HASHTAG POPULARITY INDEX (HPI): ELIMINASI *SPAM* DI *HASHTAG*

Yumarsono Muhyi[1], Said Mirza Pahlevi[2]

[1]Jurusan Sistem Komputer, STMIK Indonesia Jakarta
[2]Badan Pusat Statistik, Jakarta
muhyi@stmik-indonesia.ac.id[1], **smirza@bps.go.id[2]**

Abstrak

Twitter yang menjadi sistem media sosial yang sangat populer saat ini, mengundang para *spammer* untuk mengekspliotasinya. Salah satu teknik *spam* yang paling sering dilakukan di Twitter adalah dengan membanjiri Twitter dengan *posting tweet* yang sangat banyak dan memasang *hashtag* yang sedang *trending* dengan jumlah yang banyak. Artikel ini memberikan sebuah metode baru bernama *Hashtag Popularity Index* (HPI), yang mampu mengeliminasi *spam* pada *hashtag* dan ekspliotasi *hashtag*. Dengan menggunakan fitur-fitur internal dari *tweet*, HPI dapat digunakan dan diimplementasikan sebagai alat yang handal dan *robust* dalam mengeliminasi *spam* pada *hashtag*.

Kata Kunci: *hashtag*, rekomendasi, *spam*, *tweet*, seleksi

Abstract

Twitter currently has become one the most popular social media system, is tempting spammers to exploit it. One of the spam technique being conducted in Twitter is by flooding Twitter with large amount of tweets and by putting many trending hashtags. This article gives a novelty named Hashtag Popularity Index (HPI) which is able to eliminate spams in hashtags and hashtag exploitation. Using internal features of tweets, HPI is suitable to use and to implement as a reliable and robust tool in eliminating spams in hashtag.

***Keywords*: hashtag, recommendation, spam, tweet, selection**

1. Pendahuluan

Twitter adalah salah satu dari sistem media sosial daring (dalam jaringan atau *online*) yang terpopuler saat ini. Twitter memiliki 313 juta akun pengguna yang aktif tiap bulannya dan selalu ada 500 juta pesan *tweet* baru setiap harinya [3].

Pada Juni 2012, lebih dari 500 juta akun pengguna terdaftar di Twitter, dengan Indonesia menduduki urutan ke-5 terbanyak di seluruh dunia untuk pengguna Twitter. Jakarta adalah kota teraktif dalam membuat *posting* di Twitter dengan lebih dari 24,3 juta *tweet* publik yang dikirimkan oleh para pengguna Twitter di Jakarta setiap harinya [8]. Untuk mempermudah pencarian *tweet*, seorang pembaca *tweet* disarankan memilih pesan-pesan *tweet* dengan menggunakan *hashtag* untuk membuang *tweet* yang tidak diinginkan. Sebagai hasilnya, bagi para penulis *tweet*, memilih *hashtag* yang tepat agar dapat meraih para pembaca yang ditujunya [11].

Popularitas Twitter yang tinggi ini mengundang para *spammer* untuk melakukan *posting* bersifat *spam* demi berbagai tujuan. Pada tahun 2014 ditengarai tidak kurang dari 8,5% akun pengguna Twitter (atau setara dengan 23 juta akun pengguna) yang diduga sebagai *spammer*, yang dicirikan dengan kegiatan mereka yang sangat rutin dan terpola [3]. Karena *hashtag* menjadi alat bagi Twitter untuk menetapkan *trending topics*, maka *hashtag* pun menjadi sarana ideal bagi para *spammer* untuk melancarkan kegiatan *spamming* mereka [9].

Karena *hashtag* merupakan sarana indeks utama dalam pesan-pesan *tweet*, maka *hashtag* tetap digunakan. Artikel ini bertujuan memberikan metode baru dalam menyaring *spam* dalam pencarian *hashtag*, agar *hashtag* tetap dapat berfungsi secara objektif dan optimal sesuai tujuan asalnya. Metode baru ini bernama *Hashtag Popularity Index* (HPI) yang dapat menyaring *spam* pada *hashtag* dengan sangkil (efektif) dan mangkus (efisien).

2. Penelitian Terkait

2.1. Rekomendasi *Hashtag*

Hashtag disarankan untuk digunakan oleh Twitter sebagai alat bantu bagi Twitter untuk mengindeks pesan-pesan *tweet*. Pemilihan *hashtag* yang digunakan memiliki algoritma heuristis tertentu yang disarankan oleh Twitter [11]. Pada umumnya penelitian-penelitian untuk merekomendasikan *hashtag* menggunakan metode *supervised* (*clustering* yang menggunakan masukan dari luar *tweet*) [1,4,5,12], *unsupervised* (klasifikasi yang menggunakan fitur-fitur internal dari *tweet*) [2,6,7], atau merupakan gabungan keduanya (*hybrid* atau *semi-supervised*).

2.2. Pengaruh *Spam* di Rekomendasi *Hashtag*

Penelitian [9] menunjukkan bahwa adanya *spam* di Twitter memiliki pengaruh cukup dominan dalam mengganggu proses rekomendasi *hashtag* pada *tweet*.

Pada penelitian ini diperoleh bahwa *hashtag* pada *spam tweet* memiliki akurasi rekomendasi lebih tinggi daripada pesan-pesan *tweet* yang benar (*ham tweet*).

Hal ini berarti bahwa suatu *tweet* akan berpeluang lebih tinggi mendapatkan rekomendasi *hashtag* yang tergolong *spam* daripada yang benar. Untuk meningkatkan peluang rekomendasi *hashtag* yang benar (non-*spam*) pada *ham tweet* ini, maka efek *spam tweet* ini perlu disaring dan dihilangkan.

2.3. Metode Deteksi *Spam* di Twitter

Penelitian [3] memaparkan metode-metode yang umum digunakan dalam deteksi *spam* di Twitter, dengan menggunakan fitur-fitur Twitter berikut.

1. Akun dari *tweet*
2. Isi pesan *tweet*
3. Grafik relasi antara *tweet* dan akun
4. Gabungan atau campuran (*hybrid*) dari beberapa fitur

Penelitian ini juga memaparkan hal-hal yang digunakan Twitter dalam mendeteksi *spam*.

1. *Posting tweet* yang serupa dengan sangat banyak, baik di multi-akun atau di akun tunggal;
2. Melakukan *follow* atau *unfollow* dalam jumlah besar dan dalam rentang waktu singkat;
3. Memiliki laporan *spam* yang sangat banyak atas akun tersebut;
4. Sangat agresif dalam melakukan *like*, *follow*, dan *retweet*;
5. Memasukkan tautan (*link*) yang *malicious* (cenderung berbahaya);
6. Melakukan *posting tweet* yang isinya kebanyakan adalah tautan, bukan *update* status personal;
7. Melakukan posting *tweet* ke sebuah *trending topic*, untuk melacak sistem deteksi *spam* apa yang sedang dilakukan.

Penelitian [10] mengungkapkan bahwa salah satu hal utama yang juga menjadi ciri *spammer* adalah mereka melakukan eksploitasi *hashtag* dengan masif. Sebanyak 76% *ham tweet* tidak memiliki *hashtag* dan 97% *ham tweet* memiliki maksimum 2 buah *hashtag*. Sebaliknya, sebanyak 40% s*pam tweet* memiliki 3 buah *hashtag* atau lebih dan hanya 37% yang tidak memiliki *hashtag*.

Penelitian [3] melakukan deteksi *spam* berdasarkan fitur-fitur yang kebanyakan terkait hal eksternal dari suatu *tweet.* Hal ini menyulitkan pelacakan *spam tweet* secara sederhana dan cepat. Dari penelitian [3] ini hanya poin pertama yang merupakan fitur internal dari *tweet*, yaitu bahwa *spammer* akan melakukan posting *tweet* dengan jumlah yang sangat besar.

Sementara dari penelitan [10] diperoleh hubungan sederhana antara *hashtag* dengan *tweet* yang berkategori *spam* dan *ham*, yang dapat digunakan sebagai pelacakan *spam tweet* secara sederhana dan cepat. Dari penelitian [10] ini dapat diperoleh hubungan antara *ham tweet* dengan *spam tweet*, berupa rasio dari jumlah *hashtag* antara keduanya itu sebagai berikut.

1. Tanpa *hashtag* = 76:37 = 2,1:1
2. Maksimal 2 *hashtag* = 97:60 = 1,6:1
3. Lebih dari 2 *hashtag* = 3:40 = 1:13,3

Dari sini dapat ditarik kesimpulan sederhana bahwa *tweet* yang memiliki lebih dari 2 buah *hashtag* memiliki peluang dikategorikan sebagai *spam tweet* sebesar 13,3 kali lebih tinggi daripada *ham tweet*. Jumlah *hashtag* lebih dari dua inilah yang akan digunakan sebagai indikasi awal dan cepat, bahwa sebuah *tweet* adalah *spam* atau *ham*.

3. *Hashtag Popularity Index* (HPI)

HPI merupakan metode penyaringan *hashtag spam* yang bersifat campuran (*hybrid*), dengan mengambil fitur-fitur *tweet* dari akun dan dari isi pesannya, untuk digunakan dalam kalkulasi.

2.1. Ekstraksi Fitur-Fitur Twitter

Dimisalkan T adalah kumpulan *tweet* dalam *corpus* (himpunan atau set dokumen) yang akan dianalisis dan H adalah himpunan *hashtag* unik yang telah diekstraksi dari semua *tweet* di T. Setiap *tweet* pada T dapat memiliki *hashtag* h sejumlah nol (0) atau lebih yang bersifat unik. Pemetaan (*mapping*) ini secara matematis dinotasikan sebagai $m(t,h) \in \{0,1\}$, fungsi ini bernilai 1 jika *tweett* memiliki *hashtag* h dan 0 jika tidak.

Dinotasikan U adalah himpunan akun pengguna (*user*) yang terdapat dalam *corpus*, maka $o\ (t,u) \in \{0,1\}$ adalah fungsi kepemilikan (*ownership*) dari *tweet*, bernilai 1 jika *tweett* dimiliki akun u dan 0 jika tidak. Notasi berikutnya f_u adalah jumlah *follower* (pengikut) dari akun u.

Kemudian dirumuskan jumlah *reader* (pembaca) dari akun u adalah $r_u = f_u + 1$, yang berarti bahwa semua *tweet* yang dibuat oleh u akan dibaca oleh semua *follower* u dan u sendiri. Lalu dirumuskan jumlah pembaca dari *hashtag* h adalah nilai maksimum dari jumlah pembacanya, yang dinotasikan sebagai berikut.

$$r_h = max(m(t,h) \times o(t,u).r_u) \quad (1)$$

2.2. Formulasi HPI

HPI bertujuan untuk melakukan penyaringan *spam* di *hashtag*, dengan cara menghilangkan beberapa faktor dari *spam*.

1. *Posting tweet* yang sangat banyak [3] dari akun spammer;

2. Menggunakan *hashtag* dengan sangat masif [10] dari akun spammer.

Faktor *posting tweet* yang sangat banyak ini dieliminasi dengan rumus *User-Tweet Ratio* (UTR) untuk setiap *hashtag*, yaitu jumlah akun total di *corpus* dibagi dengan jumlah *tweet* untuk *hashtag* tersebut. Dengan UTR ini diperoleh bahwa semakin banyak *tweet* yang yang menggunakan suatu *hashtag*, maka nilainya akan semakin rendah. UTR dinotasikan sebagai berikut.

$$UTR_h = \frac{|U|}{\sum^h m(t,h)} \quad (2)$$

Faktor eksploitasi *hashtag* dieliminasi dengan formula *Hashtag-User Support* (HUS), yaitu jumlah akun untuk *hashtag* tersebut dibagi dengan total akun di *corpus*. Dengan HUS ini semakin besar pengguna suatu *hashtag*, maka semakin tinggi pula nilainya. HUS itu dinotasikan sebagai berikut.

$$HUS_h = \frac{\sum^h (m(t,h) \times o(t,u))}{|U|} \quad (3)$$

Kedua rumus untuk mengeliminasi faktor eksploitasi *hashtag ini* digabungkan menjadi *User Ratio* (UR) per *hashtag* dapat dihitung menggunakan

$$UR_h = UTR_h . HUS_h \quad (4)$$

atau menjadi

$$UR_h = \frac{\sum^h (m(t,h) \times o(t,u))}{\sum^h m(t,h)} \quad (5)$$

Elimininasi faktor pembaca dari *spam* adalah dengan mengambil nilai logaritma dari r_h, yang kemudian hasilnya digabungkan dengan UR_h untuk mendapatkan nilai HPI, menjadi formula berikut.

$$HPI_h = UR_h \quad (6)$$

4. Percobaan dan Hasil

Percobaan dilakukan dengan mengambil *corpus* dari Twitter yang berbahasa Indonesia saja, dari 12 Juni 2016 sampai 18 Juni 2016, dengan mengambil *tweet* yang berasal dari area geografis Indonesia. Dari pengambilan data diperoleh 332.000 *tweet*, dan hanya 207.153 *tweet* yang benar-benar menggunakan Bahasa Indonesia.

Dari sini diperoleh 28.543 *tweet* yang memiliki *hashtag* di dalam teksnya, dan inilah yang dijadikan *corpus*. Total terdapat 18.049 hashtag yang dipetakan ke dalam 51.034 *tweet*. Dari data ini, diperoleh 10 hashtag dengan frekuensi tertinggi sebagai berikut.

1. #pathdaily: 1069 *tweet*
2. #dimanamacetid: 606 *tweet*
3. #latepost: 584 *tweet*
4. #np: 471 *tweet*
5. #kom8: 457 *tweet*
6. #jakarta: 374 *tweet*
7. #prillyatgemaramadhansctv: 348 *tweet*
8. #bukabarenguyatrans7ke15: 310 *tweet*
9. #ramadhan: 283 *tweet*
10. #bukber: 246 *tweet*

Dari seluruh 28.543 *tweet*, yang terindikasi sebagai *spam* ada sebanyak 4.778 tweet (16,7%), karena memiliki lebih dari 2 *hashtag*, dengan jumlah maksimumnya adalah 12. Dari sepuluh *hashtag* dengan frekuensi tertinggi, total *spam tweet* berjumlah 1.622 dan rata-rata tertimbang (*weighted average*) *spam* adalah 76,08%, dengan jumlah *spam* dan porsi *spam* per masing-masing *hashtag* adalah sebagai berikut.

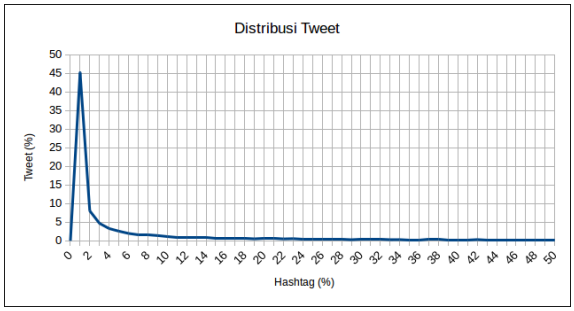
1. #pathdaily: 4 dan 0,37%
2. #dimanamacetid: 606 dan 100%
3. #latepost: 181 dan 30,99%
4. #np: 156 dan 33,69%
5. #kom8: 0 dan 0%
6. #jakarta: 364 dan 97,33%
7. #prillyatgemaramadhansctv: 16 dan 4,68%
8. #bukabarenguyatrans7ke15: 0 dan 0%
9. #ramadhan: 157 dan 55,48%
10. #bukber: 138 dan 56,10%

Setelah dilakukan eliminasi *spam* dengan HPI, diperoleh 10 *hashtag* tertinggi sebagai berikut.

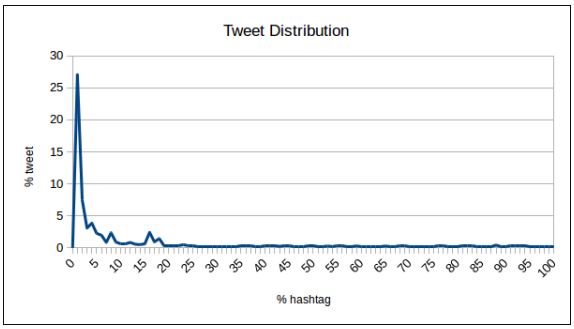
1. #pathdaily: HPI = 0,26989013
2. #latepost: HPI = 0,19814547
3. #bukber: HPI = 0,09700476
4. #ramadhan: HPI = 0,05199094
5. #likeforlike: HPI = 0,04496257
6. #euro2016: HPI = 0,03847262
7. #bukabarenguyatrans7ke15: HPI = 0,02946414
8. #repost: HPI = 0,02889165
9. #eng: HPI = 0,02579069
10. #like4like: HPI = 0,02566204

Dari kesepuluh *hashtag* setelah HPI ini, total *spam tweet* berjumlah 779 dan rata-rata tertimbang *spam* adalah 49,91%. Di sini terlihat bahwa setelah HPI *spam tweet* telah berhasil direduksi dan eksploitasi *hashtag* oleh *spammer* juga telah berhasil direduksi dengan sangat baik. Jumlah *spam* dan porsi *spam* per masing-masing *hashtag* adalah sebagai berikut.

1. #pathdaily: 4 dan 0,37%
2. #latepost: 181 dan 30,99%
3. #bukber: 138 dan 56,10%
4. #ramadhan: 157 dan 55,48%
5. #likeforlike: 93 dan 71,54%
6. #euro2016: 61 dan 34,46%
7. #bukabarenguyatrans7ke15: 0 dan 0%
8. #repost: 18 dan 8,7%
9. #eng: 41 dan 32,03%
10. #like4like: 86 dan 79,63%



Gambar 1. Distribusi *Hashtag* Sebelum HPI



Gambar 2. Distribusi *Hashtag* Setelah HPI

Distribusi *hashtag* terhadap prosentase *tweet* sebelum HPI ditampilkan pada Gambar 4.1. Dari gambar tersebut terlihat tingkat *short head* yang tinggi, yaitu 50% distribusi *tweet* tercapai hanya dengan 2% *hashtag* dengan frekuensi tertinggi pertama.

Distribusi *hashtag* terhadap prosentase *tweet* setelah dilakukan eliminasi *spam* dengan HPI ditampilkan pada Gambar 4.2. Distribusi *tweet* setelah dilakukan eliminasi *spam* dengan HPI megalami koreksi derajat *short head* yang cukup signifikan, dimana 50% *tweet* tertinggi pertama dicapai untuk *hashtag* tertinggi pertama sebanyak 10%.

5. Kesimpulan

HPI yang diajukan pada artikel ini secara tepat dapat mengeliminasi *spam* pada *hashtag* dari dua faktor utama serangan *spam* sebagai berikut.

1. *Posting tweet* yang sangat banyak dari para *spammer* [3], yang eliminasinya ditandai dengan menurunnya *spam tweet* yang terlibat dalam *hashtag* rekomendasi, dari 1.662 menjadi 779.
2. Mengekspliotasi *hashtag* dengan sangat masif [10], yang eliminasinya ditandai dengan penurunan nilai rata-rata tertimbang porsi *spam* dari *hashtag* rekomendasi, dari yang awalnya 76,08% menjadi 49,91%.

Dengan HPI ini, sistem rekomendasi *hashtag* yang disarankan oleh Twitter [11] dapat tetap dijalankan dengan konsistensi *hashtag* yang tinggi. Kalkulasi HPI ini dapat diimplementasikan dengan cepat dan *robust*, karena hanya mengambil fitur-fitur *tweet* yang internal dan inheren pada *tweet* itu sendiri, tanpa perlu intervensi dari fitur eksternal.

Daftar Pustaka

[1] Dovgopol, R., Nohelty, M., "*Twitter Hash Tag Recommendation*", Computing Research Repository, 2015.

[2] Godin, F., Slavkovikj, V., Neve, W.D., Schrauwen, B., Walle, R.V., "*Using Topic Models for Twitter Hashtag Recommendation. Proceedings of the 22nd International Conference on World Wide Web*", ACM, 2013.

[3] Kabakus, A.T., Kara, R., "*A Survey of Spam Detection Methods on Twitter*", (IJACSA) International Journal of Advanced Computer Science and Applications, 2017.

[4] Kywe, S.M., Hoang, T.A., Lim, E.P., Zhu, F. "*On Recommending Hashtags in Twitter Networks. Proceedings of the 4th International Conference on Social Informatics*", Berlin: Springer-Verlag, 2012.

[5] Mazzia, A., Juett, J. "*Suggesting Hashtags on Twitter*". Diakses pada 1 Juni 2016 dari http://www-personal.umich.edu/amazzia/pubs/545-final.pdf.

[6] Li T, Wu Y, Zhang Y. "Twitter Hash Tag Prediction Algorithm". Diakses pada 1 Juni 2016 dari http://cerc.wvu.edu/download/WORLDCOMP %2711/2011%20CD%20papers/ICM3338.pdf.

[7] Muhyi, Y., "*Rekomendasi Hashtag Untuk Tweet Berbahasa Indonesia*", Indonesia: STMIK Nusa Mandiri, 2016.

[8] Semiocast. "*Twitter Reaches Half a Billion Accounts*". Diakses pada 1 Juni 2016 dari http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US.

[9] Sedhai, S., Sun, A., "*Effect of Spam on Hashtag Recommendation for Tweets*", Canada: 25th World Wide Web Conference, 2016.

[10] Sedhai, S., Sun, A., "*HSpam14: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research*", The 38th International ACM SIGIR Conference, 2015.

[11] Twitter. "*How to Choose a Tag*". Diakses pada 1 Juni 2016 dari https://blog.twitter.com/2013/how-to-choose-a-hashtag.

[12] Zangerle, E., Gassler, W., Specht, G., "*Recommending #-Tags in Twitter. Proceedings of the Workshop on Semantic Adaptive Social Web 2011*", http://ceur-ws.org, 2011.