KINETIK, Vol.1, No.1, Mei 2016, Hal. 9-16
ISSN : 2503-2259,
E-ISSN : 2503-2267

Peringkasan Tweet Berdasarkan Trending Topic Twitter Dengan Pembobotan TF-IDF dan Single Linkage Angglomerative Hierarchical Clustering

Annisa[1], Yuda Munarko[2], Yufis Azhar[3]
[1,2,3]Universitas Muhammadiyah Malang
annisakawaii@ymail.com[1], yuda@umm.ac.id[2], yufis@umm.ac.id[3]

Abstrak

Fitur yang paling sering digunakan pada Twitter ialah Trending Topic. Trending Topic merupakan fitur yang menampilkan beberapa hashtag berisi topik yang sedang trend saat ini. Jika pengguna ingin mengetahui informasi mengenai suatu trending topic, pengguna bisa mengklik salah satu hashtag dan barulah muncul beberapa tweet terkait dengan hashtag tersebut. Agar menghemat waktu pengguna Twitter dalam membaca suatu trending topic tanpa perlu membaca beberapa tweet terlebih dahulu, maka dilakukanlah analisa dengan tujuan membuat text summarization untuk trending topic pada Twitter menggunakan algoritma TF-IDF dan Single Linkage Agglomerative Hierarchical Clustering. Penelitian ini menggunakan 100 trending topic untuk data tes pada sistem dan setiap trending topic terdiri atas 50 tweet berbahasa indonesia, sedangkan untuk pengujian digunakan 30 data trending topic diambil secara acak (data mewakili trending topic dengan sub tema minimal 2 dan maksimal 9 dari 100 data tes pada sistem). Dari 30 data pengujian, 1 data menghasilkan semua ringkasan sama persis dengan ahli, dan 29 data menghasilkan 1-4 ringkasan sama persis dengan ahli (terdiri atas 2-9 ringkasan untuk setiap trending topic).

Kata kunci:*Text Summarization, TF-IDF, Single Linkage Agglomerative Hierarchical Clustering*

Abstract

Trending topic is a feature provided by twitter that informs something widely discussed by users in a particular time. The form of trending topic is a hash tag and can be selected by clicking. However, the number of tweets for each trending topics can be very large, so it will be difficult if we want to know all the contents. So, in order to make easy when reading the topic, a small number of tweets can be selected as the main idea of the topic. In this study, we applied the Agglomerative Single Linkage Hierarchical Clustering by calculating the TF-IDF value for each words in advance. We used 100 trending topics, where each topics consists of 50 tweets in Indonesian. For testing, we provided 30 trending topics which consist of 2 until 9 sub topics. The result is that each trending topics can be summarized into shorter text contains 2 until 9 tweets. We were able to summarized 1 trending topics exactly same as the topic summarized by human expert. However, the rest of topics corresponded partially with human expert.

Keywords: *Text Summarization, TF-IDF, Single Linkage Agglomerative Hierarchical Clustering*

1. Pendahuluan

Salah satu faktor penting penunjang globalisasi ialah internet. Semakin majunya teknologi internet menyebabkan banyaknya pengembang perangkat lunak membuat berbagai macam aplikasi *online*, salah satunya yakni sosial media. Banyak sosial media yang saat ini sedang digandrungi, baik itu dari luar negeri maupun dalam negeri. Salah satu contoh sosial media yang sedang *trend* saat ini, yakni Twitter.

Pada jejaring sosial Twitter terdapat berbagai macam fitur, salah satu fitur yang paling sering digunakan ialah *Trending Topic*. *Trending Topic* merupakan fitur yang menampilkan beberapa *hashtag* yang berisi topik yang sedang *trend* saat ini di Twitter. Jika pengguna ingin mengetahui informasi atau berita mengenai salah satu *trending topic* tadi, pengguna bisa

Makalah dikirim 26 Februari 2016; Revisi 1 Mei 2016; Diterima 1 Juni 2016

mengklik salah satu *hashtag* dan barulah muncul beberapa *tweet* terkait dengan *hashtag* tersebut.

Hal tersebut tentulah merepotkan pengguna dan membuang waktu jika harus membaca beberapa *tweet* terlebih dahulu hanya untuk mengetahui informasi atau berita yang akurat pada suatu *trending topic*. Kalaupun hanya membaca *tweet* teratas atau terbaru mengenai *trending topic* tadi, belum tentu *tweet* tersebut mewakili berita atau informasi yang ingin disampaikan, karena biasanya beberapa *tweet* ada yang melenceng dari *hashtag* atau tema.

Berdasarkan permasalahan di atas mengenai fitur *trending topic*, untuk memudahkan dan menghemat waktu pengguna Twitter dalam mengetahui sebuah informasi atau berita pada salah satu *trending topic* tanpa perlu membaca beberapa *tweet* terlebih dahulu*,* maka dilakukanlah analisa untuk membuat ringkasan otomatis (*text summarization)* untuk *mult itweet* berdasarkan *trending topic* Twitter.*Text summarization* ialah suatu proses yang secara otomatis menghasilkan informasi berupa ringkasan yang berguna untuk *user* [1].

Berdasarkan literatur *"Pembangunan Perangkat Lunak Peringkas Dokumen Dari Banyak Sumber Menggunakan Sentence Scoring Dengan Metode TF-IDF",* bahwa untuk mengoptimalkan pemilihan kalimat yang digunakan sebagai ringkasan, akan lebih baik jika digabungkan dengan metode yang mencari kemiripan antar kata atau kalimat sehingga mengatasi resiko apabila terdapat kemiripan [2].

Biasanya dalam sebuah *trending topic* ada beberapa *tweet* yang melenceng dari *hashtag*, contohnya 5 data *tweet* dalam *trending topic* "PHK" pada Tabel 1 , sebagai berikut:

Tabel 1. Trending Topic PHK

Data	Tweet
T1	PHK Pekerja Toshiba
T2	Pekerja Toshiba kena PHK
T3	Jual Sony Xperia Z5 http://www.jualhp.sony/xperia/z5
T4	RT Pekerja Toshiba kena PHK
T5	Jual Sony Xperia Z5 http://www.jualhp.sony/xperia/z5

Berdasarkan Tabel 1 dapat lihat T1, T2, dan T4 membicarakan tentang PHK, sedangkan T3 dan T5 membicarakan tentang menjual hp. Oleh sebab itu, untuk mengatasi adanya sub topik yang muncul lebih dari satu dalam suatu *trending topic*, maka sebelum dilakukan peringkasan perlu klasterisasi data *tweet* terlebih dahulu.

Clustering adalah suatu teknik pengelompokkan data ke dalam objek yang serupa [3]. Ada dua metode *clustering* yang kita kenal, yaitu *hierarchical clustering* dan *partitioning*. Pada penelitian ini digunakan metode *hierarchical clustering* karena tidak memerlukan *input* berapa jumlah *cluster* data, sebab pada sebuah *trending topic* tidak diketahui berapa jumlah *tweet* yang akan melenceng dari *hashtag*.

Penelitian ini bertujuan untuk membuat sebuah sistem yang dapat menghasilkan ringkasan secara otomatis yang sesuai dengan *multi tweet* berdasarkan *trending topic* dengan menggunakan algoritma TF-IDF dan *Single Linkage Agglomerative Hierarchical Clustering*.

Batasan masalah untuk penelitian ini, yaitu *tweet* berbahasa Indonesia, setiap *trending topic* mengambil *sample* 50 *tweet* untuk dilakukan peringkasan, dan data *test* untuk sistem terdiri atas 100 *trending topic* yang dikumpulkan mulai tanggal 2 Februari 2016 – 13 Maret 2016.

Perbedaan antara penelitian ini dengan penelitian sebelumnya *"Pembangunan Perangkat Lunak Peringkas Dokumen Dari Banyak Sumber Menggunakan Sentence Scoring Dengan Metode TF-IDF"* adalah penelitian sebelumnya membangun sistem *text summarization* untuk dokumen dari banyak sumber dan tidak ada tahapan *clustering* didalam sistemnya sedangkan pada penelitian ini membangun sistem *text summarization* untuk *multi tweet* dan terdapat tahapan *clustering* sebelum dilakukan peringkasan *tweet*.

2. Metode Penelitian

Metode *text summarization* yang digunakan pada penelitian ini ialah metode *extractive summary.* Teknik ini meringkas dengan cara menyalin unit-unit teks yang dianggap paling penting dan dapat mewakili teks sumber menjadi ringkasan.

Tahapan pada sistem *text summarization* untuk *trending topic* pada Twitter dalam bahasa Indonesia terdiri atas 4 tahapan, sebagai berikut:

1. Tahapan pertama mengumpulkan data *tweet* melalui *web crawler* dengan memanfaatkan API Twitter. Setelah dokumen tersimpan dilakukan *preprocessing* yang terdiri atas 7 tahapan, sebagai berikut:
   a. Pemecahan *tweet* yaitu proses memecah *string* teks dokumen yang panjang menjadi kumpulan *twee*t. Dalam memecah dokumen menjadi beberapa *tweet* menggunakan fungsi *split*( ),dengan tanda baris baru sebagai *delimiter* untuk memotong *string* dokumen.
   b. *Case folding* ialah tahapan mengubah semua huruf pada *tweet* menjadi huruf kecil, menghilangkan karakter angka, dan menghilangkan *delimiter* seperti (.), (,), (:), (;), (?), dan (!).
   c. *Tokenizing* ialah tahapan penguraian deskripsi dari *tweet* menjadi kata-kata dengan pemisah spasi.
   d. *Editing* ialah tahapan mengubah kata-kata yang disingkat menjadi kata aslinya dan mengubah kata tidak baku menjadi kata baku.
   e. Tahapan untuk menghilangkan kata-kata yang dianggap tidak penting (*stopwords*), seperti kata di, ke, apa, dan, ini, itu, pun, agar, akan, dari, juga, oleh, pada, yang, bahwa, dapat, namun, untuk, dengan, kepada, ya, nya.
   f. Normalisasi ialah tahapan menghilangkan *tag-tag* yang tidak penting biasanya muncul pada *tweet*, seperti *link* diawali dengan http, mentioned diawali dengan simbol @, *hashtag* diawali dengan simbol #, dan *retweet* yg dilambangkan dengan RT diawal *tweet*.
   g. *Stemming* ialah tahapan pemetaan dari bentuk (*variants*) menjadi kata dasar.

   Pada penelitian ini tahapan *editing*, normalisasi, dan *stemming* dilakukan secara manual.
2. Tahapan kedua yaitu perhitung bobot kata dengan TF-IDF. TF(*Term Frequency*) ialah frekuensi banyaknya kata yang muncul pada sebuah dokumen, sedangkan IDF ialah kemunculan *term* atau kata pada kumpulan dokumen. Persamaan IDF adalah sebagai berikut.

$$IDF(t) = \log \frac{N}{df(t)} \quad (1)$$

   Dimana *df(t)* adalah banyak dokumen yang mengandung *term t*. TF*IDF merupakan kombinasi metode TF dengan metode IDF. Sehingga persamaan TF*IDF adalah sebagai berikut:

$$TF - IDF(d,t) = TF(d,t) * IDF(t) \quad (2)$$

3. Tahapan ketiga yaitu klasterisasi data dari hasil perhitungan bobot TF-IDF menggunakan metode *Single Linkage Agglomerative Hierarchical Clustering*.

   Cara kerja algoritma *Single Linkage Agglomerative Hierarchical Clustering* menggunakan prinsip jarak minimum yang diawali dengan mencari dua objek terdekat dan keduanya membentuk *cluster* yang pertama. Pada langkah selanjutnya terdapat dua kemungkinan, yaitu :
   a. Objek ketiga akan bergabung dengan *cluster* yang telah terbentuk, atau
   b. Dua objek lainnya akan membentuk *cluster* baru.

   Proses ini akan berlanjut sampai akhirnya terbentuk *cluster* tunggal. Data yang digunakan untuk menghitung *cluster* ialah *Euclidean Matrix* yang didapatkan berdasarkan persamaan berikut:

$$EuclideanDis\tan ce(X,Y) = \sqrt{\sum (Xi - Yi)^2} \quad (3)$$

   Pada penelitian ini tahap penggabungan *cluster* (memilih *level cluster* terbaik), menggunakan ukuran *dissimilarity* antar *cluster* dengan UPGMA *(Unweighted Pair Group Method Average)* [4,5], ditunjukkan pada Persamaan 4.

$$Dissimilarity(cluster1, cluster2) = \frac{\sum Euclidean(d1, d2)}{sizecluster1 * sizecluster2} \quad (4)$$

Teknik UPGMA pada penelitian ini dilihat dari ketidak miripan semua dokumen dari *cluster,* untuk nilai tengah *cluster* yang didefinisikan dengan Persamaan 5

$$Sim(X) = \sum_{dex} Euclidean(d, c) \quad (5)$$

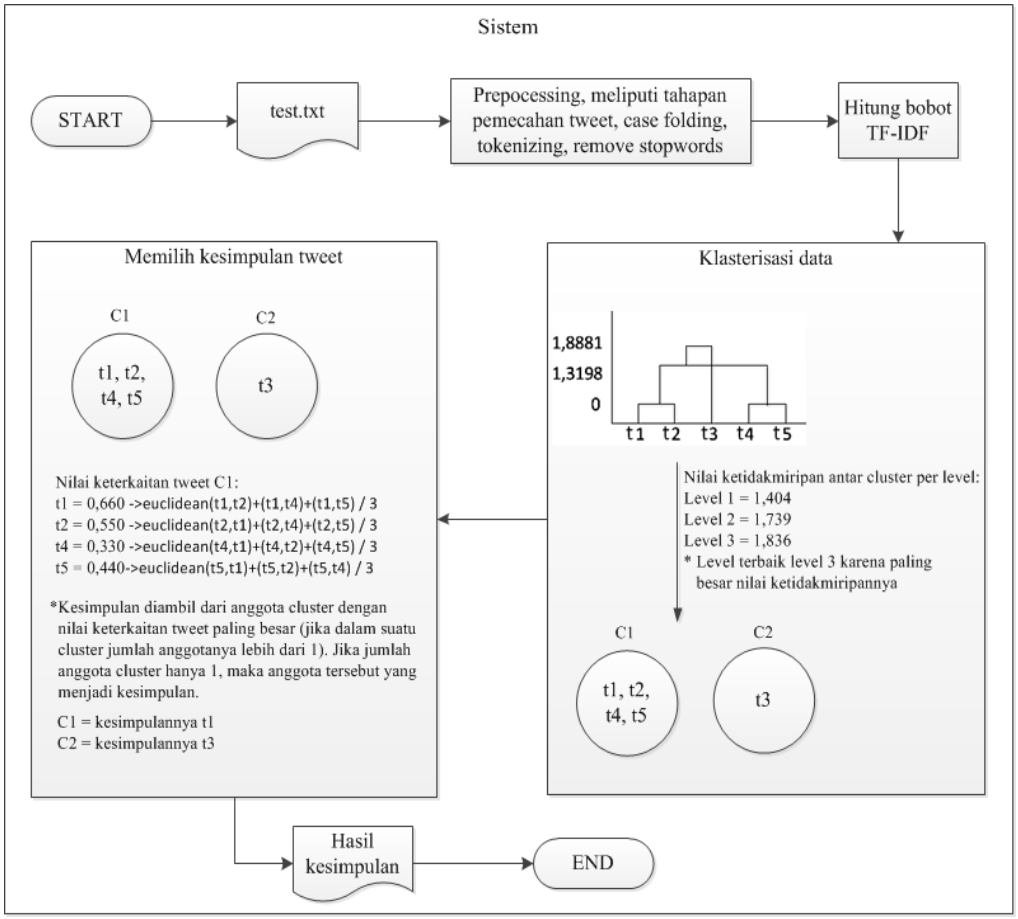
dimana *d* ialah dokumen pada *clusterX*, dan *c* ialah nilai tengah dari *cluster X* yang merupakan *mean* (nilai tengah) dari *vector* dokumen. Pemilihan untuk penggabungan *cluster* terbaik (memilih *level cluster* yang tepat) berdasarkan nilai terbesar dari Persamaan 6.

4. Tahapan keempat menentukan ringkasan untuk setiap *cluster*. Jika jumlah anggota *cluster* 1, maka anggota tersebut yang menjadi ringkasan. Sedangkan saat jumlah anggota *cluster* lebih dari 1, maka yang menjadi ringkasan ialah salah satu anggota *cluster* dengan nilai keterkaitan antar *tweet*ter kecil dengan persamaan nilai sebagai berikut:

$$nilai = \frac{\sum EuclideanDis\tan ce}{n} \quad (6)$$

Jumlah kesimpulan yang muncul sama dengan jumlah *cluster* atau sub topik yang ada pada sebuah *trending topic*.

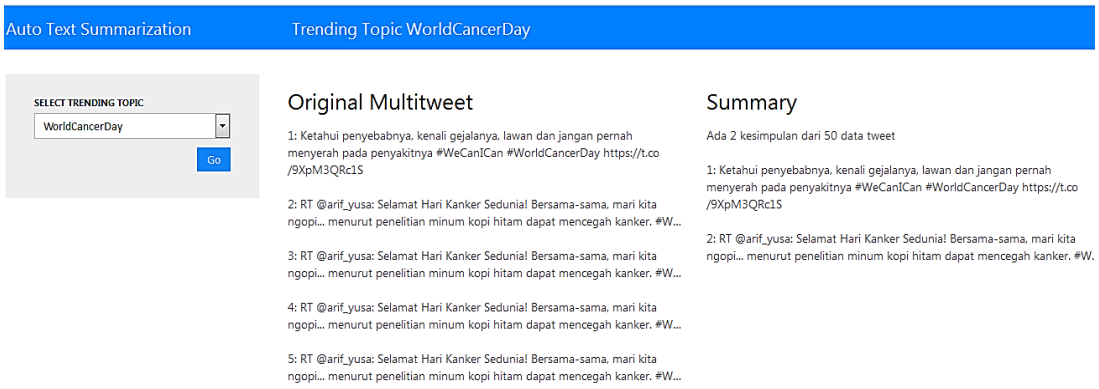
Berikut contoh gambaran sistem untuk peringkasan *trending topic* tentang “tes” pada Gambar 1:



Gambar 1. Gambaran Sistem

3. Hasil Penelitian dan Pembahasan

Berikut tampilan antarmuka halaman hasil ringkasan dari sistem *text summarization* untuk *trending topic* pada Twitter dalam bahasa Indonesia:



Gambar 2. Halaman Hasil Ringkasan

Pengujian terhadap sistem *text summarization* untuk *trending topic* pada Twitter dalam bahasa Indonesia ini ada dua, yaitu pengujian terhadap *cluster* dan pengujian terhadap hasil ringkasan *tweet*.

Data yang digunakan untuk pengujian adalah 30 data *trending topic* diambil secara acak (data mewakili *trending topic* dengan sub tema minimal 2 dan maksimal 9 dari 100 data tes pada sistem).

Pertama, pengujian *cluster* dengan *F-Measure,* merupakan metode pengujian yang menghubungkan antara nilai *precision* dan *recall,* dirumuskan sebagai berikut:

$$F-Measure = \frac{2xprecision \times recall}{precision + recall} \quad (7)$$

Skenario pengujian *cluster* pada penelitian ini adalah dengan membandingkan hasil *cluster* yang dibuat secara manual oleh ahli dengan hasil *cluster* dari sistem. Jumlah *cluster* telah ditentukan terlebih dahulu sesuai dengan keluaran dari sistem. Saat nilai *F-Measure* mendekati atau mencapai 100% , maka semakin mirip *cluster* yang dihasilkan sistem dengan yang ditentukan ahli.

Berdasarkan pengujian *cluster*yang ditunjukkan pada Tabel 2. dapat diambil kesimpulan sebagai berikut:

1. Dari 30 data uji, hanya satu yang semua klasterisasi datanya sama persis dengan ahli, yaitu *trending topic* “WorldCancerDay”.
2. Ada 1 data uji pada *trending topic* “Berhenti Langganan IndiHome” semua klasterisasinya sangat tidak mirip atau jauh berbeda dengan yang ditentukan ahli.
3. Ada 2 data uji, yaitu *trending topic* “Chris Martin” dan “HitzSirkusPagi” memperoleh satu klaterisasi data yang sama persis dengan yang ditentukan ahli (terdiri atas 2-9 *cluster* untuk setiap *trending topic*).
4. Sisa data uji 26 *trendingtopic*, rata-rata semuanya hanya mendapatkan satu klasterisasi yang hampir mirip dengan yang ditentukan ahli (terdiri atas 2-9 *cluster* untuk setiap *trending topic*).
5. Data *cluster* atau sub tema yang mirip dengan ahli dihasilkan saat terdapat banyak data *retweet* pada suatu *trending topic*.
   Contohnya pada pengujian ini terdapat nilai *F-Measure* 100% semua untuk setiap cluster atau sub tema pada *trending topic* “WorldCancerDay”, artinya semua hasil *cluster* yang dihasilkan oleh sistem dengan yang dihasilkan oleh ahli sama persis. Hasil tersebut bisa didapatkan karena, pada *trending topic* “WorldCancerDay” hanya terdapat 2 *tweet* yang berbeda, seperti yang ditunjukan pada Tabel 3.

Tabel 2. F-Measure Persen

No	Trending Topic	Jlh Cluster	F-Measure								
			Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
1	BSM Tabungan Berencana	2	6%	56%							
2	WorldCancerDay	2	100%	100%							
3	Ciledug	2	6%	56%							
4	dahSyatKamisKustik	2	5%	31%							
5	Fatwa MUI	3	0%	0%	68%						
6	HappyKyuDay	3	0%	0%	72%						
7	Niall is in Bali	3	0%	0%	89%						
8	Ahmadiyah	4	13%	10%	0%	35%					
9	ANAKJALANANSPESIAL PART 3	4	0%	18%	0%	60%					
10	AskDokArt	4	0%	13%	0%	73%					
11	HTM 10k	4	0%	0%	29%	69%					
12	Luais Suarez	4	20%	0%	18%	62%					
13	NOTRECOM	4	0%	0%	0%	63%					
14	DOAQU BUAT SINGLE IRWANDA2	4	67%	15%	15%	68%					
15	BreakoutNET_Band	4	33%	0%	13%	65%					
16	BestCutOfPiyuDiPro2FM	4	15%	0%	42%	0%					
17	Chris Martin	5	100%	15%	0%	0%	54%				
18	2.500 Pekerja Kena PHK	5	11%	0%	50%	0%	61%				
19	7ManusiaHariamueps569	5	40%	0%	0%	0%	72%				
20	AdaSHanisaJKT48LagiDiTMG	5	0%	33%	10%	0%	55%				
21	BANG IPUL DA3 IVM	5	33%	0%	0%	0%	56%				
22	BELAHANJIWAKAHRAMAN55	5	14%	0%	0%	0%	64%				
23	70tahunPMI	6	0%	17%	40%	0%	0%	70%			
24	DaruratLGBT	6	20%	0%	0%	0%	0%	58%			
25	AkhimyaRAISAkeIniTalkShow	7	40%	25%	22%	0%	0%	0%	60%		
26	Bintang	7	25%	40%	17%	0%	29%	0%	34%		
27	HitzSirkusPagi	7	100%	18%	0%	0%	0%	0%	70%		
28	Kick Andy	7	0%	0%	17%	0%	36%	20%	0%		
29	Berhenti Langganan IndiHome	8	40%	0%	33%	33%	0%	0%	0%	44%	
30	Gemini	9	20%	0%	67%	0%	0%	67%	0%	0%	43%

Keterangan:

= sama persis (semua komposisi anggota cluster sama)

= hampir mirip(lebih dari setengah komposisi anggota *cluster* sama)

Kedua, pengujian ringkasan menggunakan metode intrinsik *ROUGE-N* [6,7].Pengukuran *ROUGE-N* mengukur perbandingan *N-gram* dari dua ringkasan, dan menghitung berapa jumlah yang sesuai, dirumuskan dalam Persamaan 7:

Tabel 3. Retweet

Data Tweet	Tweet
1	Ketahui penyebabnya, kenali gejalanya, lawan dan jangan pernah menyerah pada penyakitnya #WeCanICan #WorldCancerDay https://t.co/9XpM3QRc1S”
2-50	RT @arif_yusa: Selamat Hari Kanker Sedunia! Bersama-sama, mari kita ngopi... menurut penelitian minum kopi hitam dapat mencegah kanker. # WorldCancerDay

$$ROUGE-N = \frac{\sum_{S} \in summ_{ref} \sum N-gramm \in_{s} count_{match}(N-gram)}{\sum_{S} \in summ_{ref} \sum N-gramm \in_{s} count(N-gram)} \quad (8)$$

Pada Persamaan 8, notasi N menunjukkan panjang dari *N-gram*, $Count_{match}$*(N-gram)* adalah jumlah maksimum dari *N-gram* yang muncul pada ringkasan kandidat dan ringkasan sebagai referensi. *Count(N-gram)* adalah jumlah dari *N-gram* pada ringkasan sebagai referensi.

Skenario pengujian ringkasan *tweet* pada penelitian ini adalah dengan membandingkan hasil ringkasan manual yang dibuat oleh ahli dengan hasil ringkasan yang dihasilkan oleh sistem.Pertama ahli melakukan klasterisasi data *tweet* suatu *trending topic* secara manual dengan jumlah *cluster* yang telah ditentukan terlebih dahulu. Kemudian berdasarkan *cluster* yang terbentuk tadi, ahli memilih salah satu *tweet* dalam setiap *cluster* untuk mewakili anggotanya yang lain, *tweet* yang dipilih untuk mewakili *cluster* tersebutlah yang dijadikan pembanding antara kesimpulan yang dihasilkan oleh sistem dengan kesimpulan yang ditentukan oleh ahli.

Pada penelitian ini fungsi *ROUGE-N* yang digunakan adalah *ROUGE* dengan nilai N = 1. *ROUGE* dengan nilai N=1 berarti membandingkan kesamaan hasil ringkasan dengan ringkasan referensi untuk setiap satu kata. Sebelum dilakukan pengukuran dengan *ROUGE-1*, hasil ringkasan *tweet* dilakukan preproses data terlebih dahulu, yang terdiri atas tahapan *case folding*, *tokenizing*, *editing*, menghilangkan *stopwords*, normalisasi, dan *stemming*.

Berdasarkan hasil pengujian ringkasan pada Tabel 5. dapat diambil kesimpulan sebagai berikut:

1. Dari 30 data uji, hanya satu yang semua kesimpulan untuk setiap sub tema sama persis dengan ahli, yaitu *trending topic* “WorldCancerDay”.
2. Selain data uji *trending topic* “WorldCancerDay”, sisa 29 data uji rata-rata menghasilkan 1-4 kesimpulan yang sama persis dengan ahli (terdiri atas 2-9 kesimpulan untuk setiap *trending topic*).
3. Diambil contoh salah satu data *trending topic* dari pengujian dengan nilai *ROUGE-1* paling kecil (memperoleh nilai 0), yaitu data dengan *trending topic* “dahSyatkamisKustik”. Hal tersebut bisa terjadi karena, ringkasan yang dihasilkan oleh sistem dengan yang ditentukan ahli memang berbeda jauh dari segi padanan kata dan artinya, seperti ditunjukan pada Tabel 4. sebagai berikut:

Tabel 4. Beda Ringkasan

Ringkasan	
Ahli	**Sistem**
RT @MusicaStudios: Jangan lupa nih yang lagi nonton @GEISHAIndonesia #dahsyatKamisKustik download lagunya di https://t.co/TAyig81SKG	#dahSyatKamisKustik Kalau kaum pria berkumpul, mereka saling mendengarkan, sedangkan kalau kaum wanita berkumpul, mereka saling mengamati.

4. Kesimpulan

Berdasarkan sistem yang telah dibuat dan pengujian yang telah dilakukan, dapat diambil dua kesimpulan, yaitu pertama *cluster* yang baik atau bagus terbentuk dari *trending topic* yang banyak mengandung data *retweet*. Dari 30 data uji ada 3 data *trending topic* dengan hasil salah satu *F-Measurecluster*-nya mencapai 100% berdasarkan Tabel 2. (hasil *cluster* keluaran sistem sama persis dengan *cluster* yang ditentukan ahli), yaitu *trending topic* “WordCancerDay”, “Chris Martin” dan “HitzSirkusPagi”.

Kedua, perbedaan antara komposisi anggota *cluster* dari sistem dengan komposisi anggota *cluster* dari ahli, masih memungkinkan pemilihan *tweet* yang sama sebagai ringkasan. Hal tersebut dipengaruhi oleh subjektifitas ahli dalam menentukan ringkasan.

Beberapa saran yang dapat dilakukan untuk pengembangan sistem ini, yaitu pertama menambahkan algoritma yang dapat memilih level cluster terbaik dan lebih efektif selain dengan UPGMA*(Unweighted Pair Group Method Average)*. Kedua, menambahkan algortima yang dapat menghasilkan ringkasan *tweet* dengan teknik abstraktif, yaitu menghasilkan ringkasan dengan kalimat-kalimat baru yang mewakili intisari dari setiap anggota *cluster* atau

sub tema. Ketiga, menambahkan *preprocessing* untuk tahapan *editing*, normalisasi, dan *stemming* dalam sistem, karena pada penelitian ini ketiga tahapan tersebut dilakukan secara manual.

Tabel 5. ROUGE-1

No	Trending Topic	Jlh Cluster	F-Measure								
			Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
1	BSM Tabungan Berencana	2	0,333	1,000							
2	WorldCancerDay	2	1,000	1,000							
3	Ciledug	2	1,000	0,056							
4	dahSyatKamisKustik	2	0,000	1,000							
5	Fatwa MUI	3	0,214	0,190	0,400						
6	HappyKyuDay	3	0,211	0,000	1,000						
7	Niall is in Bali	3	0,235	0,276	1,000						
8	Ahmadiyah	4	0,067	0,057	0,067	0,333					
9	ANAKJALANANSPESIAL PART 3	4	0,400	0,276	0,296	1,000					
10	AskDokArt	4	0,100	0,091	0,000	0,000					
11	HTM 10k	4	0,182	1,000	0,167	1,000					
12	Luais Suarez	4	0,173	0,182	1,000	1,000					
13	NOTRECOM	4	0,183	0,118	0,125	1,000					
14	DOAQU BUAT SINGLE IRWANDA2	4	0,588	1,000	0,556	1,000					
15	BreakoutNET_Band	4	0,000	0,000	0,000	0,463					
16	BestCutOfPiyuDiPro2FM	4	0,000	0,000	0,333	0,061					
17	Chris Martin	5	1,000	0,286	0,000	0,200	1,000				
18	2.500 Pekerja Kena PHK	5	0,071	0,267	0,273	0,214	1,000				
19	7ManusiaHariamueps569	5	0,000	0,087	0,069	0,000	0,000				
20	AdaSHanisaJKT48LagiDiTMG	5	0,000	0,000	0,000	0,000	1,000				
21	BANG IPUL DA3IVM	5	1,000	0,000	0,387	0,400	1,000				
22	BELAHANJIWAKAHRAMAN55	5	0,000	0,000	0,000	0,000	0,083				
23	70tahunPMI	6	0,000	0,000	0,167	0,286	0,000	0,667			
24	DaruratLGBT	6	0,182	0,364	0,345	0,143	0,174	1,000			
25	AkhimyaRAISAkeIniTalkShow	7	0,083	0,000	0,125	1,000	1,000	0,105	1,000		
26	Bintang	7	1,000	1,000	0,000	0,087	0,182	0,083	0,400		
27	HitzSirkusPagi	7	1,000	0,250	0,000	0,000	0,000	0,000	1,000		
28	Kick Andy	7	0,191	0,194	0,133	0,462	1,000	1,000	0,250		
29	Berhenti Langganan IndiHome	8	1,000	0,348	0,353	1,000	1,000	0,348	0,361	1,000	
30	Gemini	9	0,261	0,207	0,083	0,111	0,000	0,077	0,125	0,111	1,000

Keterangan:

= sama persis(ringkasan sistem dengan ahli sama persis)

Refrensi

[1] Erkan, Günes, dan Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." *Journal of Artificial Intelligence Research* (2004): 457-479.

[2] Móro, Róbert, dan M. Bielikov. "Personalized text summarization based on important terms identification." *Database and Expert Systems Applications (DEXA), 2012 23rd International Workshop on*. IEEE, 2012.

[3] Berkhin, Pavel. "A survey of clustering data mining techniques." *Grouping multi dimensional data*. Springer Berlin Heidelberg, 2006. 25-71.

[4] Hamzah, Amir, F.Soesianto, dan Jazi Eko Istiyanto . "Studi Kinerja Fungsi-Fungsi Jarak dan Similaritas dalam Clustering Dokumen Teks Berbahasa Indonesia." *Seminar Nasional Informatika (SEMNASIF)*. Vol. 1. No. 1. 2015.

[5] Steinbach, Michael, George Karypis, dan Vipin Kumar. "A comparison of document clustering techniques." *KDD workshop on text mining*. Vol. 400. No. 1. 2000.

[6] LIN, C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of Workshop on Text Summarization Brances Out*.

[7] Santika, Putu Praba, and Gus Nanang Syaifuddin. "Semantic Clustering Dan Pemilihan Kalimat Representatif Untuk Peringkasan Multi Dokumen." *Jurnal Teknologi Informasi dan Ilmu Komputer* 1.2 (2015).