Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202184324
Vol. 8, No. 1, Februari 2021, hlm. 199-208 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

ELIMINASI DATA *NON-TOPIC* MENGGUNAKAN PEMODELAN TOPIK UNTUK PERINGKASAN OTOMATIS DATA *TWEET* DENGAN KONTEKS COVID-19

Putri Damayanti[1], Diana Purwitasari[2], Nanik Suciati[3]

[1,2,3]Teknik Informatika, Fak. Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember
Email Penulis Korespondensi: [1]putri.19051@mhs.its.ac.id, [2]diana@if.its.ac.id, [3]nanik@if.its.ac.id,

(Naskah masuk: 08 April 2020, diterima untuk diterbitkan: 02 Februari 2021)

Abstrak

Akun *twitter*, seperti Suara Surabaya, dapat membantu menyebarkan informasi tentang COVID-19 meskipun ada bahasan lainnya seperti kecelakaan, kemacetan atau topik lain. Peringkasan teks dapat diimplementasikan pada kasus pembacaan data *twitter* karena banyaknya jumlah *tweet* yang tersedia, sehingga akan mempermudah dalam memperoleh informasi penting terkini terkait COVID-19. Jumlah variasi bahasan pada teks *tweet* mengakibatkan hasil ringkasan yang kurang baik. Oleh karena itu dibutuhkan adanya eliminasi *tweet* yang tidak berkaitan dengan konteks sebelum dilakukan peringkasan. Kontribusi penelitian ini adalah adanya metode pemodelan topik sebagai bagian tahapan dalam serangkaian proses eliminasi data. Metode pemodelan topik sebagai salah satu teknik eliminasi data dapat digunakan dalam berbagai kasus namun pada penelitian ini difokuskan pada COVID-19. Tujuannya adalah untuk mempermudah masyarakat memperoleh informasi terkini secara ringkas. Tahapan yang dilakukan adalah pra-pemrosesan, eliminasi data menggunakan pemodelan topik dan peringkasan otomatis. Penelitian ini menggunakan kombinasi beberapa metode word embedding, pemodelan topik dan peringkasan otomatis sebagai pembanding. Ringkasan diuji menggunakan metode ROUGE dari setiap kombinasi untuk ditemukan kombinasi terbaik dari penelitian ini. Hasil pengujian menunjukkan kombinasi metode Word2Vec, LSI dan TextRank memiliki nilai ROUGE terbaik yaitu 0.67. Sedangkan kombinasi metode TFIDF, LDA dan Okapi BM25 memiliki nilai ROUGE terendah yaitu 0.35.

Kata kunci: *COVID-19, Pemodelan Topik, Peringkasan Otomatis*

NON-TOPIC DATA ELIMINATION USING TOPIC MODELLING FOR AUTOMATIC SUMMARIZATION ON TWEETS WITH COVID-19 CONTEXT

Abstract

Twitter accounts, such as Suara Surabaya, can help spread information about COVID-19 even though there are other topics such as accidents, traffic jams or other topics. Text summarization can be implemented in the case of reading Twitter data because of the large number of tweets available, making it easier to obtain the latest important information related to COVID-19. The number of discussion variations in the tweet text results in poor summary results. Therefore, it is necessary to eliminate tweets that are not related to the context before summarization is carried out. The contribution to this research is the topic modeling method as part of a series of data elimination processes. The topic modeling method as a data elimination technique can be used in various cases, but this research focuses on COVID-19. The aim is to make it easier for the public to obtain current information in a concise manner. The steps taken in this study were pre-processing, data elimination using topic modeling and automatic summarization. This study uses a combination of several word embedding methods, topic modeling and automatic summarization as a comparison. The summary is tested using the ROUGE method of each combination to find the best combination of this study. The test results show that the combination of Word2Vec, LSI and TextRank methods has the best ROUGE value, 0.67. While the combination of TFIDF, LDA and Okapi BM25 methods has the lowest ROUGE value, 0.35.

Keywords: *COVID-19, Topic Modelling, Automatic Summarization*

1 PENDAHULUAN

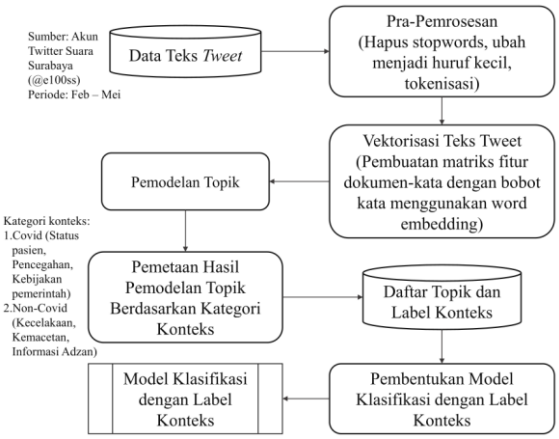
COVID-19 telah menjadi isu pandemi global sehingga sebagai media sosial yang *up-to-date* dalam penyebaran informasi, banyak akun *twitter* telah membahas topik tersebut. Sebagai contoh akun Suara Surabaya (SS) berisi informasi COVID-19 disamping informasi lain tentang Kota Surabaya seperti lalu lintas. Ini memengaruhi pengguna yang ingin mengetahui informasi COVID-19. Oleh karena itu, untuk memudahkan pengguna dalam memahami informasi penting COVID-19, maka perlu dilakukan rangkuman informasi teks *tweet* di akun SS. Metode *extractive summarization* menghasilkan ringkasan teks yang terdiri dari tiga tahap yaitu representasi teks, urutan kata, dan pemilihan kata (Wang et al., 2018) seperti yang dilakukan pada peringkasan ulasan hotel dari TripAdvisor.com untuk membantu wisatawan dalam memilih hotel (He et al., 2017). Penelitian lain juga menghasilkan ringkasan dari teks *tweet* berdasarkan tren (Sharifi et al., 2014) untuk mempermudah pengguna dalam memahami adanya banyak topik. Permasalahan peringkasan teks *tweet* adalah keterbatasan karakter namun memiliki banyak variasi topik.

Pada penelitian terdahulu untuk mengetahui informasi topik dilakukan pengklasteran k-medoid untuk memastikan *center* adalah kata dalam teks sebelum peringkasan (Purwitasari et al., 2016). Hal yang serupa dilakukan pada peringkasan dokumen dengan *Multi-Level Divisive Coefficient* untuk mengelompokkan kalimat secara hirarki pada dokumen laporan praktik industri yang memiliki topik sama (Mustamiin et al., 2018). Pada umumnya peringkasan otomatis teks *Twitter* menggunakan data *tweet* dari tagar atau *trending* (Jiwanggi & Adriani, 2016) (Gao et al., 2017) karena data *tweet* yang diperoleh dari hashtag kemungkinan besar memiliki bahasan yang sama (Gao et al., 2017). Akan tetapi data akun twitter SS terdiri dari berbagai jenis informasi, sehingga data yang tidak sesuai dengan topik akan menjadi *noise* dalam peringkasan otomatis. Pemodelan topik merupakan teknik analisis struktur tertentu dari kumpulan kata atau kalimat dalam sebuah dokumen (Prabhakar Kaila et al., 2020). Topik diartikan sebagai kumpulan kata yang memiliki frekuensi tinggi dibandingkan dengan kata lain dan dapan menjelaskan isi dokumen secara singkat (Hannigan et al., 2019).

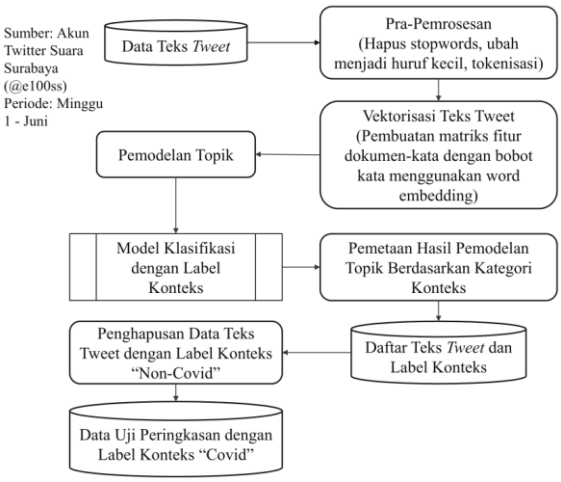
Oleh karena itu, penelitian ini mengusulkan peringkasan dengan penghapusan data *tweet* yang tidak sesuai topik COVID-19 sebelum digabung. Penghapusan data non topik memanfaatkan metode pemodelan dan klasifikasi topik.

Pada penelitian ini digunakan *Latent Semantic Indexing* (LSI), *Latent Dirichlet Allocation* (LDA), dan *Hierarchical Dirichlet Process* (HDP) sebagai metode pemodelan topik. LSI atau disebut juga *Latent Semantic Analysis* (LSA) merupakan metode ekstraksi dan representasi kata dari teks dengan menggunakan perhitungan statis pada dokumen atau data berukuran besar (Mohammed & Al-augby, 2020). LSA merupakan metode yang cepat dan paling mudah digunakan diantara metode pemodelan topik (Qomariyah et al., 2019). LDA merupakan metode yang umum digunakan pada pemodelan topik, sehingga penelitian ini membandingkan metode LDA dengan metode pemodelan topik yang lain. LDA adalah model yang menghitung probabilitas setiap kata terhadap topik secara acak (Hagen, 2018). LDA memiliki dua tahap yaitu pemodelan setiap kata kedalam topik dan perhitungan probabilitas setiap topik secara berulang. Metode LDA efektif digunakan pada penelitian teks *tweet* yang menggabungkan metode ini dan analisa sentimen (Yang & Zhang, 2018). Metode LDA juga digunakan dalam menganalisa sentimen masyarakat dalam berbagai bidang industri, termasuk ekonomi dalam hal pemasaran produk (Liu et al., 2017). HDP mengasumsikan dokumen sebagai kelompok dari kata yang diobservasi dan campuran komponen sebagai topik yang tersebar dalam dokumen dan setiap dokumen memiliki proporsi penyebaran topik yang berbeda-beda (Park & Oh, 2017). Metode HDP efektif digunakan dalam mendeteksi sub-topik atau sub-story dari sekumpulan *tweet* dalam suatu waktu (Srijith et al., 2017). HDP merupakan metode pengembangan yang lebih kompleks dari LDA.

Metode peringkasan yang digunakan pada penelitian ini adalah TextRank dan Okapi BM25. Metode TextRank merupakan metode berbasis graf yang merangkingkan kata pada pemrosesan teks (Barrios et al., 2016). TextRank digunakan dalam beberapa penelitian termasuk salah satunya pada peringkasan teks *tweet* untuk menganalisa ketertarikan pengguna. Metode ini menghasilkan rangking dari setiap kata yang selanjutnya dipetakan pada koleksi data ketertarikan pengguna (Niu & Shen, 2019). Metode TextRank merupakan metode peringkasan otomatis yang umum digunakan, oleh karena itu metode ini digunakan sebagai pembanding dengan metode yang lebih kompleks. Okapi BM25 merangkingkan kata menggunakan mesin pencari untuk mengukur tingkat relevansi data dengan kata pencarian atau *query*. BM25 menghitung jumlah kata yang muncul sesuai kata pencarian pada setiap dokumen tanpa mempertimbangkan kedekatan antar dokumen (Kadhim, 2019). Okapi BM25 dapat digunakan pada pencarian kata dalam proses pengkategorian teks dengan menggunakan KNN sebagai metode



Gambar 1. Alur Proses Pembuatan Model Klasifikasi



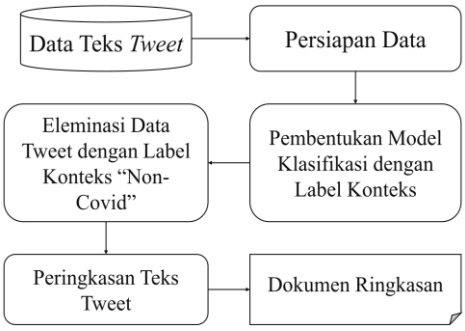
Gambar 2. Alur Proses Eliminasi Data

klasifikasi (Ghawi & Pfeffer, 2019). Metode Okapi BM25 lebih kompleks dibandingkan metode TextRank biasa, dimana Okapi BM25 memiliki bobot pada setiap kalimat penyusun ringkasan.

Masalah utama dalam penelitian ini adalah data dari akun twitter Suara Surabaya memiliki topik yang beragam. Untuk membuat informasi COVID-19 lebih mudah dipahami, dibuat ringkasan otomatis yang berfokus pada COVID-19. Berbagai topik dalam *tweet* Suara Surabaya dapat mempengaruhi hasil peringkasan sehingga dibutuhkan tahap eliminasi data yang pada penelitian ini menggunakan pemodelan topik.

2 METODE PENELITIAN

Proses utama yang diusulkan dibagi menjadi empat proses utama, (1) persiapan data, meliputi pra-pemrosesan data *tweet* dan proses vektorisasi teks atau *word embedding*. (2) Pembentukan model klasifikasi dengan label konteks. (3) Proses eliminasi terdiri dari pemodelan topik, klasifikasi topik, dan eliminasi data NON-COVID-19. (4) Penggabungan



Gambar 3. Tahapan pada peringkasan teks *tweet* dengan eliminasi data non-topik

Tabel 1. Contoh Data *Tweet* dan Karakter Khusus Teks *Tweet*

No.	*Tweet*
1.	17.23: Adzan maghrib telah berkumandang untuk wilayah Surabaya dan sekitarnya. **(odp-rt)**
2.	Kata Ririek Adriyansah Direktur Utama Telkom, pembicaraan mengarah pada melindungi pelanggan konten, termasuk kebijakan penurunan konten. Diharapkan Netflix bisa memenuhi peraturan yang berlaku, supaya bisa jalin kemitraan dengan Telkom. **(news/odp-pr)**
3.	Sudah ada pertolongan ke korban. Info sudah diteruskan ke petugas. **(hm)**

seluruh *tweet* hasil eliminasi menjadi satu dokumen dan proses peringkasan otomatis.

2.1 Persiapan Data

Data *tweet* mentah yang didapatkan dari Twitter API dibagi menjadi dua bagian untuk proses pembuatan model klasifikasi dan proses eliminasi data. Data untuk pembentukan model klasifikasi diambil dari data *tweet* bulan Februari hingga Mei dengan jumlah kurang lebih 5000 data dan data untuk proses eliminasi diambil dari data *tweet* bulan Juni. Kedua data tersebut melalui proses penghapusan beberapa karakter sebagai bentuk pembersihan data. Pra-pemrosesan tidak hanya membersihkan karakter seperti pada umumnya (angka dan tanda baca) tetapi juga karakter khusus teks *tweet*. Data *tweet* memiliki beberapa karakter khusus seperti *mention*, *hashtag* dan *link* yang dapat mengganggu proses pemodelan topik. Pemodelan topik merupakan metode yang sangat dipengaruhi oleh kata dan frekuensinya. Data *tweet* yang memiliki karakter khusus dan tersebar hampir diseluruh *tweet* dapat memperburuk hasil pemodelan topik.

Selain itu, data *tweet* Suara Surabaya juga memiliki karakter khusus tersendiri pada akun ini seperti yang ditunjukkan pada Tabel 1. Contoh Data *Tweet* dan Karakter Khusus Teks *Tweet*sehingga pra-pemrosesan dilakukan secara beberapa tahap dan memiliki banyak kondisi. Penghapusan karakter-karakter khusus tersebut dilakukan

Tabel 2. Perlakuan Penting Metode Pemodelan Topik

	LSI	**LDA**	**HDP**
Jumlah Topik	Jumlah topik yang dihasilkan untuk LSI berkisar antara 100 – 500 topik untuk dapat menghasilkan nilai koherensi yang baik	Jumlah topik untuk LDA jauh lebih kecil dibandingkan LSI, berkisar 10 – 20 untuk mendapatkan nilai koherensi yang baik dan kata-kata pada topik tidak terduplikasi. Jika jumlah topik terlalu besar menyebabkan kata-kata pada topik terduplikasi dan hal ini dapat merusak hasil klasifikasi	Pada model HDP tidak ditentukan jumlah topik yang dihasilkan karena HDP menentukan sendiri jumlah topik yang dihasilkan yaitu 150 topik. Nilai koherensi yang dihasilkan dari model ini cenderung lebih tinggi dibandingkan model yang lain.
Hasil Nilai Koherensi	Nilai koherensi LSI cenderung stabil berkisar antara 0.4 – 0.6. Apabila nilai koherensi jauh dibawah angka tersebut maka kemungkinan besar jumlah topik yang dimasukkan kurang tepat.	Sama seperti LSI, hasil nilai koherensi LDA juga cenderung stabil. Namun angka stabil belum tentu menghasilkan topik yang bagus. Ada kemungkinan angka koherensi yang tinggi menghasilkan topik yang terduplikasi. Sehingga pada metode LDA diperlukan pengecekkan hasil topik.	Berbeda dari LSI dan LDA, metode HDP menghasilkan nilai koherensi yang tinggi berkisar antara 0.7 – 0.8. Nilai tinggi juga belum tentu menghasilkan topik yang sangat baik. Dari analisa pada penelitian ini ditemukan bahwa topik yang dihasilkan metode HDP lebih umum dibandingkan metode yang lain. Umum yang dimaksudkan adalah mencakup hampir keseluruhan kata dalam dokumen. Dalam kasus ini, topik yang dihasilkan masih banyak yang bukan merupakan topik "Covid".

berdasarkan analisa manual dari *tweet* secara keseluruhan.

Penghapusan karakter juga meliputi tanda baca, angka dan beberapa karakter umum lainnya. Apabila hasil penghapusan karakter-karakter tersebut menyebabkan kekosongan data maka data row tersebut akan dihapus dari keseluruhan data. Sehingga memungkinkan jumlah data akan berkurang.

Hasil pra-pemrosesan merupakan data yang telah bersih dari karakter umum maupun khusus diproses kedalam bentuk vektor menggunakan metode *word embedding*. Dalam rangka menemukan kombinasi metode-metode yang menghasilkan peringkasan yang baik dan cocok pada data *tweet* Suara Surabaya dipilih beberapa metode *word embedding*, yaitu TFIDF dan Word2Vec. Metode TFIDF mengubah kata menjadi vektor berdasarkan jumlah frekuensi kata yang muncul dalam suatu dokumen (Maryam & Ali, 2019). Sedangkan Word2Vec menggabungkan metode CBOW dan Skip-Gram dalam mengubah kata menjadi vektor (Grant et al., 2018). Word2Vec merepresentasikan kata secara semantik. Word2Vec menggunakan metode dasar neural network yang memiliki tiga layer (*input*, *hidden*, dan *output*). Sehingga metode ini lebih kompleks dibandingkan metode konvensional TFIDF.

Metode *word embedding* digunakan sebagai bagian dari uji coba. Metode word embedding utama yang digunakan adalah TFIDF sebagai metode konvensional yang umum digunakan oleh metode pemodelan topik. Tujuan digunakannya metode Word2Vec adalah sebagai pembanding TFIDF dalam hal pemodelan topik.

Hasil perbandingan metode TFIDF dan Word2Vec dilihat dari hasil koherensi dan topik dari setiap metode pemodelan topik.

2.2 Pembentukan Model Klasifikasi

Gambar 1 menunjukkan proses pembentukan model klasifikasi yang yang digunakan pada proses eliminasi data. Seperti dijelaskan pada tahap sebelumnya data yang digunakan pada tahap ini adalah data untuk proses pembentukan model klasifikasi. Sehingga data awal dari proses ini adalah data vektor hasil *word embedding* pada tahap persiapan data.

Data vektor dari teks *tweet* diubah kedalam bentuk korpus yang merupakan inputan untuk pemodelan topik (LSI, LDA dan HDP). Hasil topik dari setiap metode kemudian digabung menjadi satu data. Pada penelitian ini ditentukan dua label konteks yaitu "Covid" dan "Non-Covid". Pelabelan konteks secara manual dilakukan terhadap data hasil topik yang telah digabung. Dari proses tersebut terbentuk data baru berisi daftar topik beserta label konteksnya. Data yang terbentuk sejumlah 850 data. Data tersebut digunakan sebagai data training pada pembuatan model klasifikasi menggunakan metode supervised learning, Neural Network. Penentuan metode supervised learning berdasarkan pengujian terhadap beberapa metode lain yaitu Decision Tree dan Random Forest. Decision Tree dipilih karena umum dan mudah digunakan dalam klasifikasi, sedangkan Random Forest dipilih karena merupakan gabungan Pohon Keputusan (Lan et al., 2020). Neural Network atau Jaringan Saraf Tiruan merupakan metode yang lebih kompleks dibandingkan Pohon Keputusan dan Random Forest dan menggunakan beberapa layer pada proses klasifikasi (Lu et al., 2019). Dari percobaan beberapa metode tersebut Neural Network menghasilkan nilai akurasi tertinggi dari setiap percobaan. Sehingga metode Neural Network dengan parameter hidden layer sebanyak 100 dipilih sebagai metode klasifikasi topik. Model klasifikasi

ini disimpan untuk digunakan pada proses klasifikasi topik dari data uji peringkasan.

2.3 Eliminasi Data

Gambar 2 menunjukkan proses eliminasi data yang menggunakan data teks *tweet* minggu pertama bulan Juni 2020. Dari tahap persiapan data telah didapatkan data vektor hasil *word embedding*. Data vektor tersebut yang menjadi inputan pada metode pemodelan topik. Pada penelitian ini digunakan beberapa metode pemodelan topik yaitu LSI, LDA dan HDP. Alasan penggunaan beberapa metode sama dengan *word embedding* adalah untuk membandingkan kombinasi metode yang tepat. *Package* gensim (Gensim, 2020) digunakan dalam memproses pemodelan topik dan word embedding.

Daftar topik yang yang dihasilkan dari proses pemodelan topik data uji menjadi data testing untuk proses klasifikasi topik. Model klasifikasi yang disimpan sebelumnya diujikan kepada data topik yang baru. Hasil klasifikasi merupakan data topik yang baru beserta label konteksnya. Hasil topik yang telah memiliki label konteks dipetakan pada setiap data teks *tweet* Data *tweet* dilabeli sesuai dengan label konteks dari topik yang merepresentasikannya. Hasil dari proses ini berupa daftar data teks *tweet* beserta label konteksnya dalam hal ini "Covid" dan "Non-Covid".

Dari daftar teks *tweet* tersebut penghapusan *tweet* yang memiliki label konteks "Non-Covid". Hasil akhir dari proses eliminasi data berupa data *tweet* dengan konteks "Covid".

2.4 Peringkasan Otomatis

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl}} \quad (1)$$

$$IDF(q_i) = ln(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}+1) \quad (2)$$

Proses pringkasan otomatis menggunakan data hasil eliminasi yang sudah berisi data "*Covid*" saja. Pada proses ini terdapat pra-pemrosesan sederhana yang meliputi pembuangan karakter khusus twitter tanpa melakukan penghapusan angka dan tanda baca.

Setiap *tweet* yang berupa kalimat digabung menjadi satu paragraph atau umumnya disebut dokumen, dan kemudian dilakukan peringkasan otomatis menggunakan metode TextRank dan Okapi BM25. Perhitungan rangking Okapi BM25 dijelaskan pada Persamaan 1 dimana D adalah panjang dokumen, Q adalah *query* yang diberikan, q_i merupakan kata kunci ke-i dari setiap dokumen, k_1 dan b merupakan variable bebas yang digunakan untuk optimasi. Persamaan 2 menjelaskan perhitungan IDF yang digunakan pada Persamaan 1, dimana N merupakan jumlah dokumen dan $n(q_i)$ merupakan jumlah dokumen yang mengandung q_i. Kombinasi metode *word embedding*, pemodelan topik dan peringkasan otomatis digunakan untuk menemukan metode yang paling tepat dalam menghasilkan ringkasan yang terbaik.

3 HASIL DAN PEMBAHASAN

3.1 Persiapan Data

Data teks *tweet* didapatkan dari akun twitter Suara Surabaya (@e100ss) melalui Twitter API. Data yang diperoleh berjumlah kurang lebih 5000 data teks tweet dari bulan Februari hingga Juni. Data tersebut dibagi menjadi dua bagian. Data pertama digunakan sebagai data mentah untuk pembentukan model klasifikasi dan data kedua digunakan sebagai data uji peringkasan.

Setiap data baik data untuk pembentukan model klasifikasi maupun data uji peringkasan melalui pra-pemrosesan dan *word embedding*. Pada tahap pra-pemrosesan, data teks *tweet* mentah diubah menjadi huruf kecil dan dihapus karakter-karakter umum seperti angka dan tanda baca atau biasa disebut *stopwords*. Selain melalui pra-pemrosesan pada umumnya, data teks *tweet* melalui proses menghilangkan karakter khusus teks *tweet* seperti *mention*, *hashtag*, dan beberapa karakter lainnya. Menurut hasil analisa manual sejauh ini terdapat 14 karakter khusus teks *tweet* yang tidak digunakan.

Setelah proses penghapusan karakter *tweet* tersebut menjadi *tweet* kosong. Kekosongan ini berakibat *error* pada proses selanjutnya sehingga dilakukan proses penghapusan data yang kosong akibat pra-pemrosesan.

Contoh kasus terjadi pada pra-pemrosesan data uji peringkasan bulan Juni minggu pertama dengan jumlah awal sebanyak 336 *tweet*. Salah satu *tweet* berisikan angka tanpa ada kata. Setelah mengalami pra-pemrosesan yang meliputi penghapusan angka, *tweet* tersebut menjadi kosong. Oleh karena itu, dilakukan pembuangan data *tweet* yang kosong, sehingga dari 336 data *tweet*, berkurang menjadi 335 data *tweet*.

Data *tweet* yang telah bersih diubah menjadi bentuk vektor atau angka menggunakan metode word embedding. Proses ini diperlukan karena metode pemodelan topik hanya dapat membaca angka bukan kata. Sehingga kata-kata *tweet* diubah menjadi angka menggunakan word embedding. Metode word embedding yang digunakan adalah TFIDF dan Word2Vec. Pemilihan beberapa metode word embedding bertujuan untuk membandingkan metode word embedding yang menghasilkan

Tabel 3. Skenario Uji Coba Eliminasi Data

No.	Skenario Uji Coba
1.	Perbandingan nilai koherensi pemodelan topik dari setiap metode *word embedding*
2.	Perbandingan nilai koherensi dari setiap metode pemodelan topik
3.	Perbandingan kumpulan kata dari setiap topik hasil pemodelan topik setiap metode *word embedding*
4.	Perbandingan hasil topik setiap metode pemodelan topik

ringkasan yang terbaik. Hal ini dilakukan karena setiap metode word embedding memberikan perlakuan yang berbeda untuk setiap katanya, dengan begitu hasil vektor setiap metode akan berbeda. Perbedaan ini yang dapat menjadi pembanding untuk proses selanjutnya, metode word embedding mana yang dapat menghasilkan topik yang terbaik. Setiap proses pada penelitian ini menentukan hasil akhir ringkasan.

3.2 Pembuatan Model Klasifikasi

Proses pembuatan model klasifikasi merupakan proses pembuatan data training dan model klasifikasi yang melibatkan atau menggunakan pemodelan topik. Pada tahap ini data yang digunakan berjumlah kurang lebih 4000 data yang melalui pra-pemrosesan dan vektorisasi. Dari 4000 data yang telah divektorisasi menjadi inputan pada pemodelan topik. Tahap ini menggunakan metode LSI, LDA dan HDP sebagai metode penghasil topik. Seluruh vektor atau yang disebut korpus diinputkan kedalam metode pemodelan topik dan menghasilkan jumlah topik yang berbeda-beda dari setiap metodenya. Keseluruhan hasil topik digabung menjadi daftar baru sejumlah 850 topik.

Data topik tersebut menjadi data belajar pada proses pembentukan model klasifikasi. Model klasifikasi menggunakan metode Neural Network disimpan untuk digunakan pada tahap klasifikasi topik hasil pemodelan topik data uji.

3.3 Eliminasi Data

Hasil word embedding konvensional atau TFIDF yang berupa vektor menjadi inputan metode pemodelan topik. Metode pemodelan topik LSI, LDA dan HDP digunakan sebagai pembanding pada setiap uji coba. Proses pemodelan topik memerlukan nilai inputan berupa jumlah topik yang dihasilkan. Jumlah topik yang dihasilkan mempengaruhi nilai koherensi dari hasil pemodelan topik. Tabel 2 menunjukkan poin-poin perbedaan perlakuan dari setiap model pada tiap datanya dan untuk mencapai hasil yang maksimal diperlukan penyesuaian dengan setiap metode.

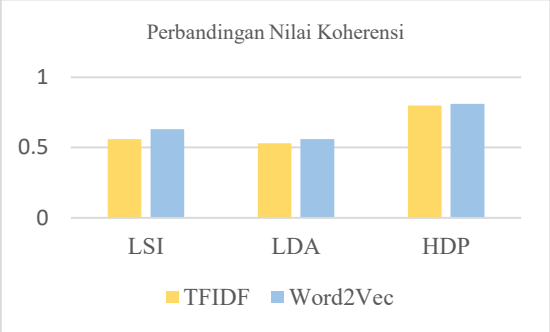
Pada penelitian ini dilakukan beberapa uji coba dan pengamatan hasil uji coba. Tabel 3 menunjukkan skenario uji coba yang dilakukan dan diamati pada proses eliminasi data. Skenario pertama adalah membandingkan hasil nilai koherensi setiap metode word embedding dan menganalisa hasilnya. Percobaan pertama menggunakan korpus vektor TFIDF sedangkan percobaan kedua dengan korpus vektor Word2Vec. Kedua percobaan ini menunjukkan hasil yang berbeda. Gambar 4 menunjukkan perbedaan antara hasil pemodelan topik dari TFIDF dan Word2Vec tidak signifikan. Artinya metode word embedding tidak banyak mempengaruhi hasil koherensi pemodelan topik.

Tabel 5 menunjukkan bahwa nilai koherensi dari TFIDF dan Word2Vec memiliki perbedaan yang kecil, perbedaan yang signifikan terjadi pada metode pemodelan topik.

Skenario kedua adalah membandingkan hasil koherensi dari setiap metode pemodelan topik.

Tabel 4. Contoh Hasil Topik Setiap Metode Pemodelan Topik

LSI		LDA		HDP	
Word2Vec	**TFIDF**	**Word2Vec**	**TFIDF**	**Word2Vec**	**TFIDF**
bypass info terbakar ss ssinfo dunia juni command pendengar km	rapid test pasien sembuh covid massal jaga jarak tangan masker	surabaya waru selamat info diteruskan test update kecelakaan normal pendengar	point check normal sim pelayanan menyesuaikan bundaran singa gisel putih	seru cari tugas takut kedurus mengatasi flashing rs ditutup virus udara pekerja	bermasker corona taman catat bi bangsa model silaturahmi virus perusahaan
angin ss km masuk kecelakaan truk dunia pendengar darah malam	hoax silakan dibaca mas mbak hubungi ceritakan rapid test padat	surabaya selamat waru info diteruskan normal rapid pendengar melaporkan	covid layanan magrib terjangkit streaming smart bersepeda warga maghrib balita	rochester kiri waspada film menunggu the daerah nico mentan rawan tokoh uk	polda posisi sesuai hubungi susanto pdp kabupatenkota tapera kencana ppdb
virus gejala darah physical orang tes tangan command dunia malam	padat waru arah tol korban motor info check point kecelakaan	surabaya selamat waru diteruskan ibadah normal covid pendengar rapid	sembuh pasien milik pandemi mas covid mencari hoax kesembuhan nilai	jatim wabah juara rujukan ibadah akbarmasjid salah bosen covid untag	haji dijalankan parit transisi rp arif total masjid memulihkan tropis



Gambar 4. Perbandingan Hasil Metode Word Embedding pada Setiap Metode Pemodelan Topik

Tabel 5. Hasil Nilai Koherensi dan Akurasi Setiap Kombinasi Word Embedding dan Pemodelan Topik

Word embedding	Pemodelan topik	Jumlah topik	Nilai koherensi	Akurasi
TFIDF	LSI	200	0.56	93.0%
TFIDF	LDA	25	0.53	88.0%
TFIDF	HDP	150	0.80	94.0%
Word2Vec	LSI	200	0.63	87%
Word2Vec	LDA	25	0.56	76%
Word2Vec	HDP	150	0.81	94.6%

Seperti pada skenario pertama, ditunjukkan bahwa metode word embedding memberikan pengaruh yang cukup kecil pada hasil pemodelan topik. Pada skenario ini hasil yang ditunjukkan lebih signifikan. Nilai koherensi metode LSI dan LDA cenderung lebih stabil dibandingkan metode HDP. Dalam hal perhitungan koherensi, nilai yang terlalu besar dan terlalu kecil menunjukkan hasil yang kurang baik. Metode HDP menunjukkan nilai koherensi yang terlalu besar jauh dibandingkan dengan LSI dan LDA.

Tabel 5 menunjukkan detail hasil koherensi tiap metode *word embedding* dan pemodelan topik dimana setiap metode HDP baik dari korpus Word2Vec maupun TFIDF memiliki nilai koherensi yang besar. Hal ini dikarenakan metode HDP tidak melakukan proses berdasarkan jumlah topik yang diinputkan. Metode HDP memproses keseluruhan data dan menghasilkan topik dengan jumlah maksimal 150. Jumlah topik yang dihasilkan oleh HDP tidak berpengaruh besar terhadap hasil koherensi. Metode HDP dapat mempertimbangkan maksimal jumlah kata yang ditampilkan pada setiap topik. Apabila tidak ada batasan, jumlah kata dalam satu topik cenderung lebih dari 15 kata. Jika dibandingkan dengan hasil topik dari metode yang lain, jumlah kata dalam satu topik yang terlalu banyak akan mengganggu hasil klasifikasi pada proses selanjutnya. Sehingga, ditentukan maksimal jumlah kata dalam satu topik adalah 12 kata. Penentuan ini dilakukan berdasarkan pada jumlah kata yang dihasilkan dalam satu topik pada hasil topik metode LSI dan LDA.

Skenario ketiga adalah membandingkan hasil topik setiap metode pemodelan topik berdasarkan metode word embedding. Dibandingkan dengan dua skenario sebelumnya yang berdasar pada hasil koherensi, perbedaan signfikan muncul pada skenario ini berdasarkan hasil topik. Tabel 4 merupakan contoh hasil topik dari setiap metode. Hasil Word2Vec menunjukkan adanya duplikasi kata pada setiap topik sedangkan metode TFIDF menghasilkan topik yang lebih bervariasi. Meski nilai koherensi Word2Vec lebih tinggi disbanding TFIDF tidak berarti hasil topik lebih baik dibandingkan TFIDF.

Skenario keempat adalah membandingkan hasil topik dari setiap metode pemodelan topik. Dari hasil topik antara LSI, LDA dan HDP hanya LDA yang memiliki duplikasi kata di setiap topiknya. Pada LDA hanya Word2Vec yang mengalami duplikasi. Berbeda dari skenario sebelumnya, pada skenario ini dapat dilihat hasil bahwa metode word embedding mempengaruhi hasil topik dari pemodelan topik.

Hasil topik dari metode pemodelan topik lain memiliki hasil yang bervariasi meskipun menggunakan korpus Word2Vec. Pengaruh Word2Vec hanya terjadi pada metode LDA. Hal dimungkinkan karena metode LDA merepresentasikan teks *tweet* kedalam satu kata, LDA mengurutkan kata dari yang terbesar ke terkecil untuk ditampilkan pada satu topik. Dalam kasus ini, LDA menemukan vektor kata "surabaya" memiliki nilai terbesar diantara semua kata, oleh karena itu kata "surabaya" muncul disetiap topik yang dihasilkan oleh LDA. Demikian juga yang terjadi pada beberapa kata dalam satu topik yang terduplikasi pada topik yang lain.

3.4 Klasifikasi Topik dengan Label Konteks

Proses selanjutnya adalah mengklasifikasi hasil topik dari setiap metode pemodelan topik menggunakan model klasifikasi yang telah disimpan sebelumnya. Pada

Tabel 5 kolom akurasi ditunjukkan hasil akurasi klasifikasi topik dari setiap kombinasi word embedding dan pemodelan topik. Hasil akurasi tersebut dapat dikatakan sangat tinggi. Pada penelitian ini dilakukan beberapa uji coba menggunakan beberapa metode supervised learning, dan hasil dari semua uji coba menunjukkan Neural Network memiliki akurasi terbaik. Oleh karena itu, selanjutnya metode Neural Network digunakan sebagai metode klasifikasi topik.

Tabel 6. Hasil Perhitungan ROUGE Setiap Ringkasan

Metode	Rouge-L
TFIDF + LSI + TextRank	0.59
TFIDF + LSI + Okapi BM25	0.56
TFIDF + LDA + TextRank	0.45
TFIDF + LDA + Okapi BM25	0.35
TFIDF + HDP + TextRank	0.50
TFIDF + HDP + Okapi BM25	0.49
Word2Vec + LSI + TextRank	0.67
Word2Vec + LSI + Okapi BM25	0.53
Word2Vec + LDA + TextRank	0.50
Word2Vec + LDA + Okapi BM25	0.52
Word2Vec + HDP + TextRank	0.65
Word2Vec + HDP + Okapi BM25	0.47

Dari Tabel 5 juga didapatkan hasil bahwa metode LDA memiliki nilai akurasi terkecil baik dari korpus TFIDF maupun Word2Vec. Hasil akurasi kecil ini dimungkinkan karena jumlah topik yang kecil dibandingkan dengan LSI dan HDP yang memiliki jumlah topik jauh lebih besar. Dari penelitian ini didapatkan bahwa jumlah topik yang kecil atau tidak seimbang dengan metode yang lain mengakibatkan perbedaan yang signifikan pada hasil pemodelan topik.

Topik yang telah diklasifikasi atau telah memiliki kelas, kemudian dicocokkan dengan data setiap *tweet*. Setiap *tweet* nilai vektor untuk setiap topiknya. Dari nilai vektor tersebut. Topik dengan nilai vektor tertinggi merupakan representasi topik dari *tweet* tersebut. Kelas dari *tweet* tersebut mengikuti topik yang merepresentasikannya.

Data *tweet* yang telah memiliki kelas, dipilah berdasarkan kelasnya. Data dengan kelas "Covid" dipindah kedalam data baru sehingga data dengan kelas "Non Covid" dibuang. Data baru yang terbentuk adalah data dengan topik "Covid".

3.5 Peringkasan Otomatis

Proses terakhir merupakan proses peringkasan otomatis dengan menggunakan data baru hasil eliminasi data. Pada tahap ini dilakukan peringkasan otomatis menggunakan metode TextRank dan Okapi BM25. Sebelum dilakukan proses peringkasan, seluruh data *tweet* hasil eliminasi, digabung menjadi satu dokumen. Penelitian ini menggunakan *package* gensim (Gensim, 2020) untuk metode TextRank dan Okapi BM25. Metode peringkasan menentukan hasil ringkasan berdasarkan maksimal kata yang diinputkan. Pada penelitian ini ditentukan maksimal jumlah kata adalah satu per tiga dari jumlah keseluruhan kata dalam satu dokumen.

Tabel 7. Skenario Uji Coba Peringkasan Otomatis

No.	Skenario Uji Coba
1.	Perbandingan tahap peringkasan otomatis dan pengaruh pada hasil ringkasan
2.	Perbandingan hasil ROUGE antara metode TextRank dan Okapi BM25

Beberapa uji coba dilakukan untuk mendapatkan metode yang tepat, Tabel 7 menunjukkan skenario uji coba pada proses peringkasan otomatis. Skenario pertama adalah membandingkan hasil ringkasan dua metode peringkasan otomatis. Menurut analisa manual hasil peringkasan otomatis menggunakan TextRank dan Okapi BM25 menunjukkan bahwa ringkasan TextRank memiliki kalimat-kalimat yang lebih banyak mengarah pada COVID-19 atau memiliki lebih sedikit *noise* dibandingkan Okapi BM25. Hal ini terjadi karena metode TextRank menghitung skor tiap kata dari keseluruhan dokumen, sedangkan Okapi BM25 menghitung bobot berdasarkan *query*. Sehingga kemungkinan adanya *noise* lebih besar pada metode Okapi BM25.

Tabel 8 merupakan hasil ringkasan metode Word2Vec, LSI dan TextRank. Dari ringkasan tersebut masih terdapat beberapa kalimat yang tidak berkaitan dengan COVID-19, contohnya informasi mogok kendaraan dan informasi adzan. Hal ini dapat terjadi karena beberapa hal yang telah disebut sebelumnya. Terlalu kecil atau besarnya nilai koherensi dapat menjadi salah satu pemicu tidak tercakupnya informasi diluar COVID-19 pada proses klasifikasi. Selain itu jumlah topik yang

Tabel 8. Contoh Hasil Ringkasan

Komposisi-komposisi hits barat yang dinyanyikan grup akan menemani malam Anda di bagian pertama dan kedua. Informasi dari Command Center Surabaya, hari ini (Minggu 7/6/20) ada Rapid Test massal Covid-19 di depan Kantor SCTV Surabaya Jl Patimura, dan di Kantor Kecamatan Kenjeran Jl HM Nur no 350. Tapera Jamin Dana Peserta yang Miliki Rumah Dikembalikan Saat Pensiun. **16.20: Info awal: Truk tronton mogok di TL Balongbendo - Krian, posisi kendaraan nyaris menutup semua lajur. 18.36: Adzan Isya untuk wilayah Surabaya dan sekitarnya**. Selamat menunaikan ibadah Sholat Isya, kawan. Andriono pendengar SS melaporkan, Truk yang terguling sudah dievakuasi**. 14.55: Adzan Ashar telah berkumandang untuk wilayah Surabaya dan sekitarnya.** BIN Temukan 136 Orang Reaktif Pada Hari Keenam Tes Cepat di Surabaya. KONI Pusat Bantu Kembalikan Atlet ke Pelatnas Saat New Normal. **17.25: Adzan Maghrib telah berkumandang untuk wilayah Surabaya dan sekitarnya.** Surabaya jadi satu diantara 25 kota yang disiapkan untuk melaksanakan tatanan normal baru. Yang berarti, kita hidup dg mematuhi protokol kesehatan untuk memutus penyebaran Covid-19. Ada dua sepeda motor yang terbakar dan satu orang tergeletak di lokasi. Keputusan tersebut diambil mempertimbangkan keselamatan jemaah dan petugas haji yang menjadi prioritas di masa pandemi Covid-19. Gisel, Singa Putih yang Lahir di Tengah Pandemi Covid-19. **18.34: Adzan Isya telah berkumandang untuk wilayah Surabaya dan sekitarnya.** 127 Anak Surabaya Terpapar Covid-19, 36 Diantaranya Balita. Dalam keputusan Menkes, orang dengan gejala batuk dilarang masuk ke area tempat kerja. Beredar informasi lewat video yang tersebar di WhatsApp yang menyebutkan bahwa mulai Senin (1/6) aktivitas harus berhenti sekitar pkl 14.00 WIB. Dokter Senior Italia Sebut Virus Corona Mulai Melemah.

rendah juga memungkinkan rendahnya akurasi hasil klasifikasi.

Hasil klasifikasi memiliki pengaruh yang besar dalam penelitian ini, sebab eliminasi data dilakukan berdasarkan kelas topik yang pelabelan konteksnya didapatkan dari hasil klasifikasi. Semakin kecil akurasi klasifikasi maka kemungkinan adanya data yang tidak ikut tereliminasi meski bukan merupakan data COVID-19 semakin tinggi.

Skenario kedua adalah membandingkan hasil pengujian ringkasan menggunakan metode ROUGE. Hasil ringkasan menunjukkan adanya *noise* meski jumlahnya berbeda. Perbedaan jumlah *noise* ini menunjukkan efektif tidaknya kombinasi metode yang digunakan. Metode ROUGE digunakan untuk menguji ringkasan. Metode ini mengguji hasil peringkasan yang menggunakan referensi sebagai pembanding (Joshi et al., 2019). Dari setiap kombinasi metode dibuat referensi secara manual. Metode ROUGE yang digunakan adalah ROUGE-L karena mempertimbangkan kemungkinan jumlah kata yang sama sebanyak mungkin (*longest common subsequence*).

Tabel 6 menunjukkan hasil pengujian dari setiap kombinasi. Kombinasi Word2Vec + LSI + TextRank memberikan hasil ringkasan terbaik sedangkan metode TFIDF + LDA + Okapi BM25 menunjukkan hasil terendah berdasarkan nilai ROUGE-L. Uji coba eliminasi data non-topik yang dilakukan cenderung menghasilkan ringkasan lebih baik dibanding proses yang sama tanpa eliminasi (Purwitasari et al., 2016). Sehingga peringkasan sejumlah teks dengan konteks bahasan yang sama akan menghasilkan ringkasan lebih baik. Konteks bahasan serupa akan tergantikan oleh pemilihan teks *tweet* dengan *trending topic* sama (Gao et al., 2017). Hal tersebut berbeda dengan uji coba yang dilakukan dalam usulan penelitian ini, dimana data yang digunakan bervariasi disbanding data dari *trending topic*.

Selain itu, Tabel 6 menunjukkan bahwa metode LDA memiliki nilai yang kecil, baik TFIDF maupun Word2Vec jika dibandingkan metode LSI dan HDP. Hal ini dapat terjadi karena jumlah topik yang dihasilkan jauh lebih sedikit dibandingkan LSI dan HDP. Jumlah topik ini mempengaruhi hasil klasifikasi. Setiap proses pada penelitian mempengaruhi satu sama lain. Dari penelitian ini letak perbedaan signifikan antara LSI, LDA dan HDP adalah jumlah topik yang dihasilkan.

4 KESIMPULAN DAN SARAN

Peringkasan teks dilakukan setelah eleminasi data "Non-Covid" berdasarkan kesesuaian hasil pemodelan topik. Hasil pengujian peringkasan dengan indikator ROUGE-L dari metode pemodelan topik LSI memiliki performa paling baik. *Word embedding* pada data input pemodelan topik tidak memiliki pengaruh yang signifikan, sehingga proses input cukup dilakukan perhitungan kemunculan kata (TFIDF). Penggunaan TextRank atau Okapi BM25 memberikan hasil ringkasan yang hampir sama dengan indikator keberhasilan ROUGE-L. Sehingga pemilihan metode pemodelan topik berpengaruh kuat dalam menghasilkan teks ringkasan yang baik dan mudah dipahami dengan topik tidak beragam.

Dari penelitian ini terdapat dua tantangan utama yaitu penghapusan karakter khusus data twitter dan data *tweet* yang memiliki lebih banyak informasi diluar COVID-19. Sehingga penelitian selanjutnya dapat melakukan eksplorasi penanganan masalah tersebut yang juga dikombinasikan dengan metode pemodelan topik.

DAFTAR PUSTAKA

BARRIOS, F., LÓPEZ, F., ARGERICH, L., & WACHENCHAUZER, R. 2016. Variations of the similarity function of textrank for automated summarization. ArXiv Preprint ArXiv:1602.03606.

GAO, D., LI, W., CAI, X., ZHANG, R., & OUYANG, Y. 2017. Sequential summarization: A full view of Twitter trending topics. Social Media Content Analysis: Natural Language Processing and Beyond, 375–399.

GENSIM. 2020. Gensim. <https://radimrehurek.com/gensim> [Diakses 1 Juli 2020]

GHAWI, R., & PFEFFER, J. 2019. Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity. Open Computer Science, 9(1), 160–180.

GRANT, R. N., KUCHER, D., LEÓN, A. M., GEMMELL, J. F., RAICU, D. S.,DAN FODEH, S. J. 2018. Automatic extraction of informal topics from online suicidal ideation. BMC Bioinformatics, 19(Suppl 8).

HAGEN, L. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? Information Processing & Management, 54(6), 1292–1307.

HANNIGAN, T. R., HAANS, R. F. J., VAKILI, K., TCHALIAN, H., GLASER, V. L., WANG, M. S., KAPLAN, S., & JENNINGS, P. D. 2019. Topic modeling in management research: Rendering new theory from textual data. Academy of Management Annals, 13(2), 586–632.

HE, R., LIU, Y., YU, G., TANG, J., HU, Q., & DANG, J. 2017. Twitter summarization with social-temporal context. World Wide Web, 20(2), 267–290.

JIWANGGI, M. A., & ADRIANI, M. 2016. Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement Algorithm. Procedia Computer Science.

JOSHI, A., FIDALGO, E., ALEGRE, E., & FERNÁNDEZ-ROBLES, L. 2019. SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Systems with Applications, 129, 200–215.

KADHIM, A. I. 2019. Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF. 2019 International Conference on Advanced Science and Engineering (ICOASE), 124–128.

LAN, T., HU, H., JIANG, C., YANG, G., & ZHAO, Z. 2020. A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. Advances in Space Research.

LIU, X., BURNS, A. C., & HOU, Y. 2017. An Investigation of Brand-Related User-Generated Content on Twitter. Journal of Advertising, 46(2), 236–247. https://doi.org/10.1080/00913367.2017.1297273

LU, S., LI, Q., BAI, L., & WANG, R. 2019. Performance predictions of ground source heat pump system based on random forest and back propagation neural network models. Energy Conversion and Management, 197, 111864.

MARYAM, A., & ALI, R. 2019. Temporal TF-IDF-Based Twitter Event Summarization Incorporating Keyword Importance. In Information and Communication Technology for Intelligent Systems (pp. 559–566). Springer.

MOHAMMED, S. H., & AL-AUGBY, S. 2020. LSA & LDA Topic Modeling Classification: Comparison study on E-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1).

MUSTAMIIN, M., GHOZALI, A. L., & SIFA, M. L. 2018. Peringkasan Multi-dokumen menggunakan Metode Pengelompokkan berbasis Hirarki dengan Multi-level Divisive Coefficient. Jurnal Teknologi Informasi Dan Ilmu Komputer, 5(6), 697.

NIU, R., & SHEN, B. 2019. Microblog User Interest Mining Based on Improved TextRank Model. Journal of Computers, 30(1), 42–51.

PARK, J., & OH, H.-J. 2017. Comparison of topic modeling methods for analyzing research trends of archives management in korea: Focused on lda and hdp. Journal of Korean Library and Information Science Society, 48(4), 235–258.

PRABHAKAR KAILA, D., PRASAD, D. A. V, & OTHERS. 2020. Informational flow on Twitter--Corona virus outbreak--topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET), 11(3).

PURWITASARI, D., FATICHAH, C., ARIESHANTI, I., & HAYATIN, N. 2016. K-medoids algorithm on Indonesian Twitter feeds for clustering trending issue as important terms in news summarization. Proceedings of 2015 International Conference on Information and Communication Technology and Systems, ICTS 2015, 95–98.

QOMARIYAH, S., IRIAWAN, N., & FITHRIASARI, K. 2019. Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. AIP Conference Proceedings, 2194(1), 20093.

SHARIFI, B. P., INOUYE, D. I., & KALITA, J. K. 2014. Summarization of twitter microblogs. Computer Journal, 57(3), 378–402.

SRIJITH, P. K., HEPPLE, M., BONTCHEVA, K., & Preotiuc-Pietro, D. 2017. Sub-story detection in Twitter with hierarchical Dirichlet processes. Information Processing & Management, 53(4), 989–1003.

WANG, F., ORTON, K., WAGENSELLER, P., & XU, K. 2018. Towards understanding community interests with topic modeling. IEEE Access, 6, 24660–24668.

YANG, S., & ZHANG, H. 2018. Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. Int. J. Comput. Inf. Eng, 12, 525–529.