalam mempromosikan produk-produknya [6].

C. *Tinjauan Pustaka*

1. *Preprocessing Text Mining*

   Menurut definisi, *Text Mining* adalah proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut [1]. *Text mining* mengekstrak informasi berguna dari sumber data melalui identifikasi dan eksplorasi yang tidak dalam bentuk *database record*, melainkan dalam data teks yang tidak terstruktur. *Preprocessing* adalah tahap proses awal *text mining* terhadap teks untuk mempersiapkan teks menjadi data yang dapat diolah lebih lanjut. Sekumpulan karakter yang bersambungan (teks) harus dipecah-pecah menjadi unsur yang lebih berarti. Suatu dokumen dapat dipecah menjadi bab, sub-bab, paragraf, kalimat, kata dan bahkan suku kata [3].

2. Status Twitter (*tweets*)

   Tweets adalah pesan yang ditulis oleh pengguna media sosial twitter. *Tweets* itu sendiri berupa teks tulisan yang dapat memuat hingga 140 karakter yang ditampilkan pada halaman profil *user* [10].

3. *Web Crawler*

   Web crawler adalah sebuah perangkat lunak yang diguankan untuk menjelajah serta mengumpulkan halaman-halaman web yang selanjutnya diindeks oleh mesin pencari [4]. Sedangkan proses c*rawling* adalah proses yang digunakan oleh mesin pencari (*search engine*) untuk mengumpulkan halaman *website* [2].

III. HASIL DAN PEMBAHASAN PENELITIAN

A. *Analisis Sistem*

1. Analisis Sistem Berjalan

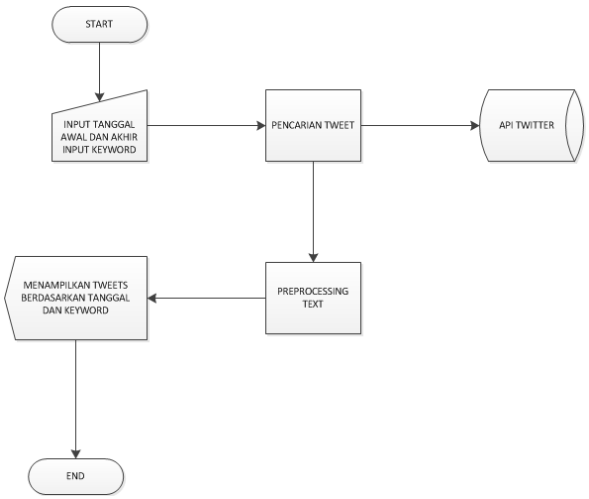
   Masyarakat Indosnesia sudah tidak asing lagi dengan fenomena penggunaan twitter sebagai media untuk mengungkapkan ide, komentar, opini atau pernyataan terhadap suatu hal. Setiap saat orang-orang menuliskan apa yang ada dalam pikirannya untuk kemudian dishare ke twitter. Selain itu orang-orang pun dapat mengomentari tweet yang ditulis orang lain. Akan tetapi *tweets* yang ada saat ini hanya berupa *tweets* biasa dan belum banyak dihimpun untuk dijadikan suatu data atau informasi.

2. Sistem yang diusulkan

   Dari permasalahan diatas, diperlukan sebuah solusi yaitu dengan mengumpulkan *tweets* dari API twitter. Sistem akan melakukan proses pencarian *tweets* berdasarkan kebutuhan yang kemudian diproses melalui metode *pre-processing text mining*. Hasilny akan direpresentasikan dalam bentuk sebuah web.

B. *Proses Penelusuran Tweets*

Proses penelusuran *tweets* merupakan proses yang perlu dilakukan untuk mengubah *tweets* menjadi sebuah data yang memiliki makna. Proses penelusuruan tweets dilakukan dengan cara pengambilan tweets dari API Twitter, kemudian dilakukan proses penyaringan tweets agar dapat diolah menjadi data yang memiliki makna. Proses penelusuran *tweets* yang diusulkan pada penelitian ini secara lebih detail dapat dilihat pada Gambar 2.



Gambar 2. Usulan Proses Penelusuran *Tweets*

Berdasarkan Gambar 2, maka dapat dideskripsikan bahwa tahap-tahap pada proses penelusuran *tweets* pada aplikasi ini yaitu.

1. *User* menginputkan tanggal awal dan akhir dan *keyword* sesuai yang diinginkan *user*.
2. Sistem akan melakukan proses pencarian *tweets* dengan cara mengambil *tweets* dari API Twitter.
3. Menggunakan *twitteroauth* agar dizinkan mengambil data dari API Twitter berupa data start_date , end_date dan q.
4. Data akan diproses melalui tahap *preprocessing tweet* yaitu proses memecah data *tweets* menjadi kata per kata dan diubah menjadi huruf kecil
5. Data yang diinginkan *user* akan di tampilkan kedalam bentuk web.

C. *Diagram Alur Preprocessing*

Pada penelitian ini, untuk proses penyaringan *tweets* menggunakan metode *pre-processing text mining*. Diagram alur untuk proses *pre-processing* dapat dilihat pada Gambar 3.

Gambar 3. Proses *Pre-Processing Text*

Berdasarkan Gambar 3, maka dapat dideskripsikan bahwa tahap-tahap pada *preprocessing* pada aplikasi ini yaitu.

1. Data berupa *tweets* yang telah didapatkan sebelumnya dari *database* akan melalui tahap *parsing* atau tiap *tweets* akan dipecah kedalam satu suku kata.
2. *Tweets* yang telah dipecah menjadi kata tunggal (unigram) akan melewati tahap *lowercase* atau diubah menjadi huruf kecil. Hal ini dilakukan karena untuk memudahkan tahap selanjutnya.
3. Setelah melewati proses diatas, aplikasi akan menampilkan hasil berupa *tweets* terurut berdasarkan kata kunci dan tanggal.

D. *Implementasi Sistem*

Berikut ini merupakan implementasi antarmuka *user* dari aplikasi yang dibangun.

1. Tampilan halaman untuk proses memasukkan *keyword* pencarian dapat dilihat pada Gambar 4.

Gambar 4. Proses Input *Keyword*

Pada proses ini *user* akan menginputkan *keyword* lalu *keyword* tersebut di *parsing* ke API Twitter melalui nilai q. Twitter akan memberikan data *keyword* yang diminta melalui nilai q yang diinputkan di php.

2. Tampilan halaman untuk proses memasukkan tanggal dapat dilihat pada Gambar 5.

Gambar 5. Proses Input Tanggal

Pada proses ini *user* akan menginputkan tanggal awal dan tanggal akhir. Aplikasi akan meminta data dari API twitter melalui nilai "*Since*" dan "*Until*". API twitter akan memberikan data "*Since*" dan "*Until*" dengan syarat rentang waktu tidak lebih dari 9 hari dari tanggal saat menggunakan aplikasi.

3. Tampilan data hasil penelusuran *tweets* dan proses *preprocessing* dapat dilihat pada Gambar 6.

Gambar 6. Tampilan Hasil Penelusuran *Tweets*

Pada proses ini akan ditampilkan *tweets* yang sesuai dengan yang diinputkan user. Data yang diambil berdasarkan data "Since","Until" dan " q" dari API twitter.

E. *Pengujian Sistem*

Hasil pengujian dari aplikasi pada penelitian ini dapat dilihat pada Tabel 3.

Tabel 3. *Black Box Testing*

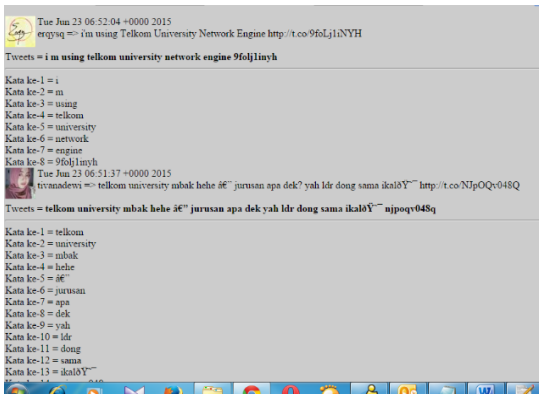
Nama Field	Tipe Masukan	*Output* yang diharapkan	*Output*	Kesimpulan
Keyword	*Keyword* yang diinginkan *user*	Menampilkan *tweets* sesuai dengan *keyword* yang diinputkan *user*	Menampilkan *tweets* sesuai dengan *keyword* yang diinputkan *user*	OK
Start Date	Kosong	Menampilkan *tweets* sesuai dengan tanggal akhir	Menampilkan *tweets* sesuai dengan tanggal akhir	OK
	Tanggal yang diinginkan *user* , dengan format YYYY-MM-DD	Menampilkan tanggal dengan sesuai dengan pilihan *user*	Menampilkan tanggal dengan sesuai dengan pilihan *user*	OK
End Date	Kosong	Menampilkan *tweets* sesuai dengan tanggal awal	Menampilkan *tweets* sesuai dengan tanggal awal	OK
	Tanggal yang diinginkan *user* , dengan format YYYY-MM-DD	Menampilkan tanggal dengan sesuai dengan pilihan *user*	Menampilkan tanggal dengan sesuai dengan pilihan *user*	OK
Start Date & End Date	Kosong	Menampilkan *tweets* sampai dengan tanggal hari ini	Menampilkan *tweets* sampai dengan tanggal hari ini	OK
	Tanggal yang diinginkan *user* , dengan format YYYY-MM-DD	Menampilkan tanggal dengan sesuai dengan pilihan *user*	Menampilkan tanggal dengan sesuai dengan pilihan *user*	OK
Submit data	-	Menampilkan *tweets* berdasarkan kriteria *filter* yang dibuat	Menampilkan *tweets* berdasarkan kriteria *filter* yang dibuat	OK

F. Simulasi Kerja Aplikasi

Hasil simulasi kerja pada aplikasi yang telah dibangun adalah seperti berikut.

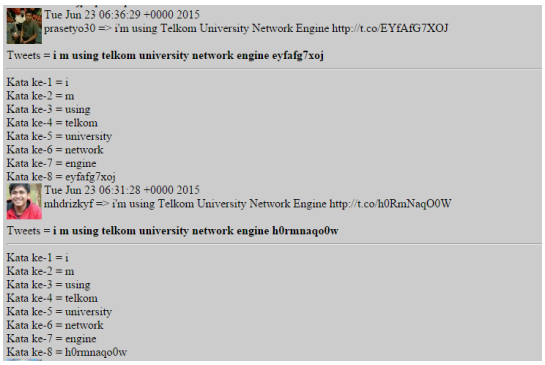
1. Menginputkan *keyword* dan tanggal

   Hasil simulasi contoh *user* menginputkan *keyword* "Telkom University" dan tanggal awal dan akhir 2015-6-23. Maka Aplikasi akan menampilkan *tweets* yang berkaitan tentang "Telkom University" di tanggal tersebut dapat dilihat pada Gambar 7-10.



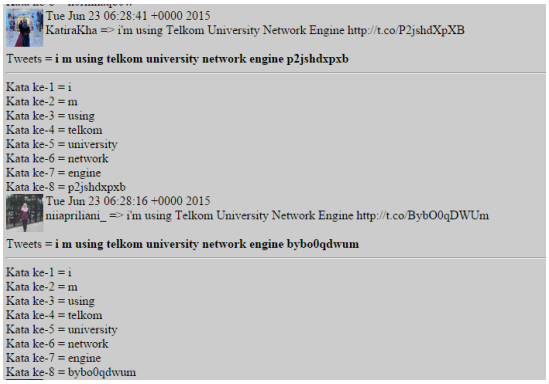
Gambar 7. Tampilan Data Bagian 1

Gambar 7 menunjukkan dua *tweets* dengan posisi pertama dan kedua dari atas tampilan aplikasi beserta hasil *pre-processing* nya. Semakin atas posisi menunjukkan keterbaruan dalam hal waktu.



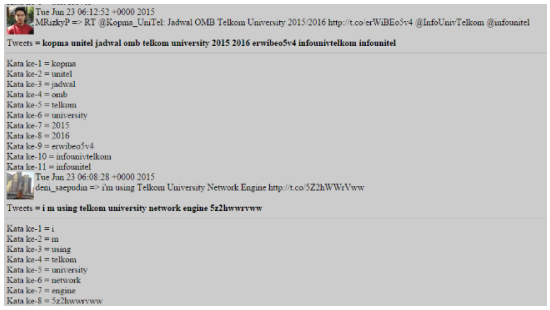
Gambar 8. Tampilan Data Bagian 2

Gambar 8 menunjukkan dua *tweets* dengan posisi ketiga dan keempat dari atas tampilan aplikasi beserta hasil *pre-processing* nya.



Gambar 9. Tampilan Data Bagian 3

Gambar 9 menunjukkan dua *tweets* dengan posisi kelima dan keenam dari atas tampilan aplikasi beserta hasil *pre-processing* nya.



Gambar 10. Tampilan Data Bagian 4

Gambar 10 menunjukkan dua *tweets* dengan posisi ketujuh dan kedelapan dari atas tampilan aplikasi beserta hasil *pre-processing* nya.

2. Menginputkan Tanggal Kosng

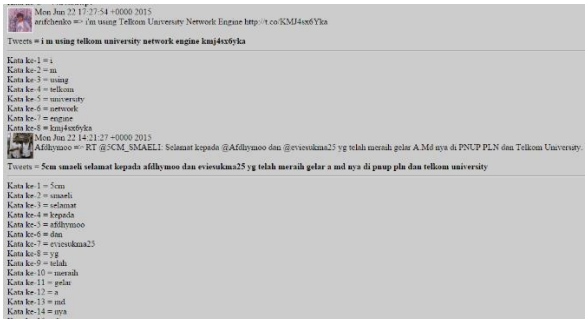
   Hasil simulasi contoh *user* tidak menginputkan tanggal dapat dilihat pada Gambar 11.

Gambar 11. Proses Input Tanggal Kosong

Pada proses ini, jika *user* tidak menginputkan tanggal awal dan tanggal akhir maka aplikasi akan memberikan notifikasi bahwa tanggal awal dan tanggal akhir tidak boleh kosong.

3. Menginputkan tanggal akhir

   Hasil contoh simulasi jika *user* menginputkan tanggal akhir 2015-6-23 tanpa menginput tanggal awal dapat dilihat pada Gambar 12.



Gambar 12. Hasil Simulasi Tanggal Awal Kosong

Gambar 12 menampilkan data *tweets* dari 9 hari yang lalu sampai tanggal 2015-6-23.

4. Menginputkan Tanggal awal

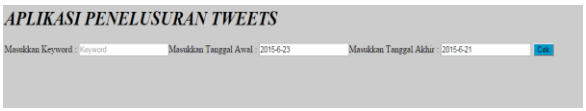
   Hasil simulasi jika *user* menginputkan tanggal awal 2015-6-22 tanpa menginputkan tanggal akhir dapat dilihat pada Gambar 13.

Gambar 13. Hasil Simulasi Tanggal Akhir Kosong

Gambar 13 menampilkan data tweets dari tanggal 2015-6-22 sampai hari ini.

5. Menginputkan tanggal awal lebih besar dari tanggal akhir.

   Hasil simulasi jika tanggal awal diinputkan 2015-6-23 dan tanggal akhir diinputkan 2015-6-21 dapat dilihat pada Gambar 14.



Gambar 14. Contoh Tanggal Awal Lebih Besar

Gambar 14 menunjukkan bahwa jika tanggal awal lebih besar dari tanggal akhir maka aplikasi tidak menampilkan hasil pencarian.

IV. PENUTUP

Kesimpulan dari penelitian ini adalah penggunaan *web crawler* pada aplikasi yang dapat membantu menghimpun data tweets dan membantu mempersiapkan data untuk diolah menjadi informasi telah berhasil dilakukan berdasarkan analisa dan perancangan yang telah dilakukan. Hasil *tweets* yang dihimpun berdasarkan kata kunci dan tanggal pencarian telah dapat direpresentasikan kembali kedalam bentuk web berupa data-data hasil proses *pre-processing*. Penelitian ini masih banyak keterbatasan, sehingga perlu dilakukan beberapa pengembangan lebih lanjut terutama dalam hal kapasitas penyimpanan aplikasi dan juga dalah hal tingkat akurasi dari proses *pre-processing*.

DAFTAR PUSTAKA

[1] A.S, Rosa and Shalahudin, M. 2011. *Modul Pembelajaran Rekayasa Perangkat Lunak (Terstruktur dan Berorientasi Objek)*. Modula, Bandung.

[2] Castillo, C. 2004. *EffectiveWeb Crawling. (p. i)*. Dept. of Science: University of Chile

[3] Feldman, R & Snager, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York

[4] Khanna, Rajiv A dan Kasliwal, Sourabh. 2007. *Designing A Web Crawler*.

[5] Nur, Muhamad Yunus., & Santika, Diaz. D. (2011). *Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine*. Konferensi Nasional Sistem dan Informatika, (Vol. 009). Bali.

[6] Pitria, Pipit. 2014. Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naïve Bayes. Hyperlink: http://elib.unikom.ac.id/index.php/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-pipitpitri-35651 diakses 7 April 2015.

[7] Pressman, S. Roger. 2010. *Software Engineering: A Practitioner's Approach*.

[8] Rosmala, Dewi & Syafei, Rizqia Riyani. 2011. *Implementasi Webcrawler pada Social Media Monitoring*. Hyperlink: http://lib.itenas.ac.id/kti/wp-content/uploads/2013/10/No.-2-Vol.-2-Mei-Agustus-2011-5.pdf diakses 7 april 2015.

[9] Sigit, Yohanes & dkk. 2012. *Analisis dan Perancangan Alat Bantu Monitor Brand Universitas Atma Jaya Yogyakarta di Situs Jejaring Sosial Twitter*. Hyperlink: http://jurnal.uajy.ac.id/jbi/2012/01/11/analisis-dan-perancangan-alat-bantu-monitor-brand-universitas-atma-jaya-yogyakarta-di-situs-jejaring-sosial-twitter/ diakses 8 April 2015.

[10] Twitter. 2014. Tweets | Twitter Developers. Online. Hyperlink:https://dev.twitter.com/docs/platform-objects/tweets) diakses 6 April 2015.