

# Penentuan Topik dengan *Opinion Mining* berbasis *Two-Pass Classifier* dan Bayesian dalam Peringkasan Teks Twitter

Muhammad Mirza Muttaqi<sup>1</sup>, Diana Purwitasari<sup>2,\*</sup>, Rizka Wakhidatus Sholikhah<sup>3</sup>

<sup>1</sup>Teknik Informatika, Institut Teknologi Sepuluh Nopember, Indonesia

<sup>2</sup>Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Indonesia

<sup>1</sup>mirza.18051@mhs.its.ac.id; <sup>2</sup>diana@if.its.ac.id; <sup>3</sup>wakhidatus@its.ac.id

\*corresponding author

## INFO ARTIKEL

### Sejarah Artikel

Diterima: 30 Juni 2022

Direvisi: 26 Desember 2022

Diterbitkan: 30 Desember 2022

### Kata Kunci

*opinion mining*  
penentuan topik  
peringkasan teks  
teks twitter  
*two-pass classifier*

## ABSTRAK

Kolom komentar berkepanjangan pada teks Twitter mempersulit masyarakat yang ingin mengetahui informasi terkini seperti topik tren Covid-19. Peringkasan teks menghasilkan ringkasan lebih terfokus ke dalam sebuah topik. Pada klustering teks Twitter ditemukan kalimat - kalimat atau satu *tweet* yang seharusnya berbeda kluster dengan topik yang dimaksud. Oleh karena itu, perlu dilakukan pemodelan topik dan penentuan topik untuk mempermudah pengguna mencari suatu informasi tertentu. Pada penelitian ini difokuskan pada topik Covid-19. Pemodelan topik digunakan untuk mengatur, mencari, memahami, dan meringkas sebuah teks. *Opinion Mining* digunakan untuk mengekstrak atau mengklasifikasikan polaritas sentimen. Polaritas sentimen ini berupa “positif” atau “negatif” pada suatu entitas atau aspek. Proses klasifikasi menggunakan metode *two pass classifier* untuk sentimen positif dan negatif, serta *Bayesian* sebagai metode pelabel entitas – entitas. Setelah itu, label-label tersebut dikelompokkan sehingga terbentuk topik-topik dan beberapa *tweet* yang mempunyai kemiripan entitas topik dikelompokkan ke dalam topik. Dari hasil evaluasi menggunakan *TextRank*, Okapi BM25 dan *PageRank*, proses *opinion mining* menghasilkan nilai yang lebih tinggi dibandingkan dengan *Latent Semantic Indexing (LSI)*, selisih rata-rata 0,53

## PENDAHULUAN

Salah satu *trending topic* di *Twitter* pada saat ini yaitu wabah pandemi Covid-19 atau biasa disebut dengan penyakit *coronavirus* 2019. Banyaknya *tweet* yang membahas Covid-19 ini, menyebabkan pengguna ingin mengetahui informasi ini secara menyeluruh. Namun, seringkali teks yang panjang dari *tweet* maupun kolom komentar membutuhkan waktu membaca lebih lama dan mempersulit mencari informasi kunci sebagai Penentuan Topik. Oleh karena itu, mulai dikembangkan peringkasan otomatis. Peringkasan otomatis digunakan untuk mempertahankan konten informasi dan makna keseluruhannya dengan menggunakan model semantik distribusi sebagai fitur akan memperoleh hasil yang lebih baik dalam menentukan sebuah topik [1]. Pendekatan ekstraksi topik sebagai salah satu cara model semantik juga digunakan untuk peringkasan otomatis dari kumpulan data *tweet*[2].

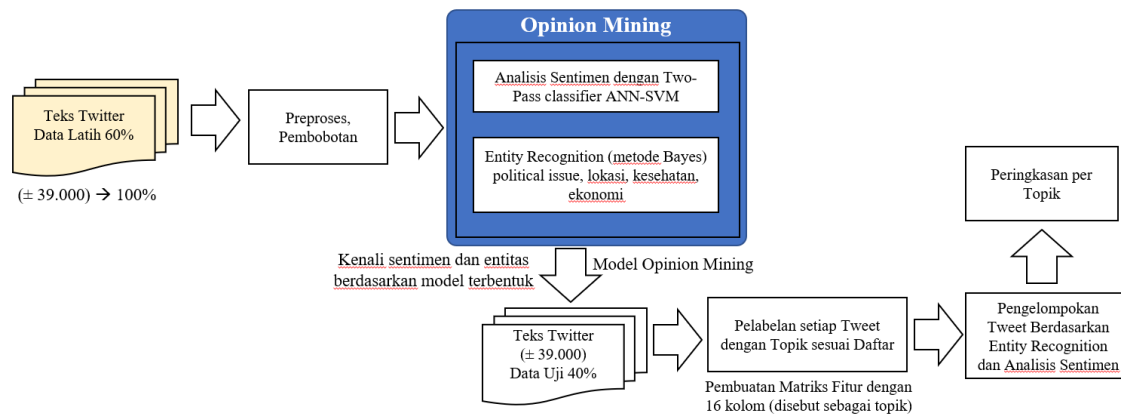
Penelitian terkait peringkasan teks membutuhkan pengklasifikasian data teks *tweet* [3], [4] dan khusus *tweet* bertema Covid-19 menggunakan melakukan eliminasi data non Covid-19 untuk memastikan kesesuaian topik [5]. Akan tetapi ada kemungkinan suatu *tweet* dapat dikategorikan menjadi multi-topik. Peringkasan otomatis tidak hanya digunakan untuk penentuan topik saja, melainkan juga dapat digunakan untuk penentuan subtopik[6]. Penentuan subtopik perlu dilakukan karena subtopik merupakan kalimat

pendukung dari topik. Sehingga beberapa kalimat yang memiliki kemiripan karakter pada topik akan dikategorikan sebagai subtopik[7]. Dalam peringkasan otomatis, data-data yang tidak sesuai dengan topik maupun topik tertentu akan menjadi noise dan merusak hasil peringkasan. Sehingga eliminasi data yang tidak sesuai dengan topik maupun topik perlu dilakukan [8], [9].

*Twitter* merupakan salah satu sumber data yang tidak terstruktur dengan jumlah data yang besar, tidak dimungkinkan untuk membaca semua teks *tweet* beserta *retweet*-nya. penggunaan *opinion mining* akan membantu menganalisis, mengelompokkan, serta menentukan secara cepat dan tepat key information [10]–[12]. Selain itu di dalam pengelompokan topik maupun topik, masih terdeteksi data *tweet* yang memuat diluar topik. Oleh karena itu, dalam penelitian ini diusulkan metode peringkasan otomatis menggunakan *opinion mining* yang dapat digunakan untuk sentiment analysis sekaligus *entity recognition*. Pada penelitian ini juga diusulkan pengelompokan topik menggunakan opinion mining menggunakan kombinasi *two-pass classifier* dan *Bayesian* agar pengelompokan topik lebih signifikan.

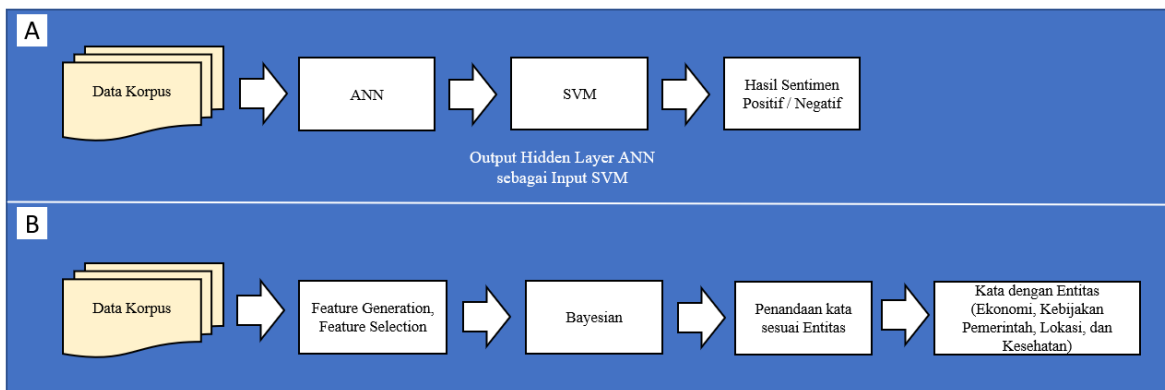
## METODE

Input dalam penelitian ini adalah kumpulan data teks *tweet* tentang COVID-19. Kemudian dilakukan preprocessing dan vektorisasi teks untuk menghasilkan korpus. Data korpus yang didapatkan, digunakan sebagai input untuk proses klasifikasi. Proses klasifikasi dibagi menjadi dua tahap yaitu sentimen positif dan negatif serta *entity recognition*. Output dari penelitian ini yaitu pengelompokan sesuai topik yang sedang dibahas dari entitas– entitas tersebut. Kemudian entitas-entitas yang sama dikelompokkan berdasarkan sentimen positif atau negatif ke dalam topik.



Gambar 1. Skema perancangan metode penelitian

Gambar 1 menjelaskan tentang perancangan metode yang digunakan pada penelitian ini. Input dalam penelitian ini adalah kumpulan data teks tweet tentang COVID-19. Kemudian dilakukan preprocessing dan vektorisasi teks untuk menghasilkan korpus. Data korpus yang didapatkan, digunakan sebagai *input* untuk proses klasifikasi. Proses klasifikasi dibagi menjadi dua tahap yaitu sentimen positif dan negatif serta *entity recognition*. *Output* dari penelitian ini yaitu pengelompokan sesuai topik yang sedang dibahas dari entitas– entitas tersebut. Kemudian entitas-entitas yang sama dikelompokkan berdasarkan sentimen positif atau negatif ke dalam topik.



Gambar 2. Skema tahap klasifikasi (a) Tahap penentuan sentimen positif dan negatif, dan (b) Tahap penentuan *entity recognition*.

Gambar 2 menjelaskan tentang tahapan klasifikasi. *Input* dalam penelitian ini adalah data *tweet* yang sudah melewati tahap *preprocessing* dan kemudian di-*word embedding*. Selanjutnya dilakukan proses klasifikasi menggunakan metode Bayesian dan *two pass classifier* untuk sentimen positif dan negatif, serta untuk memberikan label entitas – entitas. Setelah itu, label-label tersebut dikelompokkan sehingga terbentuk topik-topik dan beberapa *tweet* yang mempunyai kemiripan entitas topik. Data korpus yang didapat diolah menggunakan ANN. Kemudian dicari *output hidden layer*-nya sebagai input SVM. Hasil dari output SVM dapat mendeteksi sentimen positif dan negatif (Gambar 2.a). Sedangkan Gambar 2.b menunjukkan data korpus yang didapatkan diolah menggunakan Bayesian untuk mendapatkan *entity recognition*. *Entity recognition* dilakukan dalam beberapa tahapan yaitu pembagian data dalam hal ini data dibagi menjadi dua (data *training* dan data *testing*). Kemudian dilakukan uji kesesuaian data *training* dan data *testing*. Pencarian topik kalimat misalnya ekonomi, political issue, lokasi, dan kesehatan. Kemudian dilakukan pemodelan data baru menggunakan Bayesian. Bayesian digunakan untuk menguji tingkat keakuratan dalam klasifikasi labelisasi data baru. Entitas-entitas yang didapatkan kemudian dijadikan topik. *Tweet-tweet* yang mempunyai karakteristik yang mirip dengan suatu topik tertentu.

Tabel 1 Contoh Pelabelan Sentimen Analisis

No.	Teks	Analisis Sentimen
1	Dapat Target dari Presiden, Pemprov Jatim Mulai Menata Sistem Rujukan Rumah Sakit	Positif
2	Sebelum Jembatan Branjangan - Greges MACET. Imbas ada antrean di depo kontainer Greges.	Positif
3	'STIMUNO memanfaatkan Meniran sebagai imunomodulator untuk menjaga daya tahan tubuh agar berfungsi optimal dalam melindungi tubuh dari penyebab infeksi. Dapatkan STIMUNO melalui Official Store GoApotik di Shopee. <a href="https://bit.ly/ShopeeStimuno">https://bit.ly/ShopeeStimuno</a> #ADV	Negatif
4	Facebook Kenalkan Model Gelap untuk Aplikasi Seluler.	Positif
5	Purwodadi arah Surabaya MACET, ada evakuasi kecelakaan truk dan pickup yang terperosok ke sungai. Iptu Marti Kani Laka Lantas Polres Pasuruan mengatakan, kecelakaannya terjadi empat hari lalu di depan Mapolsek Purwodadi.	Positif
6	Jenis BBM, yang pertama masih non subsidi terlebih dahulu, tapi tidak kemungkinan Premium dan Solar juga akan bisa dilayani dengan non tunai.	Negatif

Pada proses ini data tweet yang akan digunakan adalah dataset dari kaggle.com. Pelabelan sentimen analisis pada penelitian ini menggunakan SVM dan ANN. SVM dan ANN dikombinasikan untuk menghasilkan output. ANN membutuhkan dataset training

yang besar untuk melatih ketepatan mesin. Untuk mengatasi hal ini ANN di hibridkan menggunakan SVM untuk pelatihan sehingga dari kombinasi ini mesin tidak perlu memproses lebih banyak data training. Namun, mesin yang menggunakan kombinasi SVM dan ANN ini akan menghasilkan keputusan yang tepat. Contoh pelabelan sentimen analisis tertera pada Tabel 1.

Tabel 2 Contoh Pelabelan *Entity Recognition*

No.	Teks	Entitas	Kategori
1	Kabar, baiknya pemerintah Aceh menyalurkan bantuan, pokok, utk, Masyarakat terdampak Covid - 19	<i>Political issue</i> , lokasi, dan Kesehatan.	O, O B-Pol B-Lok B-Pol O O O O B-Kes I-Kes O O
2	Wii . . . Yg berdampak Covid - 19 itu tdk hanya warga miskin . Berita ini judulnya benar .	Kesehatan dan ekonomi.	O O O O O B-Kes I-Kes O O O O O B-Eko I-Eko O O O O O
3	Tiga daerah di Papua Barat Zona hijau dari Covid - 19	Lokasi dan kesehatan.	O O O B-Lok I-Lok O O O B-Kes O O

Pada Tabel 2 menunjukkan bahwa data training dan data testing untuk pelabelan entitas dimana kategori B-Lok dan I-Lok, “B” menunjukkan begin (token awal) sedangkan “I” menunjukkan inside (token isi) begitu juga seterusnya untuk kategori yang mengandung “B” dan “I”. Pada kategori B-Lok dan I-Lok, “Lok” merupakan entitas lokasi sedangkan pada kategori B-Pol dan I-Pol, “Pol” menunjukkan entitas political issue. Kategori B-Kes dan I-Kes, “Kes” menunjukkan entitas kesehatan sedangkan B-Eko dan I-Eko, “Eko” merupakan entitas ekonomi. Kategori O dalam Tabel 2 merupakan kalimat yang tidak termasuk di dalam entitas.

## HASIL DAN PEMBAHASAN

Data *tweet* yang digunakan pada penelitian ini diperoleh dari kaggle.com yang memuat dataset *tweet* berbahasa Indonesia menggunakan keyword “Korona dan Pemerintah” atau “Covid dan Pemerintah” pada bulan Mei sampai Agustus yang berjumlah 36.923 data *tweet*. Data *tweet* memiliki karakter khusus yang dapat merusak hasil pemodelan topik seperti “(odp-rt)”, “(hm)”, “(odp-pr)”, dan sebagainya. Pada tahap ini dilakukan penghapusan karakter-karakter tersebut secara otomatis. Tahap pra-pemrosesan meliputi penghapusan karakter khusus, penghapusan tanda baca dan angka, penghapusan url, mention, dan hastag serta penghapusan stopwords bahasa Indonesia. Apabila ada data *tweet* yang kosong pada proses tersebut maka data itu akan dihapus. Sehingga terdapat pengurangan jumlah data dari data awal. Data kosong tersebut dapat mengakibatkan gagalnya proses pemodelan topik. Contoh data *tweet* berbahasa Indonesia dari kaggle.com yang bertemakan Covid-19. Hasil pra-pemrosesan dilakukan proses pengubahan teks

menjadi angka atau vektorisasi teks menggunakan *TFIDF*. Data yang dihasilkan berupa vektor dari setiap inputan data sehingga terdapat data vektor.

Hasil pra-pemrosesan dilakukan proses pengubahan teks menjadi angka atau vektorisasi teks menggunakan *TFIDF*. Data yang dihasilkan berupa vektor dari setiap inputan data sehingga terdapat data vektor. Pada proses ini data *tweet* yang akan digunakan adalah dataset dari *kaggle.com*. Pelabelan sentimen analisis pada penelitian ini menggunakan *SVM* dan *ANN*. *SVM* dan *ANN* dikombinasikan untuk menghasilkan output. *ANN* membutuhkan dataset training yang besar untuk melatih ketepatan mesin. Untuk mengatasi hal ini *ANN* di hibridkan menggunakan *SVM* untuk pelatihan sehingga dari kombinasi ini mesin tidak perlu memproses lebih banyak data training. Namun, mesin yang menggunakan kombinasi *SVM* dan *ANN* ini akan menghasilkan keputusan yang tepat. Hasil dari sentimen positif dan negatif berdasarkan beberapa parameter inputannya. Dari Tabel 3 didapatkan hasil dari pengkombinasian *ANN* dan *SVM* yang dapat membaca sentimen positif atau negatif sesuai dengan analisis secara manual. Akurasi klasifikasi *SVM* diperoleh 0.742 serta kombinasi *ANN* dan *SVM* diperoleh 0.769.

Tabel 3. Hasil Klasifikasi

No.	Metode Klasifikasi	Akurasi
1.	SVM	0.74
2.	ANN + SVM	0.77

Bayesian digunakan untuk mendapatkan *entity recognition*. Pelabelan *entity recognition* dilakukan dengan pembagian data menjadi data training dan data testing. Kemudian dilakukan uji kesesuaian keduanya. Sehingga akan dikelompokkan menjadi entitas ekonomi, political issue, lokasi, dan kesehatan. Poses selanjutnya pengelompokan dokumen yang memiliki label yang sama sesuai dengan Tabel 3.

Tabel 4. Kombinasi Entitas

Topik	Kombinasi Entitas				Sentimen
1	Kesehatan				Positif
2	Kesehatan	Lokasi			Positif
3	Kesehatan	Kebijakan Pemerintah			Positif
4	Kesehatan	Ekonomi			Positif
5	Kesehatan	Lokasi	Kebijakan Pemerintah		Positif
6	Kesehatan	Lokasi	Ekonomi		Positif
7	Kesehatan	Kebijakan Pemerintah	Ekonomi		Positif
8	Kesehatan	Lokasi	Political issue	Ekonomi	Positif
9	Kesehatan				Negatif
10	Kesehatan	Lokasi			Negatif
11	Kesehatan	Kebijakan Pemerintah			Negatif
12	Kesehatan	Ekonomi			Negatif
13	Kesehatan	Lokasi	Kebijakan Pemerintah		Negatif
14	Kesehatan	Lokasi	Ekonomi		Negatif
15	Kesehatan	Kebijakan Pemerintah	Ekonomi		Negatif
16	Kesehatan	Lokasi	Political issue	Ekonomi	Negatif

Tabel 5. Hasil Pemodelan Topik

Topik	Dokumen
1	dapat target dari presiden pemprov jatim mulai menata sistem rujukan rumah sakit. purwodadi arah surabaya macet ada evakuasi kecelakaan truk dan pickup yang terperosok ke sungai iptu marti kanit laka lantas polres pasuruan mengatakan kecelakaannya terjadi empat hari lalu di depan mapolsek purwodadi.
2	sebelum jembatan branjangan greges macet imbas ada antrean di depo kontainer greges odprt. facebook kealkan model gelap untuk aplikasi seluler. tempat wisata mulai dibuka tnipolri pastikan keamanan pengunjung. google meet tambah fitur baru mirip zoom.
3	pihak keluarga silakan menghubungi gkssfm. jenis bbm yang pertama masih non subsidi terlebih dahulu tapi tidak kemungkinan premium dan solar juga akan bisa dilayani dengan non tunai.
4	stimuno memanfaatkan meniran sebagai imunomodulator untuk menjaga daya tahan tubuh agar berfungsi optimal dalam melindungi tubuh dari penyebab infeksi dapatkan stimuno melalui official store goapotik di shopee.
5	tidak bawa masker dan ktp dihukum beri makanan ke odgj. hari ini forkopimda jatim bagibagi jutaan masker ke masyarakat. ylki kombinasi obat covid yang dirilis gugus tugas belum memenuhi syarat konsumsi.
6	tapi warga diminta tidak lengah tetap menerapkan pakai masker dan jaga jarak newsodppr. panglima tnikapolri pimpin rapat penanganan covid di jatim.
7	pasien covid di jatim sembuh total kasus capai . pemprov jatim dukung penelitian obat covid oleh tim peneliti unair. dokter sampaikan kiat aman pulang ke rumah di tengah pandemi covid.
10	selanjutnya semua jalur akan pakai non tunai target di minggu terakhir bulan juli. sekdaprov jatim tegaskan data covid di jatim valid.
11	presiden melantik pimpinan pusat dan dewan pertimbangan legiun veteran republik indonesia. masih bekerjasama dengan halodoc gojek juga menyediakan layanan rapid test atau pemeriksaan pcr untuk para mitranya
12	gate tol waru perak yang sebelumnya ditutup karena ada sterilisasi jalur sudah dibuka lalu lintas mulai bergerak odprt. risma minta kepala sekolah rumuskan protokol kesehatan di sekolah
13	covid hunter ala pemprov jatim mulai terjun ke daerah untuk tes massal. gugus tugas covid apresiasi langkah pemprov jatim dan pemkot surabaya bangun rs darurat
14	pihaknya melibatkan babinsa babinkamtibmas dan satpol pp untuk keliling kampung dan pasar tradisional untuk membagikan masker
15	pasmr dan pemkab sidoarjo kerjasama percepatan penanganan pandemi covid

Sehingga pemodelan topik berjumlah 16 topik dan menghapus sebagian dokumen yang tidak memiliki kemiripan dengan total jumlah dokumen 36.923 yang dapat dilihat pada Tabel 5. Pemodelan topik juga menggunakan metode TFIDF dan LSI sebagai pembanding. Teknik peringkasan otomatis dilakukan uji coba dengan 3 metode seperti : *TextRank*, *Okapi BM25*, dan *Weight Pagerank*. Metode peringkasan yang dilakukan secara otomatis menggunakan *TextRank*, *Okapi BM25*, dan *PageRank* menghasilkan nilai yang cukup berbeda disetiap topik data.

Metode pemodelan *Opinion Mining* dan *K-Means* menghasilkan nilai yang lebih baik daripada metode pemodelan *TFIDF* dan *LSI* untuk metode peringkasan menggunakan *TextRank*, *PageRank*, dan *Okapi BM25* hal ini digambarkan seperti Tabel 5. Nilai pada topik 1 sampai 8 untuk pemodelan *Opinion Mining* lebih tinggi, tapi ada beberapa topik yang tidak bisa ter summary.

Tabel 6. Hasil Perbandingan Peringkasan Otomatis

Topik	PageRank		Okapi BM25		TextRank	
	Opinion mining	TFIDF + LSI	Opinion mining	TFIDF + LSI	Opinion Mining	TFIDF + LSI
1	0,53	-	0,95	0,051	0,43	0,036
2	0,18	0,011	0,97	0,037	0,37	0,02
3	0,51	0,024	0,97	0,031	0,42	0,038
4	0,28	0,032	0,93	0,053	0,55	0,017

5	0,49	0,054	0,93	0,017	0,17	0,051
6	0,32	0,013	0,88	0,049	0,73	0,059
7	0,49	0,02	0,924	0,039	0,41	0,028
8	0,39	0,022	0,84	0,021	0,22	0,017
9	-	0,03	0,86	0,019	-	0,016
10	-	0,031	0,93	-	-	-
11	-	0,028	0,87	-	-	-
12	-	0,029	0,63	0,006	-	0,02
13	-	0,018	-	-	-	0,014
14	-	0,035	-	0,032	-	0,022
15	-	0,02	-	-	-	0,012
16	-	-	-	-	-	-

## KESIMPULAN

Data *tweet* yang digunakan pada penelitian ini diperoleh dari kaggle.com yang memuat dataset *tweet* berbahasa Indonesia menggunakan kata kunci “Korona dan Pemerintah” atau “Covid dan Pemerintah” pada bulan Mei sampai Agustus yang berjumlah 36.923 data *tweet*. Dalam proses klasifikasi untuk menentukan sentiment terhadap setiap dokumen menggunakan metode *SVM* menggunakan *ANN* yang memperoleh akurasi sebesar 0.76 lebih baik dibandingkan hanya menggunakan *SVM* saja dengan jumlah akurasi 0.74. Dokumen yang telah memiliki nilai sentiment, selanjutnya diberikan *entity recognition* berupa entitas ekonomi, *political issue*, lokasi, dan kesehatan. Dokumen yang tidak memiliki salah satu dari 4 entitas tersebut maka tidak akan diproses dalam tahap pemodelan topik. Dimana tahap pemodelan topik ini menggunakan metode *Opinion Mining* dan *clustering K-Means* dengan total 16 topik. Selain menggunakan metode *Opinion Mining* dan *K-Means* pemodelan juga menggunakan metode *TFIDF* dan *LSI*. Data 16 topik hasil dari pemodelan topik akan digunakan untuk proses peringkasan dokumen yang menggunakan 3 metode yaitu *TextRank*, *Okapi BM25*, dan *Weight Pagerank*. Pada tahap peringkasan ini menggunakan evaluasi *Rogue 1*, *Rogue 2*, *Rogue L* dan *Rogue Mean*. Sehingga diperoleh nilai rata-rata peringkasan mean *TextRank* pada pemodelan *Opinion Mining* dan *K-means* lebih tinggi yaitu 0,41 daripada pemodelan *TFIDF* dan *LSI* yaitu 0,027. Nilai rata-rata peringkasan mean *Okapi BM25* pada pemodelan *Opinion Mining* dan *K-means* lebih tinggi yaitu 0,89 daripada pemodelan *TFIDF* dan *LSI* yaitu 0,029. Nilai rata-rata peringkasan mean *TextRank* pada pemodelan *Opinion Mining* dan *K-means* lebih tinggi yaitu 0,39 daripada pemodelan *TFIDF* dan *LSI* yaitu 0,026. Dari ketiga metode peringkasan pemodelan topik menggunakan *Opinion Mining* dan *K-Means* menghasilkan nilai yang lebih tinggi daripada pemodelan topik menggunakan *TFIDF* dan *LSI*. Tapi pemodelan *Opinion Mining* dan *K-Means* hanya dapat teringkaskan pada beberapa topik saja.

## SARAN

Penelitian ini merupakan pengembangan metode untuk mengoptimalkan *Opinion Mining* pada peringkasan otomatis sehingga masih banyak kekurangan di dalamnya. Diharapkan pada penelitian selanjutnya *Opinion Mining* dapat dimaksimalkan seperti penentuan *Sentiment Analysis* dan *Entity Recognition* lebih baik lagi terutama dalam penentuan entitas di dalam penelitian ini hanya ada 4 entitas sehingga peringkasan otomatis dapat digunakan untuk membantu dalam menyimpulkan *text tweet* yang setiap kalimatnya bukan kalimat baku.

**REFERENSI**

- [1] M. Mohd, R. Jan, and M. Shah, "Text document summarization using word embedding," *Expert Syst. Appl.*, vol. 143, Apr. 2020, doi: 10.1016/j.eswa.2019.112958.
- [2] J. Rashid, S. M. A. Shah, and A. Irtaza, "Fuzzy topic modeling approach for text mining over short text," *Inf. Process. Manag.*, vol. 56, no. 6, Nov. 2019, doi: 10.1016/j.ipm.2019.102060.
- [3] M. Asghari, D. Sierra-Sosa, and A. S. Elmaghraby, "A topic modeling framework for spatio-temporal information management," *Inf. Process. Manag.*, vol. 57, no. 6, Nov. 2020, doi: 10.1016/j.ipm.2020.102340.
- [4] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny)*, vol. 477, pp. 15–29, Mar. 2019, doi: 10.1016/j.ins.2018.10.006.
- [5] P. Damayanti, D. Purwitasari, and N. Suciati, "Eliminasi Non-Topic Menggunakan Pemodelan Topik untuk Peringkasan Otomatis Data Tweet dengan Konteks Covid-19," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 199, 2021, doi: 10.25126/jtiik.0814324.
- [6] X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, "Extractive summarization using supervised and unsupervised learning," *Expert Syst. Appl.*, vol. 133, pp. 173–181, Nov. 2019, doi: 10.1016/j.eswa.2019.05.011.
- [7] Y. Du, Q. Li, L. Wang, and Y. He, "Biomedical-domain pre-trained language model for extractive summarization," *Knowledge-Based Syst.*, vol. 199, Jul. 2020, doi: 10.1016/j.knosys.2020.105964.
- [8] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165. Elsevier Ltd, Mar. 01, 2021. doi: 10.1016/j.eswa.2020.113679.
- [9] G. Chandra and S. K. Dwivedi, "Query expansion based on term selection for Hindi – English cross lingual IR," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 3, pp. 310–319, Mar. 2020, doi: 10.1016/j.jksuci.2017.09.002.
- [10] A. Sharma and U. Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India," in *Procedia Computer Science*, 2020, vol. 173, pp. 325–334. doi: 10.1016/j.procs.2020.06.038.
- [11] J. Serrano-Guerrero, F. P. Romero, and J. A. Olivias, "Fuzzy logic applied to opinion mining: A review," *Knowledge-Based Syst.*, vol. 222, Jun. 2021, doi: 10.1016/j.knosys.2021.107018.
- [12] S. Zervoudakis, E. Marakakis, H. Kondylakis, and S. Goumas, "OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data," *Mach. Learn. with Appl.*, vol. 3, p. 100018, Mar. 2021, doi: 10.1016/j.mlwa.2020.100018.