REPOSITOR, Vol. 2, No. 6, Juni 2020, Pp. 815-822
ISSN : 2714-7975
E-ISSN : 2716-1382

Metode Cluster Importance Untuk Peringkasan Dokumen Pada Data Tweet Berbahasa Indonesia (Study Kasus Pilkada Dki Jakarta)

Dyah Hestiningtyas[*1], Nur Hayatin[2], Yuda Munarko[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
dhestiningtyas@gmail.com[*1],noorhayatin@gmail.com[2],yuda@umm.ac.id[3]

Abstrak

Pemanfaatan media elektronik sebagai media informasi berkembang sangat pesat di era sekarang. Terbukti meningkatnya jumlah informasi dan data. Banyaknya data yang ada diharapkan dapat memberikan manfaat yang banyak pula. Automatic Text Summarization merupakan salah satu proses peringkasan teks dokumen yang dilakukan secara otomatis melalui mesin komputer. Pada penelitian ini penulis membahas tentang peringkasan dokumen pada tweet, dimana data yang digunakan dengam mengumpulkan tweet melalui web crawler dengan memanfaatkan API Twitter. Pada paper ini diajukan sebuah metode Cluster Importance dan melakukan pemilihan tweet representatif pada setiap cluster berdasarkan bobot tweet terpenting pada suatu cluster. Yang nantinya bobot tweet tertinggi dipilih sebagai tweet penyusun ringkasan. Penelitian ini menggunakan 25 topik. hasil dari perhitungan Rouge-N yaitu: Dari 25 topik data uji ada 3 topik yang mempunyai Iterasi 1 yaitu topik HUT DKI, Kinerja Djarot, dan Pemimpin dikarenakan jumlah tweet pada topik tersebut sama - sama memiliki 2 tweet. Pengujian perbandingan antara hasil manual dan hasil sistem menunjukan hasil 100% pada topik Dukungan. Perhitungan Rouge-N menyimpulkan bahwa sistem dapat merangkum tweet minimal 3 tweet, dari perbandingan hasil sistem dan manual di dapatkan tweet minimal mempunyai nilai 96% yang di mana sistem bekerja dengan baik, serta pada topik yang memiliki 18 tweet menunjukan hasil 89%.

Kata Kunci: *Cluster Importance, Twitter, Peringkasan Tweet*

Abstract

The use of electronic media as an information media is developing very rapidly in the present era. Proven increasing amount of information and data. The amount of data available is expected to provide many benefits too. Automatic Text Summarization is one of the process of summarizing document text that is automatically through a computer machine. In this study the author discusses about the summarization of documents on tweets, where the data is used by collecting tweets through web crawlers by utilizing the Twitter API. This paper proposes a Cluster Importance method and selected representative tweets on each cluster based on the weight of the most important tweets in a cluster. The highest tweet will be selected as summary compiler tweets. This study uses 25 topics. the results of Rouge-N are: in the 25 topics of the test data there are 3 topics that have Iteration 1 namely the topic of HUT DKI, Kinerja Djarot, and Pemimpin *because the same number of tweets on the topic has 2 tweets. Testing the comparison between manual results and system results shows 100% results on the topics* Ahok, Korupsi, Dukungan, and Pajak.

Keywords: *Cluster Importance, Twitter, Tweet Summarization*

1. Pendahuluan

1.1 Latar Belakang

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. Salah satu contoh sosial media yang sedang *trend* saat ini yaitu *Twitter*. Pada jejaring sosial *Twitter* terdapat berbagai macam fitur, salah satunya ialah *Trending Topic. Trending Topic* merupakan fitur yang menampilkan beberapa hashtag yang berisi topik yang sedang trend saat ini di Twitter [1].

Automatic Text Summarization adalah proses peringkasan dokumen teks yang dilakukan secara otomatis melalui mesin komputer. Otomatisasi ringkasan dapat dikenakan terhadap satu dokumen (*single document summarization*) atau beberapa dokument (*multi-document*

summarization) [2]. Ringkasan yang baik merupakan ringkasan yang mampu mencakup (*coverage*) sebanyak mungkin konsep – konsep penting (*saliency*) yang ada pada dokumen sumber [3].

Coverage dan *saliency* adalah masalah utama dalam metode peringkasan dimana strategi pemilihan kalimat pada metode extraktive menjadi sangat penting karena harus mampu memilih kalimat – kalimat utama (penting) dan terhindar dari redudansi (*redudancy*) sehingga mencakup banyak konsep. Salah satu fase penting yang banyak digunakan pada peringkasan secara ekstraktif adalah fase pembobotan kalimat (*sentence scoring*) [3].

Pada kasus sebelumnya telah dilakukan penelitian seperti berikut: melakukan peringkasan dokumen pada tweet berdasarkan trending topic dan mencari kemiripan antar kalimat dengan metode TF-IDF dan *Single Linkage Anggromerative Hierarchical Clustering* [1], serta penelitian yang membahas peringkasan multi dokumen berita menggunakan *Cluster Importance* [2].

Merujuk pada penelitian – penelitian sebelumnya, ide dari penulisan ini yaitu melakukan peringkasan tweet menggunakan metode *Cluster Importance* dan melakukan pemilihan tweet representatif pada setiap *cluster* berdasarkan bobot tweet terpenting pada suatu *cluster*. Yang nantinya bobot *tweet* tertinggi dipilih sebagai *tweet* penyusun ringkasan. Dari kombinasi teknik pembobotan tersebut diharapkan dapat menyeleksi *tweet* secara lebih efisien, sehingga mampu menghasilkan peringkasan *tweet* dengan lebih koheren.

1.2 Rumusan Masalah

Sesuai dengan latar belakang yang dipaparkan diatas, maka rumusan masalah dalam penelitian ini, adalah cara melakukan peringkasan *multi tweet* dengan menggunakan metode *Cluster Importance*.

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dalam pembuatan Tugas Akhir ini adalah melakukan peringkasan *multi tweet* dengan menggunakan metode *Cluster Importance*.

1.4 Batasan Masalah

Dalam penelitian ini, penulis membatasi masalah sebagai berikut :

1. Tweet menggunakan bahasa Indonesia.
2. Uji coba dilakukan pada data tweet Pilkada DKI Jakarta.

2. Metodelogi

Penelitian yang dilakukan untuk merancang sistem diperoleh dari pengamatan data-data yang ada. Tahap-tahap yang dilakukan untuk penelitian guna perancangan sistem tersebut secara terstruktur adalah:

1. Studi pustaka

Tahapan untuk memahami konsep dari pembangunan sistem berupa jurnal, buku referensi, artikel, tutorial, dan sumber yang berkaitan *twitter, text summarization, cluster importance*, dan teknik pengujian *ROUGE*.

2. Pengumpulan data

Tahapan ini merupakan proses pengumpulan data yang diperoleh dari proses *crawling* data melalui web *crawler* untuk mengakses API dari *twitter* [4, 5]. *Tweets* yang di ambil berdasarkan *trending issue* yang sedang berkembang (*hot topic*) pada periode waktu yang sama.

3. Analisa dan perancangan sistem

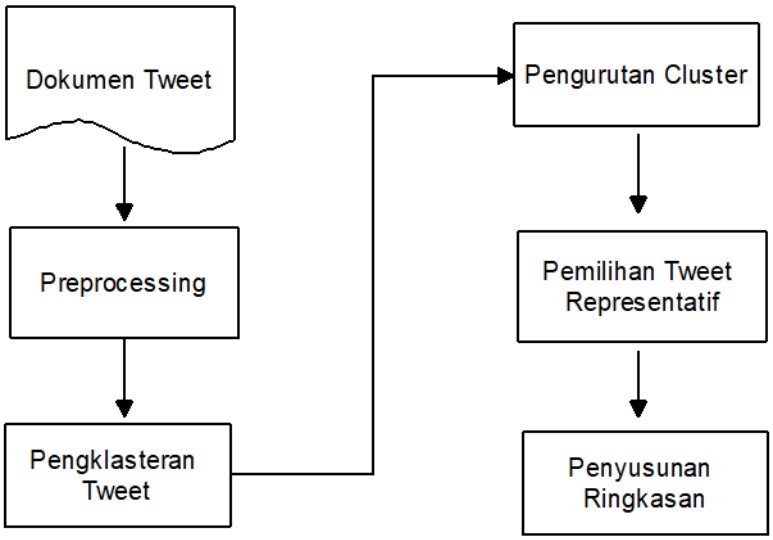
Perancangan sistem untuk pra proses data, yang terdiri dari pemecahan *tweet, case folding, tokenizing, editing,* menghilangkan *stopwords,* normalisasi, dan *steamming*. Hasil dari *text preprocessing* akan menjadi *input* untuk melakukan pembentukan sebuah *cluster* kalimat, yang selanjutnya dilakukan pembobotan kalimat untuk menyusun sebuah ringkasan [6].

4. Pengujian

Proses pengujian dilakukan untuk mengukur kualitas hasil rigkasan berdasarkan kesesuaian antar unit dengan menggunakan pengujian *ROUGE-N*, yaitu mengukur perbandingan *N-gram* dari ringkasan manual dan ringkasan daari sistem.

2.2 Analisa Masalah

Pada Gambar 1 menunjukkan permasalahan yang telah dijelaskan pada penulisan skripsi mengenai *tweet summarization* pada bagian latar belakang, peringkasan multi *tweet* dilakukan untuk menghindari kalimat redundan untuk pemilihan kalimat utama (penting).



Gambar 1. Alur Sistem

2.3 Penyelesaian Masalah

2.3.1 Tahap Preprosesing

Text Preprocesing merupakan tahap pertama yang dilakukan sebelum *input* dokumen diolah lebih lanjut menjadi pengelompokkan kalimat berdasarkan perhitungan algoritma. Pada penelitian ini, preprocessing yang digunakan meliputi pemecahan *tweet, case folding, tokenizing, editing, stopwords, normalisasi, dan stemming* [7, 8, 9].

2.3.1.1 Tahap Case Folding

Case folding merupkan tahapan mengubah semua huruf besar pada *tweet* menjadi huruf kecil, menghilangkan karakter angka, dan menghilangkan delimeter seperti : (.), (,), (:), (;), (!), (?).

2.3.1.2 Tahap Tokenizing

Tahapan *tokenizing* ialah tahapan menguraikan deskripsi dari *tweet* menjadi kata- kata dengan pemisah spasi

2.3.1.3 Tahap Stopword Removal

Tahap *stopword removal* merupakan tahap menghilangkan kata yang tidak penting dalam teks, seperti: di, yang, tidak, untuk, ini, itu, dari, dan, ke, atau, tak, akan.

2.3.1.4 Tahap Steaming

Tahap *Stemming* ialah tahapan yang berfungsi untuk menghilangkan imbuhan seperti awalan dan akhiran.

2.3.2 Analisa Algoritma Cluster Importance

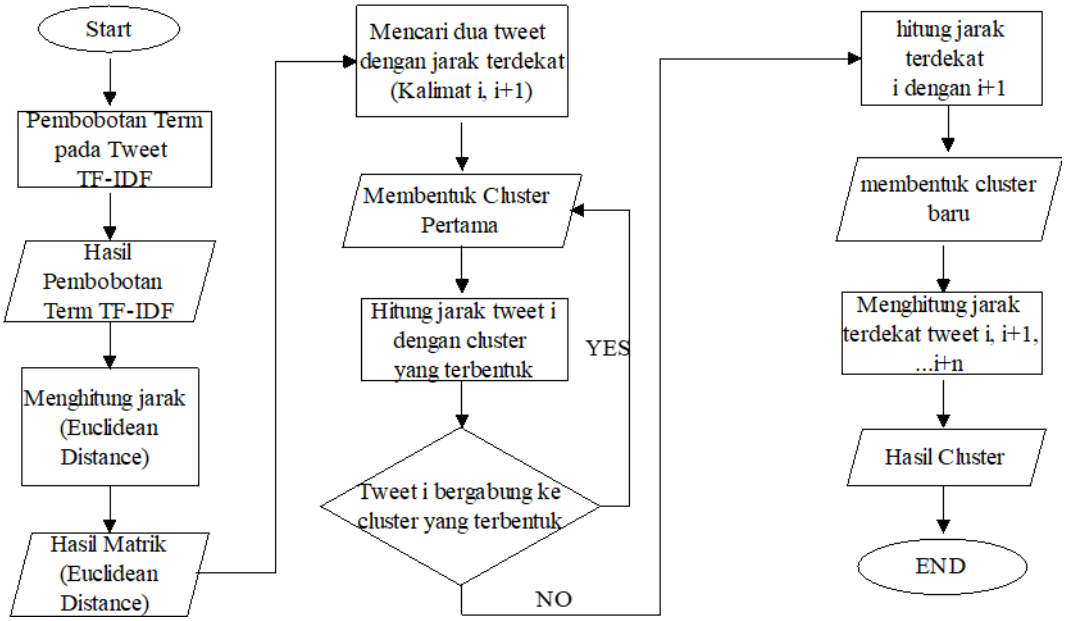
2.3.2.1 Klasterisasi Tweet

Dalam perhitungan bobot menggunakan TF-IDF, dihitung terlebih dahulu nilai TF perkata. Dimana nilai IDF adalah nilai dari setiap kata yang akan di cari, TF adalah jumlah keseluruhan dokumen yang ada, DF jumlah kemuculan kata pada semua dokumen. Setelah mendapat nilai TF dan IDF, maka untuk mendapatkan bobot akhir dari TF-IDF, dimana w(*word*) adalah nilai bobot dari setiap kata, TF (*word*) adalah hasil perhitungan dari TF. IDFi adalah hasil dari perhitungan IDF.

Setelah semua variabel yang dibutuhkan untuk proses pembobotan didapatkan, maka nilai bobot dari Kata Penting tersebut dihitung dengan rumus (TF * IDF) yang kemudian disimpan

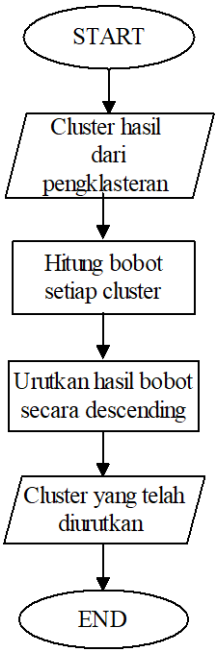
dalam matrik Kata Penting-document. proses ini dilakukan pada semua Kata Penting yang terdapat pada database Kata Penting.

1. Tahap pertama dari proses perhitungan bobot menggunakan tf-idf adalah menghitung IDF.
2. Tahap kedua dari proses perhitung bobot tf-idf adalah tahap perkalian antara tf (*term frequensi*) atau jumlah kemunculan kata tiap dokumen dikali dengan idf.
3. Tahapan ketiga yaitu dengan mencari nilai dua *tweet* terdekat hingga menghasilkan *cluster* tunggal yang didapat dari matrik *Euclidean Distance*.
4. Tahapan keempat menghitung dan memilih level *cluster* yang tepat untuk menentukan bahan peringkasan



Gambar 2. Flowchart Pengklasteran Menggunakan Metode Single Linkage [8[]9]

2.3.2.2. Ordering Clustering



Gambar 3. Flowchart Pengurutan Cluster

Pada Gambar 3 menunjukkan proses dalam cluster ordering terdiri dari perhitungan bobot setiap kalimat dan perhitungan score akhir.

2.3.2.3 Pemilihan Kalimat Representatif

Gambar 4. Flowchart Pemilihan Kalimat Representative

Berdasarkan Gambar 4 diatas, proses dari pemilihan kalimat representative dari setiap cluster dimana untuk mendapatkan nilai satu kalimat representatif dilakukan perhitungan *local importance* dan *global importance* sebagai bobot setiap kalimat pada cluster. Setelah nilai setiap kalimat pada cluster diperoleh, dipilih nilai tertinggi untuk dijadikan ringkasan akhir.

2.3.2.4 Penyusunan Ringkasan

Setelah mendapatkan nilai score akhir maka dilakukan proses pengurutan dari nilai score tertinggi sampai terendah

3. Hasil Penelitian dan pembahasan

Peringkasan kalimat pada tweet pertama pada tahap text pre-procesing yaitu pengelompokan kalimat berdasarkan hitungan algoritma tahap text pre-procesing meliputi pemecahan tweet, case folding, tokenizing, editing, stopwords, normalisasi, dan stemming. Analisa algoritma Cluster importance di lakukan setelah tahap text pre-procesing, di lakukan denagn beberapa tahap yaitu klasterisasi tweet untuk pembobotan TF-IDF, order clustering untuk perhitungan bobot setiap kalimat serta perhitungan skor akhir, pemilihan kata representative untuk tahap peringkasan berdasarkan skor, dan penyusuna ringkasan untuk menyusun menjadi 1 kalimat. Metode Rouge-N di gunakan untuk menghitung persentase persamaan antara ringkasan manual dan ringkasan dari sistem berikut adalah tabel dari perbandingan dari sistem dan manual.

Tabel 1. Nilai Hasil Perbandingan Menggunakan Rouge-N

No	Topik	N	Persentase kemiripan
1	Ahok	18	89%
2	Sumber waras	4	42%
3	Pilkada	4	96%
4	Dukungan parpol	5	77%
5	Korupsi	5	45%
6	Fitnah	5	58%
7	KTP	9	70%
8	Dukungan	11	100%
9	Generasi muda	17	88%
10	Reklamasi	14	71%
11	Sandiaga uno	16	47%
12	Ulama	5	95%
13	HUT DKI	2	-
14	Pemimpin	2	-
15	Kafir	10	85%
16	Nasionalis	5	84%
17	Gubernur baru	6	78%
18	Bahagia	4	52%
19	Pilgub	10	50%
20	Kampanye	10	57%
21	Kinerja djarot	2	-
22	Parpol	6	70%
23	Pajak	8	56%
24	Demokrasi	3	96%
25	Musuh	12	38%

Hasil perhitungan Rouge-N dari Tabel 1 menunjukan bahwa Dari 25 topik data uji ada 3topik yang mempunyai Iterasi 1 yaitu topik HUT DKI, Kinerja Djarot, dan Pemimpin dikarenakan jumlah tweet pada topik tersebut sama - sama memiliki 2 tweet. Pengujian perbandingan antara hasil manual dan hasil sistem menunjukan hasil 100% pada topik Dukungan. Perhitungan Rouge-N menyimpulkan bahwa sistem dapat merangkum tweet minimal 3 tweet, dari perbandingan hasil sistem dan manual di dapatkan tweet minimal mempunyai nilai 96% yang di mana sistem bekerja dengan baik, serta pada topik yang memiliki 18 tweet menunjukan hasil 89%.

4.Kesimpulan dan Saran

4.1 Kesimpulan

Peringkasan kalimat pada tweet pertama pada tahap *text pre-procesing* yaitu pengelompokan kalimat berdasarkan hitungan algoritma tahap *text pre-procesing* meliputi pemecahan *tweet, case folding, tokenizing, editing, stopwords, normalisasi, dan stemming.* Analisa *algoritma Cluster importance* di lakukan setelah tahap text pre-procesing, di lakukan denagn beberapa tahap yaitu klasterisasi tweet untuk pembobotan TF-IDF, order clustering untuk perhitungan bobot setiap kalimat serta perhitungan skor akhir, pemilihan kata representative untuk tahap peringkasan berdasarkan skor, dan penyusuna ringkasan untuk menyusun menjadi 1 kalimat. Metode Rouge-N di gunakan untuk menghitung persentase persamaan antara ringkasan manual dan ringkasan dari sistem, berikut adalah hasil dari perhitungan Rouge-N yaitu: Dari 25 topik data uji ada 3topik yang mempunyai Iterasi 1 yaitu topik HUT DKI, Kinerja Djarot, dan Pemimpin dikarenakan jumlah tweet pada topik tersebut sama - sama memiliki 2 tweet. Pengujian perbandingan antara hasil manual dan hasil sistem menunjukan hasil 100% pada topik Dukungan. Perhitungan Rouge-N menyimpulkan bahwa sistem dapat merangkum tweet minimal 3 tweet, dari perbandingan hasil sistem dan manual di dapatkan tweet minimal mempunyai nilai 96% yang di mana sistem bekerja dengan baik, serta pada topik yang memiliki 18 tweet menunjukan hasil 89%.

4.2 Saran

Beberapa saran yang dapat dilakukan untuk pengembangan system ini, sebagai berikut:

1. Mengkombinasi atau menambahkan algoritma yang dapat memilih level cluster terbaik dan lebih efektif dan akurat dalam memilih cluster.
2. Menambahkan algortima yang dapat menghasilkan ringkasan dengan teknik menyerupai kecerdasan buatan, yaitu menghasilkan ringkasan dengan kalimat-kalimat baru yang mewakili topik pada tweet.

Referensi

[1] Annisa, "Peringkasan Tweet Berdasarkan Trending Topic Twitter dengan Pembobotan TF-IDF dan Single Linkage Anggromerative Hierarchical Clustering", 2016.
[2] OUYANG, Y., Li, W., Zhang, R., Li, S. & Lu, Q. 2012. A Progressive Sentence Selection Strategy for Document Summarization. *Information Processing and Management*.
[3] Sarkar, K, "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, 2009.
[4] Raffi, K. 2013. *New Tweets per second record, and how!. Twitter Official Blog.* [diakses 6 September 2017]
[5] Blanchette, J. 2008. *The little Manual of API Design. Trolltech, a Nokia company.*
[6] Sebastian, F. 2002. *Machine Learning in automated text categorization. ACM Computing Surveys,* Vol. 34, No. 1
[7] Jurafsky and Martin. 2006. Speech and Language Processing, Computational Linguistics, And Speech Recognition 2nd Edition. New Jersey: Pearsin Prentice Hall.
[8] Hovy, E.H. Automated Text Summarization. In R. Mitkov (Ed), *Handbook of computation linguistics*. Oxford University Press. 2001.
[9] Radev, D. R., Hovy, E.H., & McKeown, K., "Introduction to the Special Issue on Summarization" Computational Linguistics, vol. 28, no. 4, hal. 399-408, 2002. Ferreira, R., Freitas, F., Cabral, L. d. Lins, R. D., Lima, R., Franc. A, G., Favaro, L., "A Context Based Text Summarization System", 11th IAPR International Workshop on Document Analysis Systems, IEEE. 2014.