PEMILIHAN KATA KUNCI UNTUK DETEKSI KEJADIAN *TRIVIAL* PADA DOKUMEN TWITTER MENGGUNAKAN *AUTOCORRELATION WAVELET COEFFICIENTS*

Rizal Setya Perdana[1)], Chastine Fatichah[2)], dan Diana Purwitasari[3)]
[1)]Program Studi Informatika, Fakultas Ilmu Komputer
Universitas Brawijaya, Malang
[2,3)]Jurusan Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember, Surabaya
e-mail: rizalespe@ub.ac.id[1)], chastine@cs.its.ac.id[2)], diana@if.its.ac.id[3)]

ABSTRAK

Pada penelitian ini diajukan sebuah sistem pendeteksian kejadian yang berulang secara periodik (trivial) dengan pemilihan kata kunci kejadian penting menggunakan perhitungan korelasi (autocorrelation) pada wavelet coefficient. Pemilihan kata kunci dilakukan untuk menemukan kata yang berulang secara periodik yang dianggap sebagai kejadian trivial. Hasil penelitian menunjukkan pemilihan kata kunci dengan nilai confidence boundary yang paling optimal adalah 0.20 pada nilai autocorrelation sebesar 31. Proses yang dilakukan oleh pengguna untuk menemukan kata kunci dari suatu kejadian, secara manual pengguna harus membaca banyak tweet dalam jumlah tertentu. Kata kunci yang merepresentasikan suatu kejadian penting menentukan tingkat penting atau tidaknya suatu kejadian. Pengguna twitter memiliki keterbatasan untuk membaca seluruh tweet yang ada untuk mengetahui adanya suatu kejadian. Sistem deteksi kejadian pada twitter telah dilakukan oleh para peneliti dalam bidang analisis sosial media. Pendeteksian kejadian trivial atau tidak penting yang terpisah dari kejadian penting diperlukan untuk memisahkan dua kejadian tersebut. Proses eliminasi terhadap kejadian trivial akan menyisakan tweet kejadian penting. Salah satu kejadian trivial adalah kejadian yang berulang secara periodik dimana membutuhkan suatu cara spesifik untuk mendeteksi kemunculannya. Pendeteksian kejadian dilakukan dengan memanfaatkan pola-pola temporal atau sinyal dari data Twitter dalam bentuk sinyal wavelet untuk mendeteksi kemunculan kejadian penting. Pada penelitian ini melakukan pendeteksian kejadian yang berulang secara periodik dengan pemilihan kata kunci untuk kejadian penting.

Sistem pendeteksian kejadian penting melakukan perhitungan terhadap autocorrelation pada koefisien wavelet. Hasil perhitungan menunjukkan bahwa pemilihan kata kunci paling optimal pada nilai confidence boundary sebesar 0.20 dan nilai autocorrelation sebesar 31.

***Kata Kunci**: Autocorrelation, Deteksi kejadian (event detection), Twitter, Wavelet*

ABSTRACT

In this study proposed a detection of recurrent events periodically for keyword selection of important events by calculating the correlation (autocorrelation) in the wavelet coefficients. The results showed the selection of keywords will be optimal with confidence boundary in 0.20 and autocorrelation value 31. In order to recognize event keywords, user manually reading tweets in a twitter stream becomes necessary. Keywords to represent an important event determines event critical level. However user has a limited time to read all of those tweets. Event detection system on Twitter draws researchers' attention in the field of social media analysis. In an event detection system, regular trivial events are separated from important non-trivial events. Trivial tweet elimination makes remaining tweets are the important ones. Periodical recurring events considered as trivial requires a specific method to detect the event occurrence for further elimination. Event detection uses the temporal patterns or wavelet signals of Twitter data to identify changes and peaks in a signal as occurrence signs of important events. This proposed study detected periodical recurrent events for selecting keywords of important events.

The proposed event detection system calculated autocorrelation in the wavelet coefficients. The results showed that the keywords selection will be optimal with confidence boundary value in 0.20 and autocorrelation value in 31.

***Keywords**: Autocorrelation, Event Detection, Twitter, Wavelet.*

I. PENDAHULUAN

TWITTER merupakan layanan jejaring sosial yang memiliki perbedaan dengan jejaring sosial media yang lain yaitu memiliki ukuran panjang teks terbatas 140 karakter [1]. Pesan yang dikirim cenderung ditulis secara singkat dan langsung pada inti dari informasi apa yang ingin disampaikan. Namun, data twitter mengandung banyak tweet yang tidak memiliki arti atau tidak merujuk pada kejadian tertentu (*pointless babbles*) [2].

Beberapa penelitian melakukan pendeteksian kejadian dengan memanfaatkan pola-pola secara temporal atau sinyal dari data twitter seperti penelitian yang dilakukan pada [1][3][4] yang memanfaatkan sinyal wavelet untuk mendeteksi munculnya kejadian penting. Berdasarkan penelitian sebelumnya, sinyal individu dari masing-masing kata atau term dibentuk dengan berdasarkan wavelet analysis pada frekuensi masing-masing kemunculan kata

[3]. Dalam pemrosesan sinyal, wavelet analysis merupakan metode yang sering digunakan untuk mendeteksi perubahan dan puncak pada sinyal sehingga dapat dimanfaatkan untuk mendeteksi kemunculan suatu kejadian. Inti dari wavelet analysis adalah pada wavelet transformation yaitu merubah sinyal dari time domain menjadi time-scale domain [3].

Kontribusi pada penelitian ini adalah melakukan pencarian kejadian yang berulang secara periodik dengan melakukan perhitungan korelasi antar *coefficient (autocorrelation)* pada *wavelet* kemunculan kata kunci data twitter. Tujuan utama dari penelitian ini adalah untuk melakukan pemilihan kata kunci yang dianggap sebagai representasi kejadian *trivial* atau berulang secara periodik untuk kemudian dieliminasi. Penggunaan *wavelet analysis* masih belum dapat mendeteksi kejadian yang berulang secara periodik yang dapat dianggap sebagai kejadian trivial. Salah satu penelitian yang berusaha menemukan sinyal yang berulang secara periodik dilakukan pada [5] yang diterapkan pada bidang mekanik untuk mendeteksi kerusakan atau gangguan pada mesin.

Oleh karena itu, dibutuhkan sebuah langkah untuk mendeteksi kata kunci yang berulang secara periodik untuk mendapatkan kejadian penting pada Twitter. Tweet yang mengandung kata kunci yang termasuk kedalam kejadian trivial tidak akan sebagai hasil deteksi kejadian pada twitter sehingga deteksi yang dihasilkan mampu mencakup informasi penting pada kumpulan tweet. Struktur makalah ini selanjutnya dapat digambarkan sebagai berikut. Bagian II menjelaskan penelitian yang berhubungan dengan dengan metode yang diusulkan. Bagian III adalah detail metode yang diusulkan dengan hasil percobaan pada bagian IV. Terakhir, kesimpulan terdapat pada bagian V.

II. Deteksi Kejadian Pada Twitter

Deteksi kejadian memiliki tujuan untuk menemukan peristiwa yang telah terjadi dimana masing-masing kejadian tersebut merujuk pada kejadian yang memiliki waktu dan tempat spesifik [6]. Dilihat dari jenis kejadian yang dideteksi, penelitian dibagi menjadi dua bagian yaitu kejadian yang terspesifik dan kejadian yang tidak terspesifik. Twitter berisi pesan singkat yang berisi reportase dari seluruh kejadian baik dari lingkup sempit atau lokal maupun global. Pesan-pesan yang tergolong tidak memiliki arti (*trivial*) atau merujuk pada kejadian tertentu sebagai contoh adalah iklan, konten pornografi, pengguna yang hanya sekedar ingin menaikkan reputasi saja, dan kejadian yang berulang secara periodik. Hal ini menjadi tantangan utama dalam melakukan deteksi kejadian pada twitter untuk memisahkan kejadian biasa dan pesan yang tidak memiliki arti dengan pesan singkat yang mengandung kejadian atau peristiwa dalam dunia nyata.

A. Karakteristik Kejadian pada Twitter

Twitter merupakan layanan jejaring sosial yang memiliki perbedaan dengan jejaring sosial media yang lain yaitu memiliki ukuran panjang teks terbatas 140 karakter [1]. Batasan tersebut menyebabkan pengguna dengan mudah mengirim tweet dengan cepat tentang informasi yang akan disampaikan. Pengguna mengirim pesan singkat berisi kritik, saran, opini, kabar berita, suasana hati penulis, peristiwa, fakta, dan hal lain yang tidak terkategorikan. Pesan yang dikirim cenderung ditulis secara singkat dan langsung pada inti dari informasi apa yang ingin disampaikan.

Saat ini jumlah pengguna twitter telah mencapai angka 140 juta pengguna aktif yang rata-rata per hari mengirimkan pesan singkat sejumlah 400 juta pesan [7]. Angka-angka tersebut menunjukkan bahwa twitter banyak digunakan karena beberapa hal seperti portabilitas, mudah dalam penggunaan, berisi pesan yang singkat, dan tidak ada batasan pengguna untuk menyebarkan informasi melalui media tersebut. Dari sekian banyak pesan singkat yang dikirimkan tersebut, sebagian besar tweet merupakan laporan peristiwa atau kejadian yang dialami atau diketahui oleh pengguna [8]. Sebagai contoh peristiwa yang dibahas tersebut adalah kejadian terkait sosial seperti adanya kejadian yang menimpa tokoh di masyarakat, pesta olahraga, pemilu presiden, kemacetan di suatu wilayah, bencana alam, dan sebagainya.

Beberapa penelitian tentang deteksi kejadian pada twitter secara umum dapat diklasifikasikan menjadi tiga fokus penelitian yaitu kejadian yang sudah spesifik ditentukan, kejadian yang terfokus pada seorang tokoh, dan deteksi kejadian umum atau tidak spesifik [9]. Pada penelitian yang akan dilakukan termasuk ke dalam kategori yang ketiga yaitu melakukan deteksi kejadian yang tidak spesifik atau yang bersifat umum. Oleh karena tidak memiliki informasi tentang kejadian yang akan dideteksi, beberapa penelitian melakukan pendeteksian kejadian dengan memanfaatkan pola-pola secara temporal atau sinyal dari data twitter seperti penelitian yang dilakukan pada [1][3]. Selain memanfaatkan pola-pola sinyal, metode lain yang banyak digunakan adalah metode klaster seperti yang dilakukan pada penelitian-penelitian [9][10][11]. Metode klaster mengelompokkan kata-kata yang sering muncul ke dalam klaster tertentu dimana kata-kata yang terdapat pada satu klaster dianggap sebagai representasi kejadian yang sama.

B. Analisis Wavelet untuk Deteksi Kejadian Trivial

Kebutuhan akan resolusi tinggi dalam analisis sinyal non-stasioner telah mengakibatkan perkembangan berbagai sarana (tools) untuk menganalisis data-data sinyal non-stasioner (yaitu sinyal yang kandungan frekuensinya bervariasi terhadap waktu). Metode Transformasi berbasis Wavelet merupakan suatu sarana yang dapat digunakan untuk menganalisis sinyal-sinyal non-stasioner. Dalam beberapa tahun terakhir ini, metode ini telah dibuktikan kegunaannya dan sangat populer di berbagai bidang ilmu. Analisis Wavelet dapat digunakan untuk menunjukkan kelakuan secara temporal pada suatu sinyal. Metode Transformasi Wavelet dapat digunakan untuk menyaring data, menghilangkan sinyal-sinyal yang tidak diinginkan serta mendeteksi kejadian-kejadian tertentu pada sinyal [12].

Transformasi Wavelet juga sangat berguna untuk menganalisis sinyal-sinyal non-stasioner karena berkaitan dengan kemampuannya untuk memisahkan berbagai macam karakteristik pada berbagai skala [13]. Pada data twitter yang akan diproses frekuensi kata kunci yang muncul tidak konstan atau non-stasioner sehingga penggunaan Wavelet sesuai dengan data yang digunakan. Proses yang dilakukan dalam transformasi wavelet adalah pertama kali membentuk sinyal yang berasal dari data frekuensi terhadap waktu. Tahap selanjutnya adalah mendekomposisi sinyal menggunakan beberapa jenis wavelet yang salah satunya adalah db1. Hasil dari proses dekomposisi adalah coefficients yang nantinya akan dilakukan proses scaling atau translating sehingga coefficients disusun kembali pada tahap constructing.

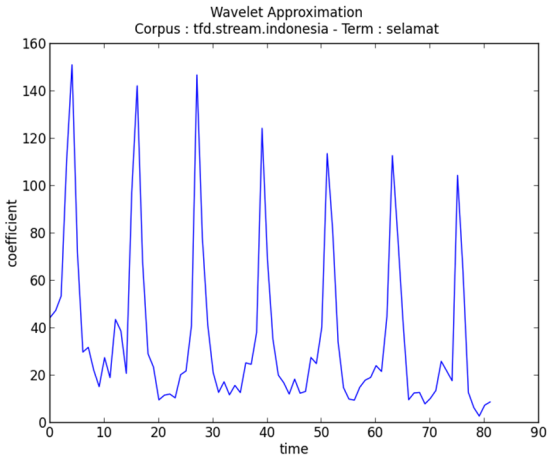
Secara khusus Wavelet digunakan dalam dua cara yaitu sebagai alat untuk mengekstraksi informasi suatu data dan sebagai penyajian atau karakterisasi suatu data. Dalam mengekstraksi informasi, merujuk pada sifat utama Wavelet yaitu time-frequency localization dimana analisis sinyal menggunakan Wavelet adalah bahwa dapat dipelajarinya karakteristik sinyal secara lokal dan detail, sesuai dengan skalanya. Penyajian data menggunakan Wavelet dilakukan dengan cara ekspansi tak berhingga dari Wavelet yang diulur (dilated) dan digeser (translated) [14]. Wavelet merupakan fungsi matematik yang membagi-bagi data menjadi beberapa komponen frekuensi yang berbeda-beda, kemudian dilakukan analisis untuk masing-masing komponen menggunakan resolusi yang sesuai dengan skalanya [15].

Wavelet analysis merupakan metode untuk melakukan pengukuran terkait kapan dan bagaimana frekuensi dari sinyal berubah terhadap waktu [3]. Apabila dibandingkan dengan Fourier, kedua metode ini dalam menganalisis sinyal diawali dengan memecah sinyal mencadi potongan-potongan sinyal. Wavelet baik digunakan untuk memproses sinyal yang tidak teratur dan berubah-ubah bentuk. Keunggulan wavelet adalah mampu menemukan korelasi atau hubungan antara waktu dan frekuensi pada domain sinyal. Inti dari wavelet analysis adalah wavelet transformation yaitu merubah sinyal dari time domain menjadi time-scale domain [3]. Proses pemecahan sinyal akan menghasilkan wavelet coefficients dan himpunan basis fungsi. Himpunan basis fungsi disebut sebagai wavelet family terbentuk dari proses scaling dan translating dari mother wavelet $\psi(t)$. Proses scaling pada wavelet adalah melakukan peregangan atau penyusutan $\psi(t)$, sedangkan proses translation hanya memindahkan posisi temporal tanpa melakukan perubahan pada sinyal itu sendiri. Pada penelitian yang dilakukan oleh [3] perhitungan *Wavelet Family* dituliskan seperti pada persamaan 1.

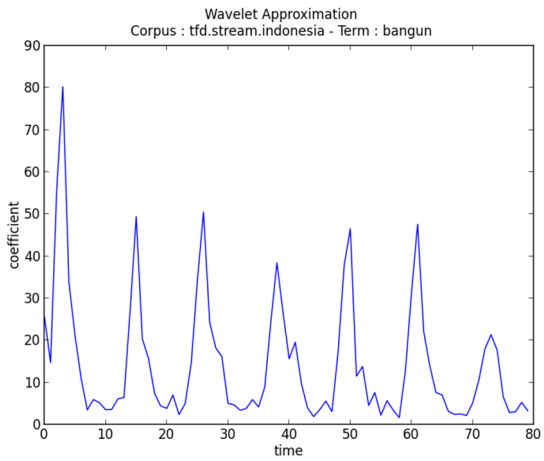
$$\varphi_{a,b} = |a|^{-1/2} \varphi\left(\frac{t-b}{a}\right) \quad (1)$$

Pada persamaan 1, $a, b \in \Re$ dimana *a* dan *b* adalah nilai scaling dan translating dan *t* adalah waktu.

Wavelet transformation dibagi menjadi continuous wavelet transformation (CWT) dan discrete wavelet transformation (DWT). Secara umum, pada saat proses analisis CWT menghasilkan representasi sinyal yang berulang atau redundant. Selain itu CWT apabila dilakukan pemrosesan atau transformasi secara langsung membutuhkan waktu yang cukup lama. Berkebalikan dengan DWT, proses yang dilakukan akan menghasilkan representasi sinyal yang tidak berulang atau non-redundant dan tidak membutuhkan waktu yang banyak ketika melakukan tranformasi. Pada penelitian akan digunakan DWT sebagai pemroses Wavelet karena data frekuensi twitter berupa data diskrit. Gambar 1 menunjukkan *wavelet* kata kunci yang memiliki tingkat keperiodikan yang cukup tinggi.



Gambar. 1 (a). *Wavelet* pada kata kunci "selamat"



Gambar. 1 (b). *Wavelet* pada kata kunci "bangun"

III. METODE YANG DIUSULKAN

Penelitian ini melakukan pemilihan kata kunci untuk deteksi kejadian *non-trivial* dengan memanfaatkan pola temporal dari kata kunci (*term*) yang berulang secara periodik. Proses penentuan kata kunci berulang secara periodik atau tidak adalah dengan menghitung *autocorrelation* dari wavelet seperti yang digambarkan pada diagram alir tahapan metode pada Gambar 2. Proses dimulai dengan pengumpulan dokumen tweets yang selanjutnya dilakukan preproses teks, transformasi teks dalam periode tertentu, transformasi sinyal wavelet, perhitungan *autocorrelation,* dan penentuan nilai ambang untuk nilai korelasi.

A. Transformasi sinyal wavelet frekuensi kata kunci

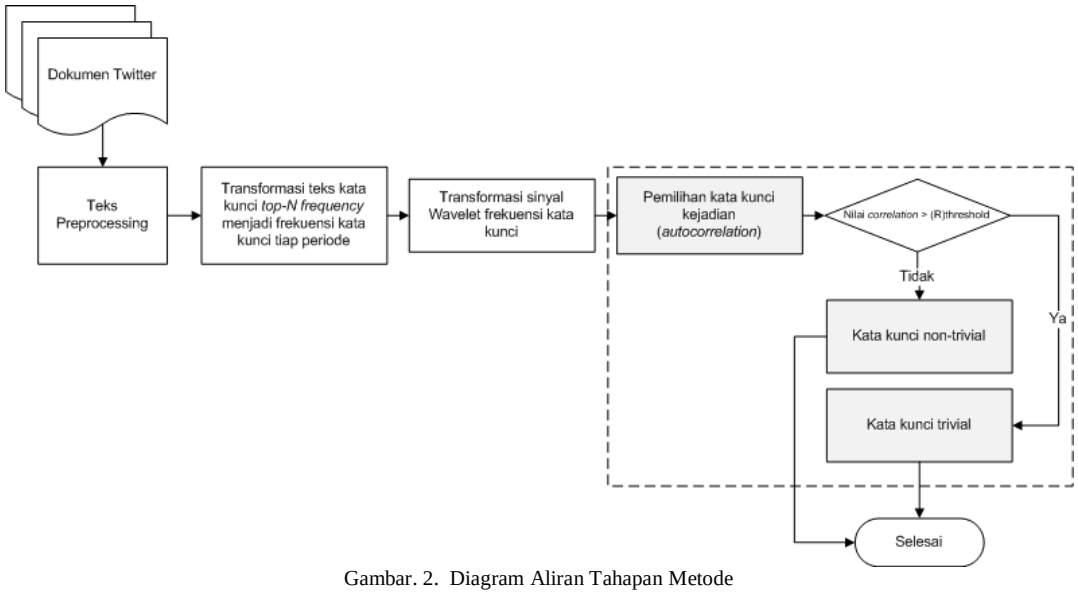
Transformasi sinyal *wavelet* dimulai dnegan proses dekomposisi sinyal yaitu proses merubah data frekuensi kemunculan kata yang sudah tersusun dalam urutan waktu menjadi koefisien. Proses ini dilakukan pada kata-kata yang memiliki document frequency N besar dari keseluruhan kemunculan kata pada rentang interval waktu ternetu. Pada penelitian ini jenis *wavelet* yang dihasilkan adalah bersinyal diskrit karena kejadian kemunculan bersifat diskrit dengan jenis *wavelet* Coifman atau yang disebut sebagai Coiflet. Jenis *wavelet* ini memiliki memiliki sifat pemampatan yang sama baik untuk coefficient aproksimasi maupun detail.

B. Pencarian Nilai Korelasi Sinyal Terhadap Sinyal itu Sendiri (Autocorrelation)

Korelasi merupakan operasi matematika yang mirip dengan proses konvolusi. Sebagaimana pada konvolusi, korelasi menggunakan dua siyal untuk menghasilkan sinyal ketiga. Sinyal ketiga ini disebut dengan cross correlation dari dua sinyal masukan. Jika sinyal dikorelasikan dengan dirinya sendiri, maka hasilnya disebut *autocorrelation.* Pendeteksian sinyal berulang secara periodik pada sinyal non-stasioner dapat dilakukan dengan metode *autocorrelation* pada Wavelet. Sinyal yang berulang secara periodik merepresentasikan kejadian yang berulang pada data twitter sehingga tidak diperlukan dalam proses peringkasan. Autocorrelation digunakan dalam analisis sinyal dengan membaca perubahan sinyal berdasarkan waktu menggunakan hubungan probabilitas. Perhitungan dilakukan dengan membandingkan coefficients yang berasal dari proses transformasi sinyal Wavelet sehingga dapat membandingkan apakah antar coefficient [5]. Prinsip kerja pada *autocorrelation* adalah dengan menggeser sinyal dengan beberapa penyesuaian pada waktu atau periode yang dinyatakan pada persamaan 2.

$$C(m) = \frac{1}{n-m} \sum_{n=0}^{n-m-1} x(n)x(n-m) \quad (2)$$

Pada persamaan 2, *C* adalah fungsi autocorrelation, *x (n)* adalah koefisien Wavelet dan m adalah pergeseran waktu. Setelah ditemukan sinyal-sinyal yang berulang pada Wavelet maka kata kunci tersebut tidak diperhitungkan dalam melakukan peringkasan dokumen yang dilakukan selanjutnya. Perhitungan autocorrelation ditampilkan dalam diagram correlogram seperti nampak pada gambar dibawah ini pada Lag pertama nilai korelasi melebihi batas threshold yang telah ditentukan sehingga hal tersebut dapat disimpulkan sebagai kejadian berulang karena memiliki korelasi tinggi pada sebuah *Wavelet*.



Gambar. 2. Diagram Aliran Tahapan Metode

TABEL I
FREKUENSI KATA KUNCI

No	Kata Kunci	Frekuensi	No	Kata Kunci	Frekuensi	No	Kata Kunci	Frekuensi
1	pagi	12560	18	cinta	4090	35	bangun	3171
2	selamat	8411	19	langor	4080	36	semangat	2996
3	orang	6956	20	jalan	3942	37	kali	2840
4	tidur	6231	21	sakit	3939	38	indah	2807
5	hati	5561	22	banget	3791	39	mata	2729
6	makan	5545	23	moga	3715	40	main	2718
7	kalo	5479	24	suka	3442	41	hidup	2709
8	morning	5239	25	time	3382	42	lupa	2707
9	kuala	5163	26	savekpk	3374	43	dont	2694
10	happy	4753	27	negeri	3349	44	follback	2673
11	malam	4713	28	biar	3345	45	salah	2668
12	lumpur	4675	29	kerja	3259	46	sekolah	2627
13	sayang	4641	30	follow	3257	47	alhamdulillah	2500
14	love	4522	31	jakarta	3256	48	hujan	2460
15	rumah	4399	32	wkwk	3221	49	bambang	2447
16	polri	4375	33	indonesia	3192	50	rindu	2440
17	good	4228	34	anak	3184			

IV. ANALISA DAN HASIL UJI COBA

Tweet diambil dengan melakukan proses *crawling* menggunakan Twitter API dengan memanfaatkan *library* Python twitter 1.16.0. Dokumen *tweets* diambil menggunakan metode *Streaming APIs* Twitter dimana proses *crawling* tidak melakukan penyaringan atau pemilihan terhadap kata kunci tertentu. Proses pembatasan *crawling* pada tweet berbahasa Indonesia dilakukan dengan membatasi proses Stream dengan konfigurasi *geolocation* atau lokasi geografis Negara Indonesia yaitu 94,-11,141,6.

Crawling dilakukan secara kontinyu atau berkelanjutan selama sepuluh hari (15-24 Januari 2015) dengan jumlah total tweet sebesar ±600000. Proses crawling secara langsung memproses setiap tweet dengan melakukan tahap praproses seperti *segmentation*, *stopword removal*, dan *stemming*. Tabel I menunjukkan tabel kata kunci dan frekuensi kemunculan yang telah dikumpulkan pada proses *crawling* dan diurutkan berdasarkan frekuensi.
Kata kunci yang memiliki frekuensi kemunculan tinggi dapat diartikan bahwa kata kunci tersebut sering digunakan atau ditulis oleh pengguna Twitter. Pada penelitian ini diambil lima puluh kata kunci yang memiliki frekuensi tertinggi untuk dilakukan proses perhitungan apakah kata kunci tersebut merupakan kata kunci yang berulang secara periodik atau tidak.

TABEL II
DETAIL WAKTU KEMUNCULAN KATA KUNCI ("PAGI")

No	Kata Kunci	Waktu
1	pagi	2015-01-15 00:07:21
2	pagi	2015-01-15 00:20:09
3	pagi	2015-01-15 00:22:15
4	pagi	2015-01-15 00:25:44
5	pagi	2015-01-15 00:25:44
6	pagi	2015-01-15 00:28:43
7	pagi	2015-01-15 00:33:15
8	pagi	2015-01-15 00:34:33
9	pagi	2015-01-15 00:35:54
10	pagi	2015-01-15 00:37:53
11	pagi	2015-01-15 00:39:28
12	pagi	2015-01-15 00:42:16
13	pagi	2015-01-15 00:47:20
14	pagi	2015-01-15 00:53:40
15	pagi	2015-01-15 00:59:01
16	pagi	2015-01-15 00:59:12
17	pagi	2015-01-15 01:00:11

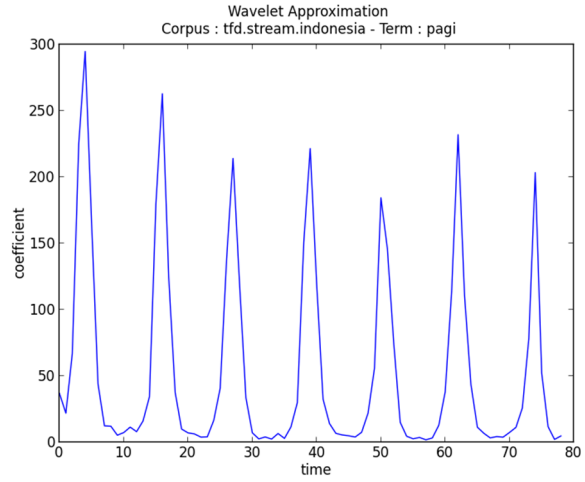
TABEL III
FREKUENSI KATA KUNCI ("PAGI") TIAP INTERVAL

No	Interval waktu	Frekuensi
1	2015-01-15 00:00:00	16
2	2015-01-15 01:00:00	25
3	2015-01-15 02:00:00	50
4	2015-01-15 03:00:00	81
5	2015-01-15 04:00:00	141
6	2015-01-15 05:00:00	253
7	2015-01-15 06:00:00	186
8	2015-01-15 07:00:00	183
9	2015-01-15 08:00:00	119
10	2015-01-15 09:00:00	49
11	2015-01-15 10:00:00	37
12	2015-01-15 11:00:00	14
13	2015-01-15 12:00:00	10
14	2015-01-15 13:00:00	7
15	2015-01-15 14:00:00	10
16	2015-01-15 15:00:00	6
17	2015-01-15 00:00:00	16

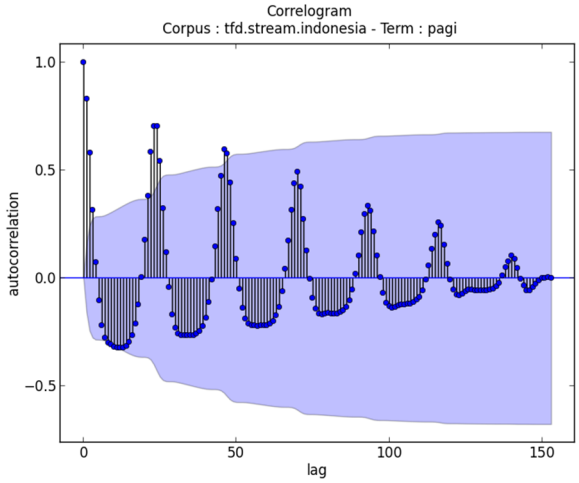
Transformasi *wavelet* diawali dengan melakukan perhitungan jumlah kata pada interval waktu tertentu untuk mendapatkan jumlah kemunculan tiap interval waktu. Pada Tabel II merupakan cuplikan detail waktu kemunculan kata kunci "pagi" untuk kemudian diproses pengelompokan berdasarkan interval sebesar 60 menit. Setelah dikelompokkan setiap 60 menit maka akan didapatkan frekuensi kemunculan tiap interval yang ditransformasi sinyal kata kunci *wavelet* dengan *mother wavelet* berjenis Coiflet. Proses transformasi akan menghasilkan koefisien *wavelet* yang akan dibentuk menjadi grafik *wavelet* untuk diperhitungkan apakah sinyal kata kunci tersebut berulang secara periodik atau tidak. Pada proses pengujian, perhitungan dilakukan dengan menggunakan 4000 *tweet* frekuensi tertinggi untuk perhitungan *autocorrelation.* Hasil perhitungan *autocorrelation* akan ditampilkan secara grafis dengan diagrap yang disebut dengan *correlogram.*

Pengujian *autocorrelation* adalah dengan menentukan nilai *confidence boundary* atau nilai ambang batas apakah nilai dari sebuah *wavelet* tergolong periodik atau tidak. Apabila sebuah *wavelet* dari kata kunci melewati batas tersebut maka term tergolong *statistically significant*. Pada penelitian ini digunakan nilai *Bartlett's formula* sebagai *confidence boundary* yaitu sebesar 0.05 sebagai acuan dan akan dicoba dengan beberapa nilai lain yaitu 0.10, 0.15, 0.20, 0.25. Gambar 3-8 menampilkan hasil pengujian beberapa kata kunci yang memiliki tingkat periodik tinggi dan rendah berupa *wavelet* dan *correlogram* yang menunjukkan tingkat keperiodikan kemunculan kata kunci tweet.

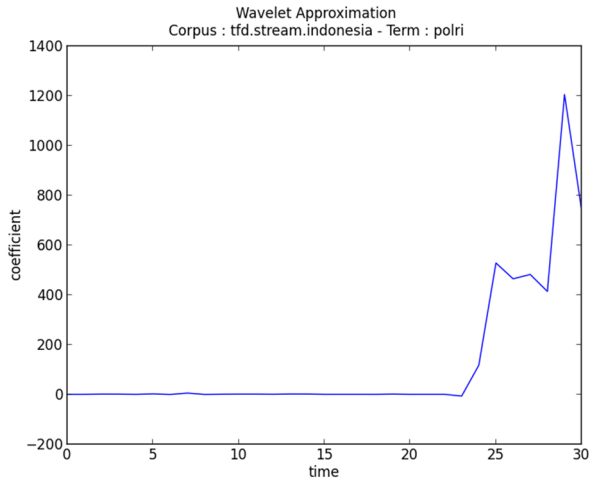
Gambar. 3 sampai 8 adalah untuk menunjukkan perbedaan yang ada pada kata kunci berulang (*trivial*), kata kunci yang tidak memiliki pola, dan kata kunci penting (*non-trivial*). Gambar. 3 dan Gambar. 4 merupakan pasangan *wavelet* dan *correlogram* yang memiliki frekuensi tinggi dan berulang secara periodik. Gambar. 5 dan Gambar. 6 merupakan pasangan yang memiliki frekuensi tinggi namun terjadi lonjakan yang drastis pada waktu tertentu saja.



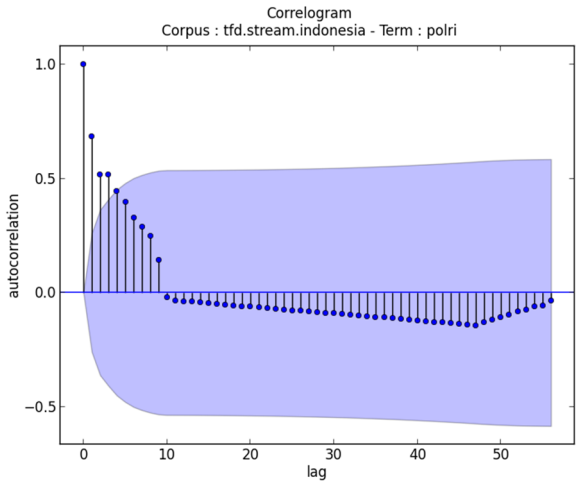
Gambar. 3. *Wavelet* kata kunci "pagi" yang memiliki frekuensi tinggi dan memiliki pola keteraturan yang tinggi



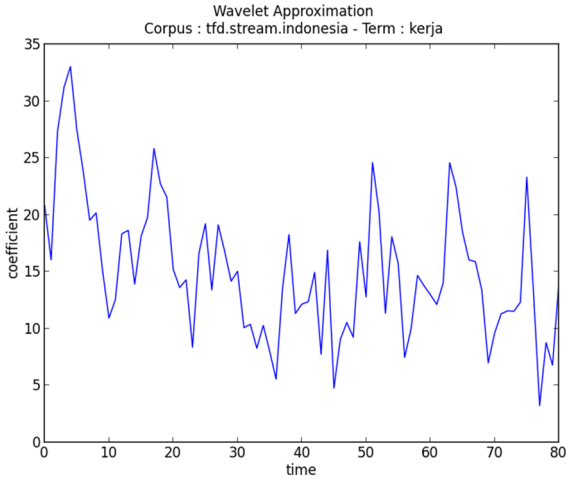
Gambar. 4. *Correlogram* kata kunci "pagi" yang berulang secara periodik



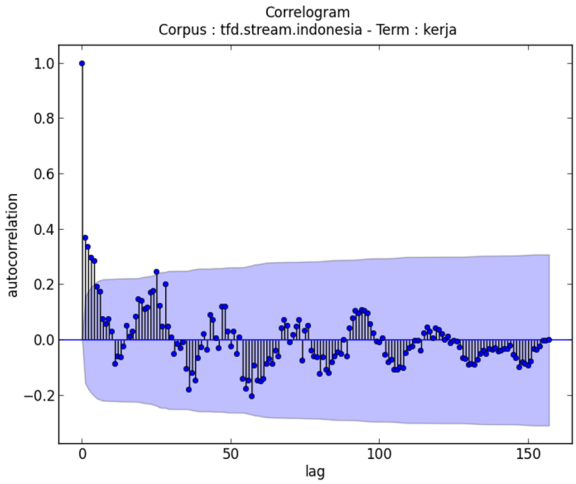
Gambar. 5. *Wavelet* kata kunci "polri" yang memiliki frekuensi tinggi namun terjadi lonjakan pada waktu tertentu saja



Gambar. 6. *Correlogram* kata kunci "polri" yang tidak berulang secara periodik atau tidak memiliki keteraturan



Gambar. 7. *Wavelet* kata kunci "kerja" yang memiliki frekuensi tinggi namun terjadi fluktuasi frekuensi yang tidak teratur



Gambar. 8. *Correlogram* kata kunci "kerja" yang memiliki nilai korelasi sangat rendah sehingga bukan merupakan suatu kejadian penting atau tidak terjadi secara periodik

Gambar. 7 dan Gambar. 8 merupakan pasangan yang memiliki frekuensi tinggi dan fluktuasi yang tidak teratur, sehingga dianggap bukan merupakan representasi dari kejadian *non-trivial*. Berdasarkan *groundtruth* pada lima kali pengujian nilai *confidence boundary* (0.05, 0.10, 0.15, 0.20, 0.25) maka nilai yang terbaik adalah penggunaan *confidence boundary* 0.20 dengan batas minimum *autocorrelation* sebesar 31. Penyimpulan nilai optimal pada pemilihan kedua nilai tersebut adalah berdasarkan pada kata kunci yang dihasilkan pada penggunaan *confidence boundary* dan *autocorrelation* menghasilkan kata kunci *trivial* dan tidak bercampur dengan kata kunci lain yang merupakan kata kunci penting. Tabel IV merupakan rekapitulasi jumlah kata kunci pada masing-masing pengujian dengan batas bawah nilai *autocorrelation* sebesar 31. Tabel V menjelaskan detail kata kunci pada *confidence boundary* 0.20 dan nilai minimum *autocorrelation* sebesar 31 yang akan digunakan pada proses eliminasi.

TABEL IV
TABEL REKAPITULASI CONFIDENCE BOUNDARY PADA AUTOCORRELATION 31

No	*Confidence Boundary*	Jumlah Kata Kunci
1	0.05	1849
2	0.10	151
3	0.15	42
4	0.20	18
5	0.25	9

TABEL V
CONFIDENCE BOUNDARY 0.20

No	Kata Kunci	*Autocorrelation*
1	tidur	61
2	night	61
3	good	60
4	pagi	58
5	malam	52
6	bangun	51
7	sleep	45
8	mall	45
9	selamat	43
10	morning	38
11	cafe	38
12	ayam	38
13	bismillah	36
14	sarap	36
15	lunch	34
16	semangat	31
17	ngantuk	31

V. KESIMPULAN

Pada penelitian ini diusulkan penggunaan perhitungan *autocorrelation wavelet* dalam pola sinyal *wavelet* untuk melakukan deteksi kejadian *non-trivial*. Kejadian *trivial* ditandai dengan adanya kata kunci yang berulang secara periodik sehingga dibutuhkan eliminasi untuk kejadian-kejadian yang berulang secara periodik. Pada proses pengujian dilakukan beberapa variasi nilai *confidence boundary* dan nilai yang paling akurat terhadap *groundtruth* adalah 0.20 pada nilai *autocorrelation* sebesar 31. Penelitian dapat dikembangkan dengan memperhitungkan urutan kemunculan kata dalam mendeteksi kejadian yang berulang secara periodik untuk meningkatkan hasil deteksi.

DAFTAR PUSTAKA

[1] Cordeiro, Mário. "Twitter event detection: Combining wavelet analysis and topic inference summarization." Doctoral Symposium on Informatics Engineering, DSIE. 2012.
[2] Hurlock, J., & Wilson, M. L. (2011, May). Searching Twitter: Separating the Tweet from the Chaff. In ICWSM (pp. 161-168).
[3] Weng, Jianshu, and Bu-Sung Lee. "Event Detection in Twitter." ICWSM 11 (2011): 401-408.
[4] Chen, L., & Roy, A. (2009, November). Event detection from flickr data through wavelet-based spatial analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 523-532). ACM.
[5] Rafiee, J., & Tse, P. W. (2009). Use of autocorrelation of wavelet coefficients for fault diagnosis. Mechanical Systems and Signal Processing, 23(5), 1554
[6] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study final report.
[7] Atefeh, F., & Khreich, W. (2013). A survey of techniques for event detection in Twitter. Computational Intelligence.
[8] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.
[9] Zhao, J., Wang, X., & Ma, Z. (2014). Towards Events Detection from Microblog Messages. International Journal of Hybrid Information Technology, 7(1).
[10] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM, 11, 438-441.
[11] Petrović, S., Osborne, M., & Lavrenko, V. (2010, June). Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 181-189). Association for Computational Linguistics.
[12] Foster, D. J., C.C. Mosher, dan S. Hassanazadeh, (1994). Wavelet Transform Methods for Geophsical Application, 64th Annual International Meeting, Soc. Expl. Geophys., Ecpanded Abstract, 1465 – 1468.
[13] Anant, K. S. dan F.U. Dowla, (1997). Wavelet Transform Methods for Phase Identification in Three-Component Seismograms, Bulletins of Seismological Society America, Vol. 87, No.5, 1598 – 1612.
[14] Kumar, P., & Foufoula Georglou, E. (1997). Wavelet analysis for geophysical application. Reviews of Geophysics, 35(4), 385-412.
[15] Graps, A., (1995), "An Introduction to Wavelets, IEEE Computational Science and Engineering, vol.2, Wavelet in Geophysics, Academic, USA, 1-43.