ANALISIS KECENDERUNGAN INFORMASI DENGAN MENGGUNAKAN METODE *TEXT MINING*

(Studi Kasus: Akun *twitter* @detikcom)

SKRIPSI

Oleh:

SYAIFUDIN KARYADI

NIM. 24010212130030

DEPARTEMEN STATISTIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

SEMARANG

2016

ANALISIS KECENDERUNGAN INFORMASI DENGAN MENGGUNAKAN METODE *TEXT MINING*

(Studi Kasus: Akun *twitter* @detikcom)

Oleh:

Syaifudin Karyadi

24010212130030

Tugas Akhir sebagai salah satu syarat untuk memperoleh
gelar Sarjana Sains pada Departemen Statistika

DEPARTEMEN STATISTIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

SEMARANG

2016

HALAMAN PENGESAHAN I

Judul Skripsi : Analisis Kecenderungan Informasi dengan Menggunakan Metode *Text Mining*
(Studi Kasus: Akun *twitter* @detikcom)

Nama : Syaifudin Karyadi

NIM : 24010212130030

Departemen : Statistika

Telah diujikan pada sidang Tugas Akhir dan dinyatakan lulus pada tanggal 16 Agustus 2016

Semarang, 16 Agustus 2016

Mengetahui,	Panitia Penguji Ujian Tugas Akhir
Ketua Departemen Statistika	Ketua,
Fakultas Sains dan Matematika Undip	
Dra. Dwi Ispriyanti, M.Si.	Dra. Suparti, M.Si
NIP. 195709141986032001	NIP. 196509131990032001

HALAMAN PENGESAHAN II

Judul Skripsi : Analisis Kecenderungan Informasi dengan Menggunakan Metode *Text Mining*
(Studi Kasus: Akun *twitter* @detikcom)

Nama : Syaifudin Karyadi

NIM : 24010212130030

Departemen : Statistika

Telah diujikan pada sidang Tugas Akhir dan dinyatakan lulus pada tanggal 16 Agustus 2016

Semarang, 16 Agustus 2016

Dosen Pembimbing I	Dosen Pembimbing II
Hasbi Yasin, S.Si, M.Si	Moch. Abdul Mukid, S.Si, M.Si
NIP. 198212172006041003	NIP. 197808172005011001

KATA PENGANTAR

Puji Syukur penulis ucapkan kehadirat Allah SWT yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan Tugas Akhir dengan judul "Analisis Kecenderungan Informasi dengan Menggunakan Metode *Text Mining*".

Penulis menyadari bahwa dalam penulisan Tugas Akhir ini tidak lepas dari bimbingan dan dukungan yang diberikan beberapa pihak. Oleh karena itu, penulis ingin menyampaikan terima kasih kepada:

1. Ibu Dra. Dwi Ispriyanti, M.Si. sebagai Ketua Departemen Statistika Fakultas Sains dan Matematika Universitas Diponegoro.
2. Bapak Hasbi Yasin, S.Si., M.Si. selaku dosen pembimbing I dan Bapak Moch. Abdul Mukid, S.Si, M.Si. selaku dosen pembimbing II.
3. Bapak dan Ibu dosen Departemen Statistika Fakultas Sains dan Matematika Universitas Diponegoro
4. Semua pihak yang tidak dapat disebutkan satu per satu yang telah membantu penulis dalam penulisan Tugas Akhir ini.

Penulis menyadari bahwa penulisan Tugas Akhir ini masih jauh dari sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun demi kesempurnaan penulisan selanjutnya.

Semarang, 16 Agustus 2016

Penulis

ABSTRAK

Internet merupakan suatu fenomena yang luar biasa. Berawal dari sebuah eksperimen militer di Amerika Serikat, internet telah berkembang menjadi 'kebutuhan' bagi lebih dari puluhan juta orang di seluruh dunia. Jumlah pengguna internet yang besar dan semakin berkembang, telah mewujudkan budaya internet. Salah satu yang berkembang pesat yaitu media sosial *twitter*. *Twitter* merupakan layanan *microblogging* yang menyimpan *text database* yang disebut *tweet.* Untuk memudahkan memperoleh informasi yang dominan dibicarakan, maka dicarilah topik dari *tweet twitter* dengan menggunakan *clustering.* Pada penelitian ini, dilakukan pengelompokkan 500 *tweet* dari akun *twitter* @detikcom menggunakan *k-means clustering*. Hasil dari penelitian ini menunjukkan bahwa *Dunn index* yang maksimum*,* pengelompokan terbaik k-means *Clustering* untuk memperoleh topik yang dominan yaitu sebanyak tiga *cluster*, yaitu mengenai pemerintah, Jakarta, dan politik.

Kata Kunci: *text mining, clustering, k-means , dunn index,* dan *twitter*

ABSTRACT

The Internet is an extraordinary phenomenon. Starting from a military experiment in the United States, the Internet has evolved into a 'need' for more than tens of millions of people worldwide. The number of internet users is large and growing, has been creating internet culture. One of the fast growing social media twitter. Twitter is a microblogging service that stores text database called tweets. To make it easier to obtain information that is dominant discussed, then sought the topic of twitter tweet using clustering. In this research, grouping 500 tweets from twitter account @detikcom using k-means clustering. The results of this study indicate that the maximum index Dunn, the best grouping K-means clustering to obtain the dominant topic as many as three clusters, namely the government, Jakarta, and politics.

Keywords: text mining, clustering,, k-means , dunn index, and twitter.

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang

Internet merupakan suatu fenomena yang luar biasa. Berawal dari sebuah eksperimen militer di Amerika Serikat, internet telah berkembang menjadi 'kebutuhan' bagi lebih dari puluhan juta orang di seluruh dunia. Jumlah pengguna internet yang besar dan semakin berkembang, telah mewujudkan budaya internet.

Menurut Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) (2014), pengguna internet di Indonesia selalu bertambah dari tahun ke tahun. Jumlah pengguna internet di Indonesia mencapai 88 juta orang hingga akhir tahun 2014 atau mengalami kenaikkan sebesar 34,9% jika dibandingkan dengan tahun 2013.

Hasil riset tahun 2014 secara signifikan menunjukkan pengguna jejaring sosial (sosial media) menduduki peringkat tertinggi yang dimanfaatkan, mengalahkan pencarian informasi (*browsing*/*searching*) di posisi kedua. Posisi ke-3 *chatting* (*messaging*), pencarian berita (ke-4), video (ke-5), *email* (ke-6). Pencarian berita dan penggunaan *email* saat ini anjlok tak populer (APJII, 2014).

Menurut Kominfo (2013), di era globalisasi perkembangan telekomunikasi dan informatika (IT) sudah begitu pesat. Teknologi membuat jarak tak lagi jadi masalah dalam berkomunikasi. Internet tentu saja menjadi salah satu medianya. Situs jejaring sosial yang paling banyak diakses adalah *facebook* dan *twitter*. Indonesia menempati peringkat 4 pengguna *facebook* terbesar setelah USA, Brazil, dan India.

Sedangkan, untuk *twitter* Indonesia menempati peringkat 5 pengguna *twitter* terbesar di dunia. Posisi Indonesia hanya kalah dari USA, Brazil, Jepang dan Inggris. Pengguna *twitter* di Indonesia berdasarkan data PT Bakrie Telecom sebesar 19,5 juta pengguna dari total 500 juta pengguna global. *Twitter* menjadi salah satu jejaring sosial paling besar di dunia sehingga mampu meraup keuntungan mencapai USD 145 juta.

Menurut Francis dan Flynn (2010), *text mining* adalah teknologi baru yang digunakan untuk data perusahaan yang selalu bertambah sehingga data teks yang tidak terstruktur tersebut dapat dianalisis. Salah satu inovasi *software* yang dapat meringankan biaya bagi penambang teks adalah *software* yang bersifat *open source*. Dua jenis *sofware open source* yang sangat populer dan diunggulkan adalah R dan Perl. R adalah bahasa pemrograman yang mendukung hal-hal yang berkaitan dengan statistik dan digunakan pada hal-hal yang berhubungan dengan ilmu pasti, matematis.

Menurut Zhao (2012), metode *text mining* telah digunakan untuk menganalisa data pada *twitter.* Metode ini dimulai dengan mengambil *text* yang ada pada *twitter, text* yang sudah diambil kemudian diubah menjadi *document-term matrix.* Setelah itu, *frequent words* dan *assosiation* yang diperoleh dari *matrix. Wordcloud* digunakan untuk menunjukkan kata-kata penting yang ada pada dokumen. Terakhir untuk mendapatkan topik dari *tweet,* kata-kata dalam *tweet* atau biasa disebut *term* akan dikelompokkan dengan metode *k-means cluster.*

R adalah salah satu *software open source* untuk komputasi statistik dan grafik. R menyediakan berbagai varian metode stastitik dan grafik. R dapat dipermudah dengan adanya *packages*. Berdasarkan pada CRAN (2016) terdapat 8042 *packages*

yang tersedia pada CRAN *packages repository* per 5 Maret 2016. Untuk melakukan analisa dengan metode *text mining* pada sebuah akun *twitter* dibutuhkan beberapa packages, seperti *packages twitter* dan *tm* diperlukan untuk membantu mendapatkan data pada akun tersebut serta menjelmakan teks. Ada juga *packages word cloud* yang digunakan untuk merepresentasikan visual untuk data teks, biasanya untuk menggambarkan metadata kata kunci (*tag*) di situs *web*. *Tags* biasanya satu kata, dan pentingnya setiap *tag* ditunjukan dengan ukuran *font* atau warna (Zhao, 2012).

Beberapa informasi penting yang dapat diperoleh dari twitter antara lain seperti melihat sejarah perkembangan manusia, sejarah obama terpilih menjadi presiden, dll. Tersedia dalam *tweet-tweet* yang bisa dirunut di *twitter*. Penelitian ini dilakukan pengelompokkan 500 *tweet* dari akun *twitter* @detikcom menggunakan metode *k-means clustering* yang bertujuan untuk untuk mengetahui kecenderungan topik pemberitaan dan mengetahui topik yang paling sering muncul. Hasil analisis pada akun *twitter* berita tersebut akan memberikan gambaran pemberitaan akhir-akhir ini. Penelitian ini menjadi penting mengingat akun @detikcom merupakan akun berita *online* dengan *followers* terbanyak, sehingga berita yang disampaikan juga akan mempengaruhi pengetahuan dan presepsi publik terhadap suatu masalah.

Berdasarkan uraian diatas maka peneliti tertarik untuk menganalisa kecenderungan topik informasi pemberitaan yang disampaikan melalui akun *twitter* @detikcom dengan menggunakan metode *text mining.*

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang dapat dirumuskan permasalahan sebagai berikut:

1. Bagaimana kecenderungan topik informasi yang disampaikan melalui akun *twitter* @detikcom?
2. *Cluster tweet* apa saja yang terbentuk dari akun *twitter* @detikcom?

1.3 Batasan Masalah

Dalam penelitian ini, masalah dibatasi hanya pada 500 *tweets* teratas yang diambil dari *timeline* akun *twitter* @detikcom pada hari Jum'at, 3 Juni 2016 jam 18.30 WIB.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut :

1. Untuk mengetahui kecenderungan topik informasi yang disampaikan melalui akun *twitter* @detikcom?
2. Untuk mengetahui *cluster tweet* apa saja yang terbentuk dari akun *twitter* @detikcom.