Deteksi Event pada Hashtag Twitter Menggunakan DF-IDF dan Entropi Wavelet

Amin Ajaib Maggang

Program Studi Teknik Elektro Fakultas Sains dan Teknik Universitas Nusa Cendana
Email: amin_maggang@staf.undana.ac.id

Info Artikel

Histori Artikel:
Diterima Sep 04, 2021
Direvisi Okt 13, 2021
Disetujui okt 29, 2021

ABSTRACT

This study aimed to detect event from a collection of tweets having similar hashtag. Discrete Wavelet Transform (DWT) and Document Frequency-Inverse Document Frequency (DF-IDF) techniques were utilised in this research to develop and analyse the signals. Signals were developed using DF-IDF during a certain period while DWT was applied to capture the sudden change in the DF-IDF signal and displayed it in the form of entropy. Words that have a sudden change in signal value at the same period of time might represent an event related to the topic of the hashtag. Knowing the words will assist users to discover the event associated to the hashtag. Using 3473 tweets collected on the RIPLKY hashtag with the 20 most frequently occurring words, the results showed that DWT managed to capture the sudden change in DF-IDF signal related to a spike occured in the wavelet entropy. The spike was shown by three words, namely proud, sophialimx, and goals. We also found that the three words were under the same tweet, and they described an event about the funeral of Singaporean prime minister.
Keywords: *event detection, topic identification, tweets, twitter, dwt, entropy*

ABSTRAK

Penelitian ini bertujuan untuk mendeteksi event dari kumpulan tweet yang memiliki hashtag yang sama. Teknik Discrete Wavelet Transform (DWT) dan Document Frequency-Inverse Document Frequency (DF-IDF) digunakan dalam penelitian ini untuk mengembangkan dan menganalisis sinyal. Sinyal dibangun menggunakan DF-IDF selama periode tertentu, sedangkan DWT dimanfaatkan untuk menangkap perubahan mendadak pada sinyal DF-IDF dan menampilkannya dalam bentuk entropi. Kata-kata yang memiliki perubahan nilai sinyal secara tiba-tiba pada periode waktu yang sama dapat mewakili suatu peristiwa yang terkait dengan topik hashtag. Mengetahui kata-kata akan membantu pengguna untuk menemukan peristiwa yang terkait dengan hashtag. Menggunakan 3473 tweet yang dikumpulkan pada tagar RIPLKY dengan 20 kata yang paling sering muncul, hasilnya menunjukkan bahwa DWT berhasil menemukan perubahan cepat pada sinyal df-idf terkait dengan lonjakan yang terjadi pada wavelet entropi. Lonjakan itu ditunjukkan dengan tiga kata, yaitu proud, sophialimx, dan goals. Kami juga menemukan bahwa ketiga kata tersebut berada pada tweet yang sama, dan menggambarkan sebuah peristiwa tentang pemakaman perdana menteri Singapura.
Kata Kunci: *event detection,topic identification, tweets, twitter, dwt, entropy*

Penulis Korespondensi:
Amin Ajaib Maggang
Program Studi Teknik Elektro Fakultas Sains dan Teknik,
Universitas Nusa Cendana,
Jl. Adisucipto Penfui - Kupang.
Email: amin_maggang@staf.undana.ac.id

1. PENDAHULUAN

Sebagai salah satu media sosial terpopuler dengan perkiraan jumlah pengguna aktif sekitar 330 juta[1], Twitter telah menjadi objek penelitian di bidang identifikasi topik [2]. Identifikasi event digunakan untuk menentukan kejadian terkait topik yang sedang dibahas dalam aliran data Twitter[3, 4]. Deteksi event dapat

didefinisikan sebagai upaya untuk menemukan peristiwa berdasarkan perubahan volume data teks yang berhubungan dengan suatu topik pada waktu tertentu[5]. Tujuan utama dari pendeteksian event adalah untuk mengenali cerita pertama (*first story*) yang membahas suatu kejadian yang terjadi pada periode waktu dan tempat tertentu [6].

Sebagai kata kunci untuk mewakili topik pembicaraan pada twitter, hashtag (kata yang didahului dengan simbol #) telah banyak digunakan di bidang penelitian deteksi event pada media sosial. Hashtag membantu pengguna untuk dengan mudah mengetahui topik yang sedang viral, tentang apa, dan siapa orang yang pertama kali men-tweet topik tersebut. Namun, peristiwa yang terkait dengan hashtag bisa sulit ditemukan karena tweet dengan hashtag tertentu ditambahkan dengan hashtag lain atau beberapa kata ketika orang melakukan *retweet*. Oleh karena itu, memiliki algoritma yang dapat dengan cepat mendeteksi peristiwa dari topik yang diwakili oleh hashtag adalah penting.

Meskipun banyak teknik telah digunakan untuk mendeteksi peristiwa dari Twitter, sejauh yang kami ketahui, masih sedikit penelitian yang dilakukan untuk mendeteksi peristiwa atau topik hanya dari satu hashtag, dengan melihat perubahan secara tiba-tiba dari sinyal kata terhadap suatu periode waktu. Penelitian oleh [7] menerapkan Discrete Fourier Transform (DFT) untuk mendeteksi peristiwa dengan melihat *burst* dalam domain waktu yang berhubungan dengan naiknya sinyal secara signifikan dalam domain frekuensi. Meskipun demikian, DFT tidak dapat menentukan waktu saat *burst* terjadi, yang mana sangat penting dalam deteksi peristiwa atau topik. Oleh karena itu, penelitian ini menggunakan transformasi wavelet karena wavelet dapat menentukan waktu ketika terjadi perubahan nilai sinyal yang cepat.

2. METODE PENELITIAN

Penelitian ini mengombinasikan DF-IDF dan Discrete Wavelet Transform (DWT) untuk mengetahui perubahan yang terjadi pada sinyal (kata). DF-IDF digunakan untuk membangun sinyal dari kata-kata yang terkandung di dalam tweets pada suatu hashtag. Sedangkan DWT digunakan untuk mendeteksi perubahan sinyal yang terjadi secara tiba-tiba dalam bentuk entropi wavelet.

2.1 H-Measure (Normalisasi Shannon Wavelet Entropy)

H-Measure merupakan normalisasi dari Shannon wavelet entropy (SWE) [8] yang digunakan untuk mengukur distribusi energi sinyal pada skala yang berbeda (band Frekuensi). H-measure didefinisikan sebagai;

$$H(S) = \frac{SWE(S)}{SWE_{\max}} \tag{1}$$

Dimana SWE_{max} diperoleh dengan mengaplikasikan distribusi uniform dari energi sinyal pada berbagai skala misalnya $\{\rho_i\} = \left\{ \frac{1}{N_{I+1}}, \frac{1}{N_{I+1}}, ... \frac{1}{N_{I+1}} \right\}$.

SWE(S) diperoleh dengan menghitung SWE[9] dari sinyal pada distribusi ρ_i.

$$SWE(S) = -\sum_i \rho_i \log \rho_i \tag{2}$$

Dimana ρ_i menunjukkan distribusi energi sinyal wavelet pada setiap skala yang berbeda .Normalisasi nilai ρ_i ini yang juga digunakan pada perhitungan H-Measure.

2.2 Konstruksi Sinyal DF-IDF dan Analisis Wavelet

Sinyal dari setiap kata dibangun dalam dua tahap. Dengan asumsi bahwa T_c adalah waktu sekarang, tahap pertama pembuatan sinyal dari kata *w* pada T_c dapat ditulis sebagai berikut:

$$S_w = [s_w(1), s(2), ...s_w(T_c)] \tag{3}$$

Nilai dari $s_w(t)$ pada setiap sampel *t*, diberikan oleh skor DF-IDF yang didefinisikan oleh:

$$S_w(t) = \frac{N_w(t)}{N(t)} x \log \frac{\sum_{i=1}^{Tc} N(i)}{\sum_{i=1}^{Tc} N_w(i)} \tag{4}$$

$Nw(t)$ adalah jumlah tweets yang berisikan kata-kata yang sering muncul, sedangkan $N(t)$ total tweets dalam periode waktu yang sama.

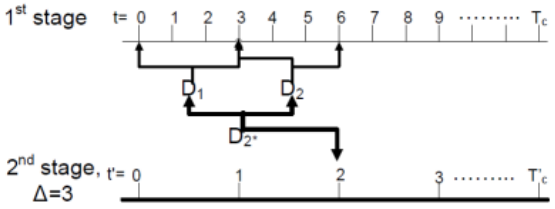
Pada tahap berikutnya, sinyal dikonstruksi menggunakan sliding window Δ yang mencakup beberapa titik sampel sinyal tahap pertama. Sinyal tahap kedua berfungsi untuk menangkap setiap perubahan pada tahap pertama. Berikut adalah sinyal tahap kedua.

$$S'_w = [s'_w(1), s'_w(2), \dots \dots, s'_w(T'_c)] \quad (5)$$

Waktu pada sinyal tahap pertama dan tahap kedua tidak harus dalam satuan yang sama. Nilai $s'_w(t')$ dapat dihitung dengan menggeser sliding window untuk mengambil sebagian sampel poin pada sinyal tahap pertama $s'_w((t' - 2) * \Delta + 1)$ to $s_w((t' - 1) * \Delta)$ dan kemudian melanjutkannya pada tahapan selanjutnya. Hal ini akan menghasilkan pemecahan sinyal pada window tersebut dan dituliskan sebagai $D_{t'-1}$. Selanjutnya H-Measure dari $D_{t'-1}$ dapat dihitung dan dituliskan dengan $H_{t'-1}$. Setelah itu sliding window digeser ke bagian titik sampel berikut $s_w((t' - 1) * \Delta)$ ke $s_w(t' * \Delta)$ untuk perhitungan $D_{t'}$ dan $H_{t'}$.

$$s'_w(t') = \begin{cases} \dfrac{H_{t*} - H_{t'-1}}{H_{t'-1}} \; if \; (H_{t*} > H_{t'-1}); & (6) \\ 0 \qquad\qquad\qquad\quad \text{lainnya} \end{cases}$$

Berikut adalah ilustrasi pembentukan kedua sinyal entropi.



Gambar 1 Analogi tahapan Konstruksi Sinyal[4]

Dari gambar 1 dapat dianalogikan bahwa D_1 merupakan $H_{t'-1}$ dan D_2 merupakan $H_{t'}$. Sedangkan $D_{2'}$ merupakan nilai $s'_w(t')$ pada persamaan(6).

2.3. Cross Correlation

Cross correlation digunakan untuk mengukur kemiripan antara dua sinyal[10]. Jika kedua sinyal direpresentasikan dengan fungsi *x(t)* dan *y(t)*, maka cross correlation dapat dituliskan sebagai.

$$(x * y)(t) = \sum x * (\tau) y(t + \tau) \quad (7)$$

3. HASIL DAN PEMBAHASAN

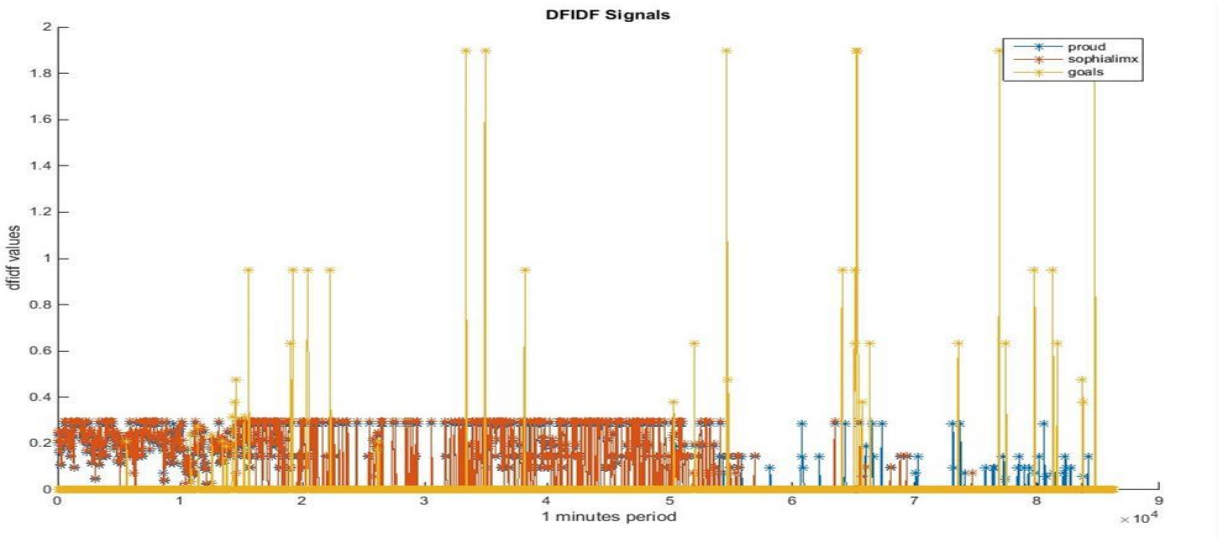
3.1. Data dan prosedur penelitian

Sumber data yang diterapkan pada percobaan pertama ini dikumpulkan dengan langkah-langkah sebagai berikut

1. Tweet diambil menggunakan hashtag RIPLKY pada tanggal 30-31 Maret 2015 menggunakan NCapture, yang merupakan ekstensi web browser dari aplikasi NVivo v.10.
2. Tweet yang dikumpulkan kemudian di-token-kan menjadi kata-kata dan kemudian digunakan untuk membangun sinyal kata-kata individual. Stop words dan kata-kata yang jarang dibuang melalui proses filter karena tidak memiliki arti khusus dan tidak terkait dengan suatu topik.
3. Setelah difilter, hanya 20 kata yang paling banyak muncul dalam dokumen yang digunakan untuk membangun sinyal. Jumlah total tweet yang diterapkan dalam percobaan ini adalah 3474 dokumen.

3.2. Konstruksi Sinyal

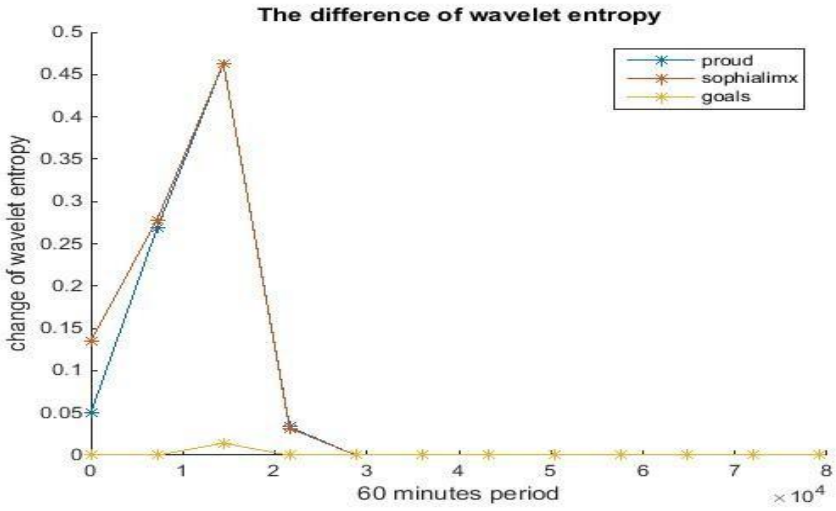
Sinyal df-idf pada Gambar 1 menampilkan pola sinyal dari tiga kata yang memiliki pola sinyal yang sama selama 24 jam, yang diukur menggunakan cross correlation. Sinyal tersebut dibangun dengan menerapkan persamaan (4) dalam interval waktu satu menit. Sinyal memiliki 1440 titik sampel karena sinyal diambil selama 24 jam.



Gambar 2 Nilai DF-IDF (Sinyal Tahap 1)

Untuk mendeteksi perubahan nilai sinyal df-idf setiap jam, maka dua jenis wavelet entropi $H_{t'-1}$ dan H_{t*} telah dihitung. $H_{t'-1}$ berfungsi untuk mendeteksi perubahan df-idf per jam (sliding window Δ = 60), sedangkan H_{t*} dibangun untuk mendeteksi perubahan sinyal df-idf setiap dua jam.

Pola sinyal dari nilai *df-idf* (sinyal tahap pertama) pada Gambar 2 menunjukkan bagaimana tiga keywords berubah-ubah terhadap waktu secara cepat mulai dari 0 sampai sekitar 20000 sample. Namun, hanya kata, "proud" dan "sophialimx" yang hampir sama pola sinyalnya. Nilai perubahan entropi wavelet pada Gambar 3 diperoleh dengan menggunakan (6). Terlihat untuk sampel pertama sampai sampel ke-4 dari ketiga kata (proud, sophialimx, dan goals) nilai $H_{t*} > H_{t'-1}$. Angka H_{t*} yang dibuat tebal adalah angka yang nilainya lebih besar dari $H_{t'-1}$. Angka yang dibuat tebal ini bersama pasangan $H_{t'-1}$ dimasukan pada (6) untuk menghasilkan nilai perubahan entropi wavelet yang ditampilkan pada Gambar 3.

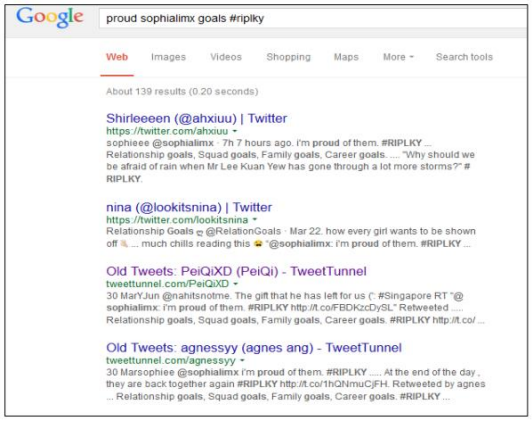


Gambar 3 Wavelet Entropi (Sinyal Tahap 2, delta =6)

Tabel 1 Nilai dua entropi ($H_t'_{-1}$ dan H_{t*} adalah entropi I dan II)

Urutan Sampel	proud		sophialimx		goals	
	$H_t'_{-1}$	H_{t*}	$H_t'_{-1}$	H_{t*}	$H_t'_{-1}$	H_{t*}
1	0.1635	**0.1716**	0.1621	**0.1841**	0	**0.7013**
2	0.2624	**0.3329**	0.2684	**0.343**	0	**0.7263**
3	0.3419	**0.5002**	0.3419	**0.5002**	0.7939	**0.8051**
4	0.7937	**0.8204**	0.7937	**0.8189**	0.7495	**0.7067**
5	0.873	0.7339	0.873	0.7365	0	0
6	0.6118	0.5839	0.6213	0.5904	0.7495	0.7013
7	0.7298	0.7013	0.7298	0.7013	0	0
8	0.8384	0.7772	0.8437	0.7775	0.7495	0.7095
9	0.8274	0.8148	0	0.7013	0	0.7013
10	0.7715	0.7376	0.744	0.7094	0.872	0.8532
11	0.7408	0.7387	0.7495	0.7013	0.7495	0.7094
12	0.8199	0.7586	0	0	0.7517	0.7121

Gambar 3 menunjukan perubahan pola sinyal *df-idf* secara cepat dan menghasilkan lonjakan nilai entropi wavelet. Seperti dapat dilihat bahwa ketiga kata tersebut memiliki pola spike yang serupa antara 10.000 hingga 20.000 sampel. Hal ini menunjukkan bahwa ketiga kata tersebut meningkat penggunaannya selama waktu tersebut. Dari sudut pandang deteksi event, peningkatan penggunaan kata menunjukkan bahwa kemungkinan besar ada peristiwa atau topik penting yang telah dibahas pada hashtag. Hasil ini sesuai dengan apa yang dikemukakan oleh [4, 5]. Dengan melakukan pencarian menggunakan google search engine, didapati bahwa kata "proud", "sophialimx", dan "goal" sebenarnya berasal dari tweets yang sama.



Gambar 4 Google Search untuk proud, sophialimx, dan goals

Hal ini menunjukan bahwa implementasi DF-IDF dan Entropi Wavelet dalam deteksi event pada satu hashtag akan menghasilkan sinyal dari kata-kata yang berada pada tweet yang sama. Dengan demikian pengguna Twitter terbantu untuk dengan mudah menemukan kejadian yang dibahas dalam hashtag tersebut dan juga untuk melacak *first story* peristiwa seperti penelitian yang dilakukan oleh [6].

Oleh karena itu, dengan penggunaan sinyal entropi wavelet, 1440 jumlah titik sampel dalam sinyal pertama hanya dapat diwakili oleh 12 sampel dalam entropi wavelet. Dengan demikian, hanya sedikit memori yang diperlukan untuk menyimpan sampel sinyal dan pada saat yang sama, masih dapat mendeteksi topik yang ditunjukkan oleh *spike* dalam periode tertentu.

4. KESIMPULAN

Penelitian ini telah berhasil mengimplementasi DWT dan DF-IDF di dalam mendeteksi kejadian dari sekumpulan tweets dengan tagar RIPLKY. df-idf berhasil membangun sinyal dari 20 kata yang dipilih setelah melalui proses filter. Setelah itu wavelet entropy digunakan untuk mendeteksi perubahan yang terjadi secara cepat pada sinyal df-idf. DWT berhasil mendeteksi perubahan secara cepat pada sinyal df-idf, dengan membuat *spike* (lonjakan nilai yang tinggi) pada nilai entropi wavelet. Tiga kata yang memiliki pola yang sama baik pada sinyal df-idf dan entropi wavelet adalah kata proud, goal, dan sophialimx.

Setelah dilakukan pencarian menggunakan google search engine, didapati bahwa ketiga kata tersebut berasal dari tweets yang sama dan menjelaskan peristiwa pemakaman perdana menteri Singapura. Yang perlu dikembangkan pada penelitian ini adalah pada pencarian deskripsi peristiwa yang masih manual menggunakan google search engine agar bisa dikembangkan melalui *coding* program yang baik untuk dapat melakukan search secara otomatis.

DAFTAR PUSTAKA

[1] Statista. (2020, 26 Juni 2020). *Number of Twitter users worldwide from 2014 to 2020(in millions)*. Available: https://www.statista.com/statistics/303681/twitter-users-worldwide/

[2] R. Kusumawardani and M. Basri, "Topic Identification and Categorization of Public Information in Community-Based Social Media," in *Journal of Physics: Conference Series*, 2017, p. 012075.

[3] M. Cordeiro, "Twitter event detection: combining wavelet analysis and topic inference summarization," in *Doctoral symposium on informatics engineering*, 2012, pp. 11-16.

[4] J. Weng and B.-S. Lee, "Event detection in twitter," in *Fifth international AAAI conference on weblogs and social media*, 2011.

[5] W. Dou, X. Wang, W. Ribarsky, and M. Zhou, "Event detection in social media data," in *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, 2012, pp. 971-980.

[6] S. Petrović, M. Osborne, and V. Lavrenko, "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*, 2010, pp. 181-189.

[7] Q. He, K. Chang, and E.-P. Lim, "Analyzing feature trajectories for event detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 207-214.

[8] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, *et al.*, "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," *Journal of neuroscience methods*, vol. 105, pp. 65-75, 2001.

[9] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, pp. 3-55, 2001.

[10] S. J. Orfanidis, *Optimum signal processing: an introduction*: Macmillan publishing company, 1988.