EKSTRAKSI TRENDING ISSUE DENGAN PENDEKATAN DISTRIBUSI KATA PADA PEMBOBOTAN TERM UNTUK PERINGKASAN MULTI-DOKUMEN BERITA

Christian Sri Kusuma Aditya [1)], Chastine Fatichah [2)], dan Diana Purwitasari [3)]
[1, 2, 3)] Teknik Informatika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: christian.s.k.aditya@gmail.com[1)], chastine.fatichah@gmail.com[2)], diana.purwitasari@gmail.com[3)]

ABSTRAK

Penggunaan trending issue dari media sosial Twitter sebagai kalimat penting efektif dalam proses peringkasan dokumen dikarenakan trending issue memiliki kedekatan kata kunci terhadap sebuah kejadian berita yang sedang berlangsung. Pembobotan term dengan TFIDF yang hanya berbasis pada dokumen itu tidak cukup untuk menentukan indeks dari suatu dokumen. Penentuan indeks yang akurat juga bergantung pada nilai informatif suatu term terhadap kelas atau cluster. Term yang sering muncul di banyak kelas atau cluster seharusnya tidak menjadi term yang penting meskipun nilai TFIDF-nya tinggi.

Penelitian ini bertujuan untuk melakukan peringkasan multi dokumen berita menggunakan ekstraksi trending issue dengan pendekatan term distribution on centroid based (TDCB) pada pembobotan fitur dan mengintegrasikannya dengan query expansion sebagai kata kunci dalam peringkasan dokumen. Metode TDCB dilakukan dengan mempertimbangkan adanya kemunculan sub topic dari cluster hasil pengelompokan tweets yang dapat dijadikan nilai informatif tambahan dalam penentuan pembobotan kalimat penting penyusunan ringkasan.

Tahapan yang dilakukan untuk menghasilkan ringkasan multi dokumen berita antara lain ekstraksi trending issue, query expansion, auto labelling, seleksi berita, ekstraksi fitur berita, pembobotan kalimat penting dan penyusunan ringkasan. Hasil percobaan menunjukan metode peringkasan dokumen dengan menambahkan nilai informatif sub topic trending issue NeFTIS-TDCB menunjukan nilai rata-rata max-ROUGE-1 terbesar 0.8615 untuk n=30 dari seluruh varian topik berita.

***Kata Kunci**: cluster importance, peringkasan berita, term distribution on centroid based, trending issue, twitter, query expansion.*

ABSTRACT

Usage of trending issue as important sentence in the document summarization are effective because trending issue has proximity of keywords to an ongoing news events. Weighting term with TFIDF only based on documents that were not enough to determine the index of a document. The accuracy of the index is also dependent on the informative value of a term to a class or cluster. Terms that often appear in many classes or clusters should not be an important term in spite of its high-value TFIDF.

This study aims to perform multi-documents summarization using extraction trending issue with approach of the distribution term on centroid based (TDCB) in the weighting of term and integrate it with the query expansion as a keyword in document summarization. TDCB method is done by considering the emergence of sub-topic of cluster tweets that can be used as an additional informative value in determining the weighting important sentences for summary.

There are steps to produce a multi-news summary, i.e. trending issue extraction, query expansion, auto labelling, news selection, news feature extraction, sentence weighting and the generating summary. The experimental results show the method summarization document by adding the informative value of sub topic trending issue, NeFTIS-TDCB, shows the average value of largest max-ROUGE-1 0.8615 for n = 30 of all variants news topics.

***Keywords**: cluster importance, document summarization, term distribution on centroid based, trending issue, twitter, query expansion.*

I. PENDAHULUAN

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 dokumen (*multi-document summarization*). Ringkasan yang baik merupakan ringkasan dengan cakupan pembahasan (*coverage*) yang luas pada dokumen sumber serta keterhubungan antarkalimat (*coherence*) yang tinggi [1]. Terdapat dua jenis ringkasan berdasarkan dari cara pembuatannya. Jenis pertama adalah *generic summary*, dimana perwakilan dari teks asli yang mencoba untuk mempresentasikan semua fitur penting dari sebuah teks asal, pemakai menginginkan segala informasi yang penting, jenis ini mengikuti pendekatan *bottom-up*. Untuk jenis yang kedua adalah *query-driven summary*, peringkasan bersandar pada spesifikasi kebutuhan informasi pemakai, seperti ekstraksi kata kunci, mengikuti pendekatan *top-down* [2]. Ekstraksi kata kunci (*keyword extraction*) adalah sebuah cara untuk mencari perwakilan kata yang menggambarkan isi sebuah teks.

Salah satu media yang dapat digunakan untuk mengekstraksi kata kunci sebagai informasi adalah *microblog twitter* yaitu dengan mengetahui *trending issue*-nya. Informasi yang dihasilkan dan beredar melalui media *twitter* sangat bebas dan beragam seperti berita, pertanyaan, opini dan komentar. Permasalahan yang mendasari penggunaan *trending issue* sebagai kalimat penting dan efektif dalam proses peringkasan dokumen dikarenakan *trending issue* memiliki kedekatan kata kunci terhadap sebuah kejadian berita yang sedang berlangsung [3]. Pada penelitian lainnya dilakukan teknik pembobotan dengan mengkombinasikan fitur berita dan *trending issue* pada *twitter* [4]. Metode pada penelitian tersebut melakukan pengelompokan *tweets* berdasar frekuensi dan bobot *term* menggunakan *clustering k-medoids*. Selanjutnya dilakukan evaluasi untuk tiap *cluster* dan diambil satu *cluster* dengan bobot tertinggi yang akan dijadikan untuk mengekstraksi *trending issue*. *Cluster* dengan bobot tertinggi yang akan dijadikan sebagai representasi ekstraksi kata kunci *trending issue* [5]. Dari aspek teknis, terdapat kemungkinan jika pengambilan *cluster* lebih dari satu, dengan mempertimbangkan adanya kemunculan *sub topic* lain dari *cluster* hasil pengelompokan *tweets* yang dapat dijadikan nilai informatif tambahan dalam penentuan pembobotan kalimat penting penyusunan ringkasan. Penggunaan lebih dari satu *cluster* inilah dilakukan sebuah teknik pembobotan *term distribution on centroid based* [6]. Metode ini menerapkan konsep pembobotan *intra-cluster*, *inter-cluster* dan keseluruhan dokumen sehingga dapat meningkatkan bobot *term* yang diskriminatif, dengan kata lain metode ini mampu membentuk representasi bobot menjadi lebih baik atau memiliki *sense* terhadap *cluster trending issue* yang terbentuk.

Selain memperhatikan teknik pembobotan, terdapat cara lain untuk mengoptimalkan tingkat *coverage* peringkasan dokumen adalah dengan memperbaiki *query* atau menambahkan kata kunci yang disebut dengan *query expansion*. *Query expansion* merupakan teknik sebagai penghubung adanya *vocabulary gaps* antar kata kunci (*query*) dan dokumen [7]. Meskipun berbeda secara konteks, kemunculan kata kunci yang dihasilkan oleh *trending issue* yang mana kata kunci tersebut sebenarnya relevan terhadap dokumen berita dapat diperoleh dengan memperhitungkan *synonim set* kata kunci nya.

Penelitian ini mengusulkan metode baru yaitu mengembangkan metode ekstraksi *trending issue* dengan pendekatan pembobotan *term distribution centroid-based* (TDCB) dan mengintegrasikan *query expansion* untuk menambah *coverage* sistem dalam melakukan peringkasan multi-dokumen berita. Pembobotan term dengan TFIDF yang hanya berbasis kemunculan frekuensi *term* pada dokumen tidak cukup untuk menentukan indeks dari suatu dokumen. Penentuan indeks yang akurat juga bergantung pada nilai informatif suatu *term* terhadap kelas atau *cluster*. Term yang sering muncul di banyak kelas atau *cluster* seharusnya tidak menjadi *term* yang penting meskipun nilai TFIDF-nya tinggi.

II. TINJAUAN PUSTAKA

Dokumen berita merupakan salah satu media yang digunakan untuk mendapatkan informasi. Meskipun di satu sisi dokumen berita merupakan sumber informasi yang sangat dibutuhkan, namun koleksi berita dalam jumlah besar dapat memberikan dampak negatif bagi pengguna yang mana membutuhkan waktu relatif lama untuk memilah informasi yang sesuai dengan kebutuhan mereka. Ringkasan dibutuhkan untuk mendapatkan isi berdasar inti sari bacaan tanpa mengubah informasi penting wacana dari berita tersebut. Memahami isi dokumen berita melalui ringkasan teks memerlukan waktu yang lebih singkat dibandingkan membaca seluruh isi dokumen, sehingga ringkasan teks menjadi sangat penting. Namun demikian, membuat ringkasan manual dengan dokumen yang banyak akan memerlukan waktu dan biaya yang besar. Sehingga diperlukan suatu sistem peringkasan secara otomatis untuk mengatasi masalah waktu baca dan biaya [8]. Berikut penjelasan tentang landasan teori yang diambil dari pustaka penelitian sebelumnya sebagai bahan acuan dalam melakukan penelitian ini.

A. Term Distribution On Centroid Based

Term weighting merupakan proses penghitungan bobot tiap *term* yang dicari pada setiap dokumen sehingga dapat diketahui ketersediaan dan kemiripan suatu *term* di dalam dokumen. Metode TDCB menggunakan konsep distribusi *term* berdasarkan *intra-class*, *inter-class* dan keseluruhan koleksi dokumen untuk meningkatkan bobot *term* yang diskriminatif [6]. Setiap *term* mempunyai bobot sesuai dengan frekuensi dokumennya (informasi *intra-class*) dan faktor diskriminatif yang berbanding terbalik dengan jumlah kelas atau *cluster* yang berisi *term* tersebut (informasi *inter-class*). Konsep distribusi tersebut berdasarkan prinsip memninimalkan varian atau kemiripan *intra-cluster* dan memaksimalkan varian *inter-cluster*, sehingga data yang mempunyai karakteristik yang sama dikelompokan dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokan ke dalam *cluster* yang lain.

Ditetapkan tiga tipe informasi sebagai representasi pembobotan untuk metode ini, diantaranya Persamaan (1) untuk *inter-class standard deviation* (*icsd*), Persamaan (2) untuk *class standard deviation* (*csd*) dan Persamaan (3) untuk *standard deviation* (*sd*). tf_{ijk} adalah frekuensi dari term t_i dokumen d_j pada kelas C_k.

$$icsd_i = \frac{\Sigma_k \left[\overline{tf}_{ik} - \frac{\Sigma_k \overline{tf}_{ik}}{|c|}\right]^2}{|c|} \quad (1)$$

$$csd_{ik} = \sqrt{\frac{\Sigma_{d_j \epsilon C_k} \left[tf_{ijk} - \overline{tf}_{ik}\right]^2}{|C_k|}} \quad (2)$$

$$sd_i = \sqrt{\frac{\Sigma_k \Sigma_{d_j \epsilon C_k} \left[tf_{ijk} - \frac{\Sigma_k \Sigma_{d_j \epsilon C_k} tf_{ijk}}{\Sigma_k |c_k|}\right]^2}{\Sigma_k |C_k|}} \quad (3)$$

dimana,

$$\overline{tf}_{ik} = \frac{\Sigma_{d_j \epsilon C_k} tf_{ijk}}{|C_k|} \quad (4)$$

Persamaan (4) untuk $\overline{tf}_{ik}$ adalah frekuensi *term* rata-rata pada keseluruhan dokumen yang termasuk pada kategori kelas C_k. $|c|$ adalah jumlah kelas dan $|C_k|$ adalah jumlah dokumen pada kategori kelas C_k.

Dari masing-masing nilai *icsd*, *csd*, dan *sd* digunakan untuk menghitung nilai bobot TDF pada Persamaan (5), dan mengkombinasikannya dengan TF-IDF seperti pada Persamaan (6).

$$\mathrm{TDF}_{ik} = \mathrm{icsd}_i^{\alpha} \; x \; \mathrm{csd}_{ik}^{\beta} \; x \; \mathrm{sd}_i^{\gamma} \quad (5)$$

$$w_{ik} = \mathrm{tf}_{ik} \; x \; \mathrm{idf_i} \; x \; \mathrm{TDF}_{ik} \quad (6)$$

Parameter α, β, dan γ adalah pembobot dari masing-masing informasi (*icsd, csd, sd*), yang mana pembobot dengan angka positif berperan menaikkan nilai informasi, dan sebaliknya pembobot dengan angka negatif menurunkan nilai informasi. Semakin besar nilai parameter yang diberikan, semakin besar kontribusi untuk pembobotan baik untuk menaikkan (*promoter*) atau menurunkan (*demoter*) nilai informasi.

B. Query Expansion

Query Expansion atau perluasan *query* adalah proses mereformulasikan kembali *query* awal dengan melakukan penambahan beberapa term atau kata pada *query* untuk meningkatkan perfoma dalam proses *information retrieval* dan diharapkan mampu menangani masalah ketidakjelasan *query* (*disambiguate query*). Dalam konteks *web search engine*, hal ini termasuk evaluasi *input user* dan memperluas *query* pencarian untuk mendapatkan dokumen yang cocok dengan *query*. Proses perluasan dalam sistem ini dilakukan dengan menggunakan sinonim dari wordnet. Metode yang dilakukan dalam perluasan adalah dengan mencari sinonim dalam bentuk *unstemmed-term* dari *query*.

Salah satu metode *query expansion* diantaranya adalah teknik global yang melibatkan pengetahuan morfologi dan semantik dari kamus elektronik atau sumber informasi leksikal seperti *WordNet*. Metode ini dimana *query* diperluas menggunakan morfologi, derivasi leksikal dan kesamaan semantik semacam sinonim yang sangat bergantung pada sumber informasi [7].

Kateglo adalah aplikasi layanan web dan isi terbuka untuk kamus, tesaurus, dan glosarium bahasa Indonesia. Namanya diambil dari akronim unsur layanannya, ka(mus), te(saurus), dan glo(sarium) [9]. Lisensi isi Kateglo adalah CC-BY-NC-SA, secara ringkas, seluruh isi dapat disalin, disebarkan, dan diadaptasi dengan bebas asal mencantumkan sumber isi, bukan untuk tujuan komersial, dan dalam lisensi yang sama atau serupa dengan lisensi CC-BY-NC-SA. Data kamus kata dari Pusat Bahasa Departemen Pendidikan Nasional Indonesia ditandai dengan "Pusba" atau "Pusat Bahasa", merupakan hak cipta dari Pusat Bahasa dan dipergunakan di Kateglo dengan seizin Pusba. Selain itu glosarium untuk kateglo sendiri menyediakan daftar padanan istilah suatu bidang ilmu tertentu sekaligus tautan ke Wikipedia jika artikel terkaitnya telah ada di sana. Dari glosarium, dibuat juga tautan ke masing-masing lema pembentuknya untuk melakukan pengujian kesesuaian makna [9].

C. Pembobotan Kalimat

Fitur penting ringkasan sebuah dokumen berita merupakan kombinasi dari empat fitur berita diantaranya *word frequency*, TF-IDF, posisi, dan kesesuaian dengan judul berita [10]. Hasil penelitian tersebut menyatakan bahwa keempat teknik tersebut dapat merepresentasikan pembobotan kalimat penting pada keseluruhan isi dokumen ringkasan. Penelitian lain menambahkan pembobotan penggunaan *trending issue* sebagai pembobotan kalimat selain menggunakan empat kombinasi pembobotan dari fitur berita [4]. Salah satu media yang dapat digunakan untuk mengetahui *trending issue* adalah *Twitter*. *Twitter* mengijinkan penggunanya untuk menulis dan membagi pesan berupa teks singkat yang disebut dengan *tweets.*

Adapun total bobot kalimat dalam penelitian tersebut menggunakan gabungan dari 5 teknik pembobotan kalimat. Pembobotan kalimat pertama W_1 adalah *Word Frequency* (WF), nilai kemiripan kalimat S_i terhadap *WFList*

menggunakan *cosine similarity*, dimana $S=\{S_1,..., S_m\}$, Persamaan (1). Pembobotan kalimat kedua W_2 adalah *Term Frequency Inverse Document Frequency* (TFIDF), merupakan hasil penjumlahan dari seluruh bobot *term j* yang muncul pada kalimat *i* (Si), Persamaan (2). Pembobotan kalimat ketiga W_3 adalah posisi kalimat, kalimat yang posisinya berada diawal dokumen memiliki skor lebih besar dibanding kalimat yang posisinya diakhir. Bobot W_3 dihitung dengan menggunakan persamaan (3). Pembobotan kalimat keempat W_4 adalah kemiripan kalimat terhadap judul berita (*Resemblance to the Title*), perhitungan W_4 mengadopsi dari [10] yaitu dengan cara membagi antara jumlah *term* judul yang muncul pada kalimat (NTW) dengan jumlah seluruh *term* yang ada pada judul (*T*), Persamaan (4). Pembobotan kalimat kelima W_5 adalah kemiripan kalimat terhadap *trending issue* (*Resemblance to the trending issue*), Persamaan (5). Metode pengukuran kemiripan kalimat terhadap *trending issue* menggunakan *cosine similarity*. Kalimat yang memiliki skor kemiripan tinggi terhadap *trending issue* akan dianggap sebagai kalimat penting. Pembobotan kalimat W_6 adalah terkait dengan perhitungan redundansi kalimat. Redundansi kalimat diidentifikasi dari kemiripan kalimat ke-*i* (*Si*) terhadap kalimat yang lain (*Sj*) dengan mengadopsi konsep dari MMR. Dimana *j* sebanyak jumlah kalimat yang ada pada dokumen (*D*), Persamaan (6). Metode pembobotan kalimat inilah yang selanjutnya disebut dengan NeFTIS.

$$W_1(s_i) = Sim(s_i, WF_{List}) \quad (7)$$

$$W_2(s_i) = Sim\sum_{j=1}^{n} TFIDFij \quad (8)$$

$$W_3(s_i) = \frac{1}{\sqrt{POS(si)}} \quad (9)$$

$$W_4(s_i) = \frac{NTW}{T} \quad (10)$$

$$W_5(s_i) = Sim(s_i, TI) \quad (11)$$

$$W_6(s_i) = Max_{sj \epsilon D} \frac{2*(Si \cap Sj)}{Si \cup Sj} \quad (12)$$

Langkah selanjutnya adalah menghitung total bobot kalimat *i* (*score*(s_i)) dengan menjumlahkan masing-masing bobot mulai W_1 sampai W_6 pada tiap kalimat keseluruhan dokumen berita, Persamaan (13). Hasil perhitungan inilah yang akan menjadi total bobot kalimat ke-*i*.

$$score(si) = W_1(s_i) + W_2(s_i) + W_3(s_i) + W_4(s_i) + W_5(s_i) - W6(s_i) \quad (13)$$

Penelitian ini akan menyusun ringkasan sebanyak *n* buah kalimat berdasarkan kalimat yang memiliki total bobot kalimat *(score(s_i))* terbesar. Semakin besar nilai total bobot kalimat yang dimiliki maka kalimat tersebut adalah kalimat penting yang dapat dianggap sebagai perwakilan konteks isi dokumen berita.

III. METODOLOGI

Pada penelitian ini secara garis besar terdapat tahapan-tahapan yang dilakukan pada perancangan sistem peringkasan dokumen. Tahapan proses tersebut diantaranya adalah *preprocessing* data, ekstraksi *trending issue*, *query expansion, auto labelling*, seleksi berita, ekstraksi fitur berita dan pembobotan kalimat. Secara global desain model sistem keseluruhan pada penelitian ini dapat dilihat pada Gambar 1.

A. Preprocessing Data

Data uji coba yang digunakan pada penelitian ini berasal dari dua sumber dengan domain yang berbeda yaitu dokumen berita dan *tweets*. Dimana bahasa yang digunakan pada kedua sumber tersebut adalah Bahasa Indonesia. Dokumen berita dan *tweets* diambil berdasarkan topik berita yang sedang berkembang (*hot topic*) pada periode waktu yang sama. Jumlah topik yang digunakan untuk uji coba sebanyak 5 topik, yaitu "gerhana", "pilgub", "reklamasi", "bpk", dan "alibaba". Dimana total data keseluruhan dokumen berita sebanyak 52 berita sedangkan total data *tweet* sebanyak 1479.

Fase *Text preprocessing* adalah fase yang pertama kali dilakukan, baik pada dokumen *twitter* dan dokumen berita diproses sehingga terbentuk *token* atau kata yang akan dijadikan *index* pembentukan dokumen. Beberapa tahapan dari *text preprocessing* adalah *case folding*, *filtering*, *tokenizing*, dan *stopword removal* yang dilakukan secara berurutan.

Khusus untuk penanganan dokumen *twitter*, normalisasi dilakukan untuk membersihkan tags yang biasa muncul pada tweets, seperti : *hashtag*, *mentioned*, dan *link*. *Tweets* memiliki karakteristik dimana ketiga fitur tersebut selalu diawali dengan karakter khusus masing-masing, simbol # untuk menandai *hashtag*, simbol @ untuk menandai *mentioned*, dan awalan *http* untuk menandai sebuah *link*. Untuk *hashtag* akan disimpan setelah diubah menjadi kata dengan menghapus simbol # yang menjadi awalan kemudian teks yang ada dibelakang simbol akan disimpan sebagai kata. Contohnya : #kebakaran dan #cuacaSurabaya, masing-masing akan diekstraksi menjadi kata kebakaran, cuaca dan surabaya. Sedangkan untuk *mentioned* dan *link* akan langsung dihapus dikarenakan keduanya

tidak digunakan dalam penelitian ini.

B. Ekstraksi Trending Issue

Ada 3 langkah yang dilakukan untuk ekstraksi *Trending Issue* yaitu pengelompokan *tweets*, ekstraksi *issue*, dan pembobotan *issue*. Pengelompokan *tweets* bertujuan untuk mengelompokan *tweets* berdasarkan kesamaan isi dari *tweets*. *Issue* didapatkan dengan cara mengekstraksi kata kunci dari tiap hasil pengelompokan *tweets* [10]. Ekstraksi kata kunci adalah pencarian kata penting dari data *tweets* yang diidentifikasi dan diseleksi dari frekuensi kemunculan kata pada *tweets* dengan menggunakan pembobotan kata berdasarkan *Term Frequency* (TF), TFIDF dan *Word Frequency* (WF) [3]. Kata atau *term* yang hasil ekstraksi kata kunci adalah kata memiliki bobot diatas nilai ambang atau *threshold*. Sehingga hasil akhir dari proses ini adalah setiap *cluster tweets* akan memiliki kumpulan kata kunci yang merepresentasikan *issue*.

Langkah berikutnya pada proses ekstraksi *Trending Issue* adalah menyeleksi *Issue* dari setiap group tweets dengan cara memberikan bobot untuk masing-masing *issue* menggunakan konsep *Cluster Importance* [11]. *Cluster Importance* (CI) adalah sebuah metode yang sebelumnya digunakan untuk mengurutkan *cluster* berdasarkan kata penting yang muncul pada *cluster*, dimana dalam hal ini *issue* merepresentasikan sebuah *cluster*.

C. Query Expansion Trending Issue

Setelah tahap pembentukan *Trending Issue*, dilakukan *query expansion dengan* melakukan pencarian *synonim set* dari tiap *term* pada *Trending Issue* untuk menambah tingkat *coverage* peringkasan dokumen. *Query expansion* model global membutuhkan *knowledge-based*, pada penelitian ini digunakan *lexical database* Bahasa Indonesia bernama Kateglo. Proses tahap *query expansion* dapat dilihat pada Gambar 2.

D. Auto Labelling

Pada penelitian ini pembobotan *term* pada dokumen berita menggunakan metode *term distribution centroid based* (TDCB) dimana metode ini menggunakan sebaran distribusi kata terhadap *cluster* yang terbentuk oleh dokumen *twitter*, maka perlu adanya proses *labelling* kalimat dari dokumen berita terhadap *issue* dari masing-masing *cluster* pada *twitter*. Tujuan dari proses *labelling* untuk mengklasifikasikan kalimat pada dokumen berita ke salah satu *cluster issue* dengan menghitung kedekatan jaraknya menggunakan *unigram matching similarity* [11]. Metode tersebut dipilih mengingat hasil ekstraksi kata kunci dan kalimat dokumen berita adalah unit yang pendek maka nilai *similarity* akan bernilai kecil ketika dihitung dengan pengukuran *cosine similarity* [12].

E. Seleksi Berita

Proses seleksi bertujuan untuk mendapatkan sejumlah *n* berita yang relevan terhadap *Trending Issue*, $D = \{D_1, \ldots, D_n\}$ dimana proses tahapan ini adalah pencarian dokumen dengan menggunakan *Trending Issue* sebagai *query* untuk mencari kumpulan dokumen berita yang relevan.

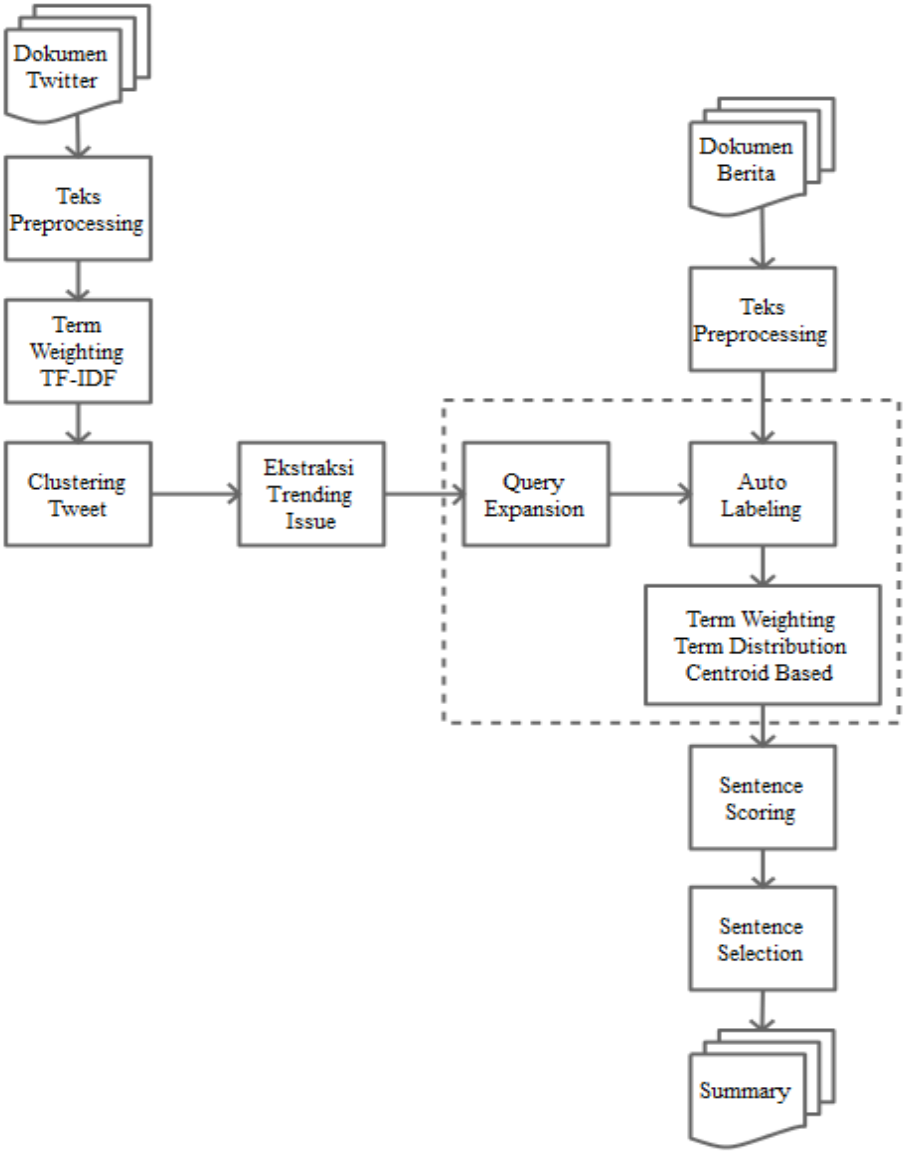
F. Ekstraksi Fitur Berita dan Pembobotan Kalimat

Ekstraksi fitur berita dilakukan terhadap sejumlah *n* berita yang didapatkan dari proses seleksi. Dari gabungan *n* berita tersebut akan didapatkan kumpulan *m* kalimat $T = \{t_1, \ldots, t_m\}$. Pada penelitian ini digunakan fitur penting dari berita yaitu *term frequency* (TF), *document frequency* (DF), *word frequency* (WF), *term distribution centroid based* (TDCB), posisi kalimat, dan judul berita. Proses perhitungan bobot TDCB dengan mendapatkan nilai dari *icsd*, *csd* dan *sd* yang kemudian digunakan untuk mencari nilai TDF. Langkah selanjutnya nilai TDF dikombinasikan dengan nilai TFIDF untuk mendapatkan bobot W_{ik} sebagai bentuk representasi dari nilai fitur TDCB, seperti yang telah dijelaskan sebelumnya oleh Persamaan (1) sampai dengan Persamaan (6).

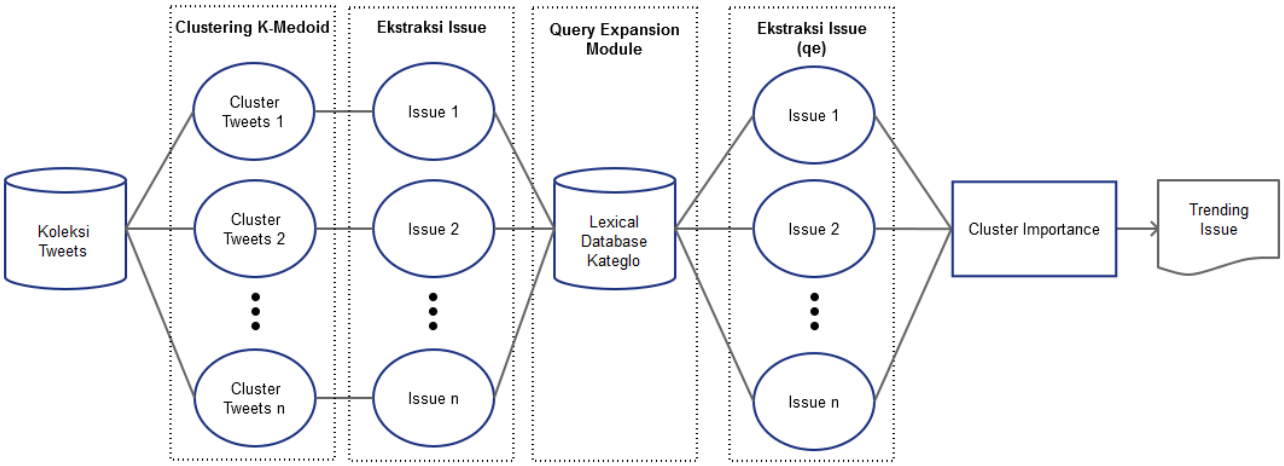
Setelah dilakukan ekstraksi fitur pada dokumen berita, hitung keseluruhan bobot kalimat untuk mendapatkan total bobot masing-masing kalimat dimana total bobot yang paling tinggi merupakan bentuk perwakilan kalimat penting dari keseluruhan isi dokumen berita, kalimat-kalimat penting inilah yang dinamakan ringkasan berita. Teknik pembobotan kalimat menggunakan metode NeFTIS [4], namun pada penelitian ini dilakukan modifikasi terhadap bobot W_2 dan W_5, dimana sebelumnya bobot kalimat W_2 pada NeFTIS yang merupakan hasil penjumlahan bobot TFIDF pada *term j* sebuah kalimat diganti dengan bobot TDCB *term j*, Persamaan (14). Sedangkan bobot kalimat W_5 yaitu kemiripan kalimat dokumen berita terhadap *trending issue* akan dilakukan proses *query expansion* agar memeperoleh *coverage* kata kunci yang lebih pada dokumen berita, Persamaan (15). Perubahan 2 informasi pembobotan inilah yang selanjutnya dinamakan metode NeFTIS-TDCB dan akan dilakukan uji komparasi hasil ringkasan dokumen berita dengan metode sebelumnya, NeFTIS.

$$W_2(s_i) = Sim \sum_{j=1}^{n} Wik_{ij} \quad (14)$$

$$W_5(s_i) = Sim(s_i, TIqe) \quad (15)$$



Gambar 1. Framework Peringkasan Dokumen NeFTIS-TDCB



Gambar 2. Alur proses ekstraksi Trending Issue dan Query Expansion

IV. HASIL DAN PEMBAHASAN

Serangkaian uji coba dilakukan untuk mengetahui apakah penelitian yang dilakukan telah dapat memenuhi tujuan penelitian sebagaimana yang telah direncanakan diantaranya, menguji kualitas *cluster Trending Issue* menggunakan metode *silhoutte coeffisien* dan membandingkan hasil peringkasan NeFTIS-TDCB terhadap metode NeFTIS dengan mendaptakan masing-masing nilai ROUGE-N.

Nilai *silhoutte coeffecient* merupakan nilai kualitas *cluster* yang menunjukkan derajat kedekatan antar objek di dalam sebuah *cluster*. Dimana pada penelitian ini objek *i* direpresentasikan dengan *tweets*. Metode ini merupakan metode validasi *cluster* yang menggabungkan metode *cohesion* dan *separation* dengan Persamaan (16), dimana *a(i)* adalah perbedaan rata-rata objek *i* ke semua objek lain pada *cluster A. b(i)* adalah rata-rata jarak dari objek *i* dengan objek yang berada di *cluster* lainnya, dari semua jarak rata-rata tersebut ambil nilai yang paling kecil.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (16)$$

Tabel I adalah hasil uji coba pengelompokan *tweets* menggunakan algoritma *clustering k-medoids* dan beberapa varian penggunaan perhitungan jarak (*euclidean*, *cosine*, *manhattan*).

Skenario untuk uji coba pertama adalah penentuan jumlah *centroid* (*k*) untuk mendapatkan satu hasil *clustering* terbaik. Dalam penelitian ini, untuk mendapatkan nilai *k* yang optimal dilakukan uji coba dari beberapa nilai *k* kemudian dipilih bentuk *clustering* dengan jumlah *k* terbaik untuk dilanjutkan tahap berikutnya, peringkasan dokumen berita. Uji coba nilai *k* dimulai dari *k*=2 sampai *k*=5. *Clustering* dengan *k* terbaik adalah yang memiliki nilai rata-rata *Silhouette* atau ASW paling besar. Pada penelitian ini, juga dipertimbangkan pemilihan bentuk *cluster* yang mempunyai sebaran data *tweet* yang cukup merata di setiap *cluster* agar masing-masing *cluster* memiliki *term* hasil ekstraksi *issue* sebagai bentuk perwakilan tiap *cluster*. Kondisi *unbalance* atau persebaran anggota *tweet* yang tidak merata tidak cocok apabila dilanjutkan ke proses pembobotan *term* TDCB, karena sebagian besar *term* hasil ekstraksi *issue* hanya akan dimiliki oleh satu *cluster* saja, tidak dimiliki oleh *cluster* lainnya.

Nilai ASW yang didapatkan untuk dokumen *tweet* masuk ke kategori "cukup baik" dimana *range* nilai nya adalah $0.26 \leq \text{ASW} < 0.51$. Hal ini disebabkan oleh karakteristik dari dokumen teks, khususnya twitter Bahasa Indonesia, dimana karakteristik data ini memiliki banyak bentuk kata unik yang disebabkan banyak kata yang tidak sesuai EYD, *typo*, *slang* dan sebagainya sehingga dimensi fitur data ini terbentuk cukup banyak.

Dari hasil percobaan penggunaan variasi perhitungan jarak *euclidean*, *cosine* dan *manhattan*, ketika menggunakan *euclidean* dan *manhattan* nilai ASW yang dihasilkan bisa dikatakan tidak terlalu berbeda secara signifikan dengan menggunakan algoritma *clustering k-medoids*. Sedangkan ketika menggunakan perhitungan jarak *cosine* nilai ASW sedikit lebih rendah namun nilai ASW cenderung lebih stabil di setiap variasi nilai *k*. Selain itu penggunaan *cosine* juga mempengaruhi sebaran data *tweet* lebih merata dibanding *euclidean* dan *manhattan* di setiap *cluster* yang terbentuk. Hal ini cukup berpengaruh dalam proses ekstraksi kata kunci selanjutnya dan menghindari kasus *unbalance clustering* walaupun nilai ASW lebih tinggi ketika menggunakan *euclidean* dan *manhattan*.

Setelah terbentuk *cluster*, langkah selanjutnya adalah ekstraksi *issue* atau kata kunci dari tiap *cluster* yang ada. *Issue* merupakan kata yang ada pada *cluster* yang diseleksi berdasarkan kombinasi nilai TF, TFIDF dan WF dengan nilai *threshold* TF > 1.0 , TFIDF > 0.2 dan WF > 2.

Selanjutnya dari hasil ekstraksi *issue* untuk tiap *cluster* dilakukan pembobotan dengan menggunakan metode *Cluster Importance* (CI) [11]. Tabel II memperlihatkan contoh hasil dari pembobotan *cluster* untuk topik berita "gerhana".

Setelah dilakukan pembobotan tiap *cluster*, langkah selanjutnya adalah mengurutkan *cluster* berdasarkan bobot yang CI terbesar. *Trending Issue* akan diekstraksi dengan cara memilih *issue* yang memiliki bobot terbesar dari seluruh *issue* yang ada. Dengan asumsi *issue* dengan bobot besar adalah *issue* yang penting atau yang paling banyak dibahas di media Twitter. Jika pemilihan *issue* lebih dari satu *cluster*, maka pembentukan *Trending Issue* merupakan gabungan dari *issue* yang digunakan. Dari masing-masing *issue* atau kata kunci yang terbentuk tiap *cluster* dilakukan tahap *query expansion* dapat dilihat hasilnya pada Tabel III.

Sebelum dilakukan proses ekstraksi fitur berita dimana akan dilakukannya pembobotan TDCB, proses *auto labelling* dilakukan untuk memetakan kemiripan tiap kalimat dokumen berita terhadap masing-masing *cluster issue* pada twitter.

Skenario uji coba berikutnya untuk mengukur performa dari hasil ringkasan menggunakan metode pembobotan NeFTIS-TDCB dan NeFTIS. Dimana untuk mengukur performa kedua metode pembobotan tersebut digunakan metode evaluasi ROUGE-N [13], Persamaan (17).

$$ROUGE-N = \frac{\Sigma_{S \epsilon Summ_{ref}} \Sigma_{N-gram \epsilon S} Count_{match}(N-gram)}{\Sigma_{S \epsilon Summ_{ref}} \Sigma_{N-gram \epsilon S} Count(N-gram)} \quad (17)$$

Penelitian ini menggunakan 2 sumber *groundtruth* yang akan dibandingkan dengan ringkasan yang dihasilkan oleh sistem, sehingga untuk mendapatkan satu nilai hasil evaluasi diambil nilai ROUGE-N yang paling maksimum dari kedua nilai ROUGE-N yang ada (max-ROUGE-N), Persamaan (18).

$$ROUGE - N_{multi} = argmax_i \, ROUGE - N(sc, rs_i) \quad (18)$$

Penentuan nilai *n* jumlah kalimat untuk menyusun hasil ringkasan, misalkan diambil sebanyak 10 kalimat sebagai penyusun ringkasan sehingga n=10. Maka urutan 10 kalimat dengan bobot tertinggi yang terpilih menjadi penyusun ringkasan. Tabel IV menunjukkan perbandingan nilai rata-rata max-ROUGE-1 untuk seluruh topik dari tiap variasi *n* jumlah kalimat. Selain melakukan pengujian terhadap metode NeFTIS-TDCB dengan NeFTIS, dilakukan uji coba pengaruh penggunaan *query expansion* terhadap *trending issue*, dengan membandingkan hasil dari NeFTIS-TDCB-qe dimana metode ini tanpa adanya tahap integrasi dengan *query expansion*.

Dari Tabel IV dapat disimpulkan hasil nilai ROUGE-1 metode NeFTIS-TDCB mengungguli nilai ROUGE-1 dibanding metode NeFTIS di keempat topik yaitu "gerhana", "pilgub", "reklamasi" dan "alibaba". Sedangkan pengaruh penggunaan *query expansion* tidak terlalu meningkatkan nilai ROUGE sistem secara signifikan. Khusus untuk topik "bpk" dimana terdapat penurunan nilai ROUGE-1 NeFTIS-TDCB dengan NeFTIS-TDCB-qe, hal ini diakibatkan perluasan *term* atau *synonim set* dari *Trending issue* yang dilakukan oleh *query expansion* banyak yang ternyata tidak terdapat di dokumen berita sehingga akan menurunkan nilai ROUGE.

Tabel V menunjukkan nilai rata-rata max-ROUGE-1 seluruh topik dari tiap variasi *n* kalimat mulai dari 5, 10, 20 dan 30. Dari tabel V tersebut dapat disimpulkan bahwa hasil dari metode pembobotan NeFTIS-TDCB memberikan hasil rata-rata max-ROUGE-1 yang lebih tinggi dibanding dengan metode NeFTIS dan NeFTIS-TDCB-qe untuk hampir seluruh variasi jumlah *n* kalimat yang menyusun ringkasan, dengan nilai rata-rata max-ROUGE-1 terbesar untuk NeFTIS-TDCB ketika n= 30 yaitu sebesar 0.8615.

Dari hasil percobaan perbandingan metode NeFTIS-TDCB dengan NeFTIS, dimana pada metode NeFTIS hanya digunakan 1 buah *cluster issue* berdasar bobot CI tertinggi. Sedangkan metode NeFTIS-TDCB menggunakan lebih dari 1 *cluster issue* (pemilihan *cluster issue* tetap berdasar urutan *descending* bobot CI tertinggi) dikarenakan *issue* yang terbentuk di tiap *cluster* lainnya dapat memiliki nilai informasi berbeda yang dapat digunakan sebagai pemilihan kalimat penting pada dokumen berita yang akan diringkas. Terbukti hal ini dapat meningkatkan nilai rata-rata max-ROUGE-1 yang ditunjukan pada Gambar 3.

Selain itu semakin banyak jumlah kalimat *n* yang terambil oleh sistem dalam menyusun ringkasan, semakin besar kemungkinan *term* atau kata yang muncul bersama pada *groundtruth*. Hal ini juga dibuktikan pada Gambar 3, semakin besar nilai *n* yang diberikan berbanding lurus dengan meningkatknya nilai rata-rata max-ROUGE dimana nilai rata-rata max-ROUGE-1 terbesar untuk NeFTIS-TDCB adalah ketika n=30. Perlu diperhatikan bahwa ringkasan yang terlalu panjang juga akan mempersulit pembaca dan dimungkinkan terdapat pengulangan yang tidak perlu, dimana balik ke tujuan dasar sebuah ringkasan adalah mendapatkan intisari atau ide pokok dalam bentuk singkat namun tidak menghilangkan unsur estetika dan tetap bisa mewakili artikel aslinya.

TABEL I
PENCARIAN NILAI K OPTIMAL UNTUK TIAP TOPIK TWEETS TERHADAP NILAI ASW

Topik	*k*	Centroid	ASW		
			Euclidean	Cosine	Manhattan
gerhana	2	**K1**: 22, **K2**: 203	0.2842	0.286	0.2842
	3	**K1**: 53, **K2**: 284, **K3**: 56	0.2925	0.2898	0.292
	4	**K1**: 18, **K2**: 22, **K3**: 240, **K4**: 267	0.2871	0.2534	0.2871
	5	**K1**: 15, **K2**: 22, **K3**: 152, **K4**: 167, **K5**: 314	0.2854	0.2699	0.2837
pilgub	2	**K1**: 132, **K2**: 220	0.3171	0.2926	0.2908
	3	**K1**: 106, **K2**: 132, **K3**: 79	0.2658	0.3233	0.2541
	4	**K1**: 52, **K2**: 195, **K3**: 86, **K4**: 236	0.3825	0.3337	0.3825
	5	**K1**: 253, **K2**: 76, **K3**: 65, **K4**: 79, **K5**: 159	0.3914	0.3298	0.299
reklamasi	2	**K1**: 36, **K2**: 107	0.3775	0.3112	0.3775
	3	**K1**: 318, **K2**: 36, **K3**: 0	0.2455	0.279	0.2455
	4	**K1**: 214, **K2**: 225, **K3**: 39, **K4**: 215	0.199	0.2322	0.2256
	5	**K1**: 260, **K2**: 296, **K3**: 168, **K4**: 244, **K5**: 312	0.2392	0.2688	0.2392
bpk	2	**K1**: 9, **K2**: 189	0.3923	0.2941	0.3923
	3	**K1**: 166, **K2**: 173, **K3**: 283	0.3196	0.3141	0.3196
	4	**K1**: 191, **K2**: 77, **K3**: 225, **K4**: 168	0.2527	0.3014	0.1663
	5	**K1**: 110, **K2**: 121, **K3**: 68, **K4**: 251, **K5**: 156	0.2017	0.2735	0.2103
alibaba	2	**K1**: 177, **K2**: 91	0.3078	0.3035	0.3078
	3	**K1**: 102, **K2**: 163, **K3**: 199	0.3068	0.3036	0.3068
	4	**K1**: 113, **K2**: 152, **K3**: 25, **K4**: 162	0.301	0.3079	0.301
	5	**K1**: 154, **K2**: 94, **K3**: 140, **K4**: 142, **K5**: 192	0.2986	0.3063	0.2986

TABEL II
HASIL EKSTRAKSI ISSUE TIAP CLUSTER DAN BOBOT CI (TOPIK "GERHANA")

Cluster	Issue	Jumlah Tweets	Bobot CI	Urutan
1	gmt detik gerhana matahari indonesia	87	13.956	2
2	detik daerah anak unik kacamata shalat fenomena pagi foto tidur	239	23.519	1
3	live detik	42	10.979	3

TABEL III
PERBANDINGAN HASIL EKSTRAKSI ISSUE

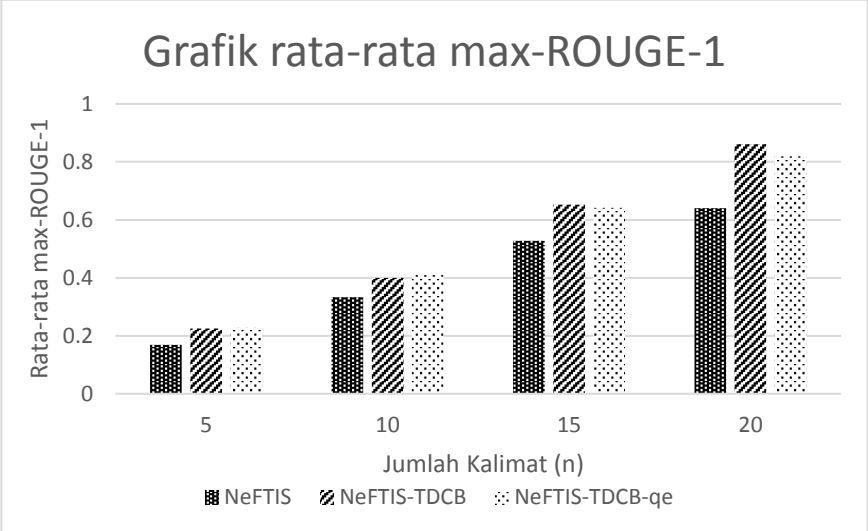
Tanpa Query Expansion	Dengan Query Expansion
Cluster 1 : ['gmt', 'detik', 'gerhana', 'matahari', 'indonesia'] **Cluster 2 :** ['detik', 'daerah', 'anak', 'unik', 'kacamata', 'shalat', 'fenomena', 'pagi', 'foto', 'tidur'] **Cluster 3 :** ['live', 'detik']	**Cluster 2 :** TI-QE CLUSTER INDEX Ke-1 = ['detik', 'daerah', 'anak', 'unik', 'kacamata', 'shalat', 'fenomena', 'pagi', 'foto', 'tidur', 'sekon', 'alam', 'area', 'kawasan', 'wilayah', 'dar', 'desa', 'domain', 'luak', 'luruh', 'medan', 'region', 'segmen', 'arek', 'bani', 'beru', 'budak', 'gulam', 'momongan', 'nak', 'pelarai demam', 'pelerai demam', 'sambungan jiwa', 'sasian', 'khusus', 'tesmak', 'fakta', 'gejala', 'kenyataan', 'awal', 'cepat', 'bayangan', 'gambaran', 'pantulan', 'potret', 'memicing', 'molor', 'berbaring'] **Cluster 1 :** TI-QE CLUSTER INDEX Ke-0 = ['gmt', 'detik', 'gerhana', 'matahari', 'indonesia', 'sekon', 'eklips', 'baskara', 'mentari', 'metari', 'rawi', 'surya', 'syamsi', 'syamsu'] **Cluster 3 :** ['live', 'detik', sekon]

TABEL IV
HASIL ROUGE-1 SELURUH TOPIK UNTUK JUMLAH KALIMAT 10 (N=10)

No.	Topik	Groundtruth 1			Groundtruth 2		
		NeFTIS	NeFTIS-TDCB	NeFTIS-TDCB-qe	NeFTIS	NeFTIS-TDCB	NeFTIS- TDCB-qe
1	gerhana	0.3331	0.3724	0.3643	0.3075	0.3445	0.3361
2	pilgub	0.3055	0.3858	0.3858	0.3087	0.3888	0.3888
3	reklamasi	0.3235	0.4426	0.4458	0.3309	0.4643	0.4605
4	bpk	0.4401	0.3660	0.4062	0.4330	0.3612	0.4008
5	alibaba	0.2516	0.4285	0.4285	0.2516	0.4285	0.4285

TABEL V
NILAI MAX-ROUGE-1 SELURUH TOPIK UNTUK TIAP VARIASI JUMLAH KALIMAT (N)

N	Rata-Rata Max-Rouge-1		
	NeFTIS	NeFTIS-TDCB	NeFTIS-TDCB-qe
5	0.1683	0.2256	0.2274
10	0.3329	0.3997	0.4097
20	0.5274	0.6523	0.6417
30	0.6405	0.8615	0.8191



Gambar 3. Rata-rata smax-ROUGE-1 dari Seluruh Topik untuk Tiap Variasi Jumlah Kalimat (n)

V. KESIMPULAN

Berdasarkan uji coba dan analisa hasil, maka dapat ditarik beberapa kesimpulan antara lain:

1. Hasil uji coba pengelompokkan *tweets* menghasilkan kualitas untuk keseluruhan topik masuk kriteria "cukup baik" dilihat dari nilai ASW yang berada pada $0.26 \leq ASW < 0.51$ dengan rata-rata nilai ASW 0.2937.
2. Berdasarkan nilai rata-rata max-ROUGE-1 metode pembobotan NeFTIS-TDCB mampu memberikan hasil yang lebih akurat dibandingkan dengan metode pembobotan NeFTIS dengan nilai rata-rata max-ROUGE-1 terbesar 0.9010 untuk n=30.
3. Pengintegrasian dengan *query expansion* tidak terlalu berdampak signifikan terhadap hasil nilai ROUGE metode NeFTIS-TDCB, perlu diperhatikan bahwa semakin banyak *term synonim set* yang terbentuk oleh proses *query expansion* dapat juga menurunkan nilai kemiripan kalimat karena dimensi dari fitur yang bertambah dan banyaknya ketidaksamaan *term* dari *synonim set* terhadap dokumen berita.

DAFTAR PUSTAKA

[1] K. Umam, F. W. Putro, G. Q. O. Pratamasunu, A. Z. Arifin and D. Purwitasari, "Optimasi Coverage, Diversity, dan Coherence Pada Peringkasan Multi-Dokumen,"*Jurnal Ilmu Komputer dan Informasi*, pp. 1-16, 2015.

[2] F. El-Ghannam and T. El-Shishtawy, "Multi-Topic Multi-Document Summarizer," *arXiv preprint arXiv:1401.0640*, 2014.

[3] D. Kim, S. Kim, M. Jo and E. Hwang, "SNS-based issue detection and related news summarization scheme," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, 2014.

[4] N. Hayatin, C. Fatichah and D. Purwitasari, "Pembobotan Kalimat Berdasarkan Fitur Berita Dan Trending Issue Untuk Peringkasan Multi Dokumen Berita," *JUTI: Jurnal Ilmiah Teknologi Informasi, 13(1)*, pp. 38-44, 2015.

[5] D. Purwitasari, C. Fatichah, I. Arieshanti and N. Hayatin, "K-medoids algorithm on Indonesian Twitter feeds for clustering trending issue as important terms in news summarization," in *Information & Communication Technology and Systems (ICTS)*, Surabaya, 2015.

[6] V. Lertnattee and T. Theeramunkong, "Effect of term distributions on centroid-based text categorization," *Information Sciences, 158*, pp. 89-115, 2004.

[7] M. A. Pasca and S. M. Harabagiu, "High performance question/answering," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001.

[8] H. Y. Aristoteles, A. Ridha and A. Julio, "ext Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algoritm. International Journal of Science Issues," *IJCSI*, pp. 1694-0814, 2012.

[9] I. Lanin, "Kateglo," 2009. [Online]. Available: https://ivanlanin.wordpress.com/2009/06/11/kateglo/. [Accessed 2015].

[10] R. Ferreira, F. Freitas, L. De Souza Cabral, R. Dueire Lins, R. Lima, G. Franca, S. Simske and L. Favaro, "A Context Based Text Summarization System," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop*, Tours, 2014.

[11] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," *nternational Journal of Computing Science and Communication Technologies, 2(1)*, pp. 325-335, 2009.

[12] I. Suputra, A. Z. Arifin and A. Yuniarti, "Pendekatan Positional Text Graph Untuk Pemilihan Kalimat Representatif Cluster Pada Peringkasan Multi-Dokumen," *Jurnal Ilmu Komputer*, p. 62, 2013.

[13] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop (Vol. 8)*, 2004.