

## Analisis Media Sosial Twitter untuk Mengetahui Pengguna Berpengaruh pada Portal Berita di Indonesia dengan Metode TSIM (Topic-based Social Influence Measurement)

Tabita Krisma Kurniasari<sup>1</sup>, Warih Maharani<sup>2</sup>, Jati Hiliamsyah Husen<sup>3</sup>

<sup>1,2,3</sup>Fakultas Informatika, Universitas Telkom, Bandung

<sup>1</sup>tebkris@students.telkomuniversity.ac.id, <sup>2</sup>wmaharani@telkomuniversity.ac.id,

<sup>3</sup>jatihusen@telkomuniversity.ac.id

---

### Abstrak

Media sosial merupakan sarana untuk melakukan interaksi dalam jaringan internet dan salah satu wadah untuk membentuk jaringan sosial. Salah satu contoh media sosial yaitu Twitter. Pengguna Twitter dapat membangun sebuah relasi antar pengguna lain dan dapat dijadikan sebagai sarana penyebaran suatu informasi dengan berbagai topik, seperti penyebaran berita, kampanye, pemasaran produk, dan penyebaran informasi lainnya. Namun, secara umum, banyak yang mengakui bahwa pengguna yang memiliki pengikut dalam jumlah besar adalah pengguna yang berpengaruh (*influencer*). Jadi, penelitian ini menggunakan metode Topic-based Social Influence Measurement (TSIM) yang bertujuan untuk merangking pengguna Twitter berbasis topik dengan bantuan metode Latent Dirichlet Allocation (LDA) dalam membantuproses pembagian topik. *Wordcloud* membantu dalam membentuk gambaran visual frekuensi kemunculan kata dan membantu dalam pemberian nama setiap topik. Pada proses Latent Dirichlet Allocation (LDA), untuk jumlah pembagian topik terbaik berdasarkan nilai *coherence* ditunjukkan pada jumlah pembagian 10 topik dengan nilai *coherency* 0.3711497. Penelitian ini menggunakan 15 pengguna Portal Berita di Indonesia berdasarkan *top site* di Alexa. Perhitungan metode TSIM dilakukan dengan memanfaatkan empat indikator, yaitu *followers scale*, *topic activity*, *topic-based attractiveness*, dan *network centrality*. Setelah perhitungan empat indikator, dilakukan agregasi untuk menunjukkan rangking setiap pengguna berdasarkan topik. Hasil peringkat dengan metode TSIM menunjukkan bahwa pengguna detikcom dan CNNIndonesia merupakan pengguna yang berpengaruh pada hampir setiap topik. Perangkingan dengan menggunakan TSIM tersebut diuji performansi prediksi kesalahannya menggunakan metode MAE dan telah memberikan hasil performansi yang baik, karena prediksi kesalahan rangking setiap topik memberikan nilai MAE yang mendakati 0. Nilai MAE terkecil ditunjukkan pada Topik Hukum, yaitu sebesar 0.016541.

Kata kunci : TSIM, Twitter, LDA, Worldcloud, coherence

---

### Abstract

Social media is a means to interact in the internet and one of the platforms to form social networks. One example of social media is Twitter. Twitter users can build relationships between other users and can be used as a means of disseminating information on various topics, such as news dissemination, campaigns, product marketing, and other information dissemination. However, in general, many recognize that users who have a large number of followers are influential influencers. So, this research uses the Topic-based Social Influence Measurement (TSIM) method which aims to rank topic-based Twitter users with the help of the Latent Dirichlet Allocation (LDA) method to help the process of topic sharing. Wordcloud helps in shaping visual images of the frequency with which words appear and helps in naming each topic. In the Latent Dirichlet Allocation (LDA) process, the best number of topics divided based on the coherency value is shown in the number of 10 topic divisions with a coherency value of 0.3711497. This study uses 15 News Portal users in Indonesia based on the top site on Alexa. The TSIM method calculation is done by utilizing four indicators, namely followers scale, topic activity, topic-based attractiveness, and network centrality. After calculating the four indicators, an aggregation is performed to show the ranking of each user by topic. The ranking results using the TSIM method show that users of detikcom and CNNIndonesia are influential users on almost every topic. Ranking by using the TSIM tested its error prediction performance using the MAE method and has provided good performance results, because the prediction error ranking of each topic gives a value of the approaching MAE 0. The smallest MAE value is shown in the Legal Topic, which is 0.016541.

Keywords: TSIM, Twitter, LDA, Worldcloud, coherence

---

## 1. Pendahuluan

### Latar Belakang

Pada era digital saat ini perkembangan teknologi informasi dan komunikasi bertumbuh semakin pesat dan memberikan kemudahan dalam mengakses suatu informasi terkini melalui media massa. Media massa merupakan alat komunikasi yang dapat menjangkau masyarakat luas [7], salah satunya adalah berita. Berita adalah sebuah informasi tentang kejadian yang baru, penting dan bermakna (signifikan), yang berpengaruh secara relevan terhadap para pendengarnya sehingga layak dinikmati oleh mereka [4]. Oleh karena itu, berita dapat membantu dalam proses penyebaran informasi.

Penyebaran informasi berita itu sendiri merupakan dampak dari suatu pemanfaatan teknologi yang bisa membentuk sebuah jaringan sosial melalui media sosial. Salah satu contoh media sosial yaitu Twitter, dimana pengguna bisa memanfaatkan media sosial ini sebagai wadah untuk mendapatkan informasi atau berita dan dapat dijadikan sebagai sarana penyebaran suatu berita. Maka dari itu, pengguna Twitter yang memerlukan suatu informasi atau berita, perlu mengetahui pengguna portal berita mana yang paling berpengaruh terhadap beberapa macam kategori berita, karena hal itu menjadi sebuah nilai jika berita itu berhasil tersampaikan ke pembaca [8]. Selain itu, sebuah portal berita memberikan informasi kepada pengikutnya yang aktual (up to date) [12]. Dengan adanya latar belakang tersebut, penulis melakukan penelitian dengan beberapa metode yang mendukung proses penelitian. Pada setiap metode penelitian yang dilakukan tentunya memiliki kekurangan dan kelebihan baik dari segi biaya, waktu proses penelitian, keakuratan hasil akhir penelitian dan beberapa faktor yang mempengaruhi suatu penelitian.

Pada penelitian sebelumnya, digunakan metode TwitterRank yang bertujuan untuk mengukur dan mengklasifikasi pengaruh pengguna berdasarkan suatu topik yang menggunakan suatu hashtag. Namun, 80% pengguna Twitter membagikan tweet tanpa menggunakan sebuah hashtag yang merujuk pada topik tertentu. Hal ini dapat menyedatkan hasil pengaruh pengguna berdasarkan suatu topik, misalnya 4 dari 5 tweet di penelitian sebelumnya tidak bisa mengukur seberapa besar pengaruhnya dengan menggunakan metode tersebut. Selain itu, TwitterRank hanya menggunakan 3 indikator untuk mengukur pengaruhnya [6]. Maka dari itu, penelitian ini menggunakan metode TSIM untuk mengukur ranking pengaruh pengguna twitter berdasarkan topik, karena metode TSIM tidak menggunakan hashtag untuk proses pencarian topik.

Topic-based Social Influence Measurement (TSIM) menggunakan cara pengukuran berdasarkan empat indikator yaitu followers scale, topic activity, topic-based attractiveness, dan network centrality. Setelah empat indikator tersebut memberikan hasil, selanjutnya akan dilakukan proses aggregating influence score yang memberikan nilai akhir untuk menentukan pengguna yang berpengaruh pada setiap topik. Arti kata berpengaruh adalah kemampuan pengguna dalam menyampaikan suatu informasi secara luas pada suatu jaringan. Oleh karena itu, pada penelitian ini memanfaatkan topic modelling Latent Dirichlet Allocation (LDA) untuk mengidentifikasi topik. Selain itu, metode TSIM juga dapat diterapkan dalam penyebaran informasi dengan biaya yang lebih murah dan efisien [1]. Pada penelitian ini penulis langsung menentukan 15 pengguna Portal Berita di Indonesia berdasarkan top sites Alexa dengan total scraping 11.760 tweet.

### Topik dan Batasannya

Berdasarkan latar belakang dari penelitian diatas, maka yang menjadi rumusan masalah pada penelitian tugas akhir ini adalah cara mengimplementasikan metode Topic-based Social Influence Measurement (TSIM) dengan bantuan pemodelan Latent Dirichlet Allocation (LDA) dalam meranking 15 pengguna media sosial Twitter Portal Berita di Indonesia. Serta, mengevaluasi uji performansi metode TSIM dan mengevaluasi pembagian topik terbaik di metode LDA pada penelitian ini.

Pada proses pengerjaan tugas akhir ini memiliki beberapa batasan masalah yaitu data diambil dari Twitter sejumlah 11.761 tweets dan dataset yang terambil dari tanggal 6 Mei 2020 hingga 13 Mei 2020, karena banyaknya konflik yang terjadi pada tanggal tersebut, mulai dari masalah ekonomi, politik, hingga bencana alam, dengan memilih 15 pengguna Portal Berita di Indonesia menggunakan twitterscraper. Dari 15 pengguna tersebut dipilih berdasarkan top site dari website Alexa.

### Tujuan

Tujuan dari penelitian ini adalah mengimplementasikan metode Topic-based Social Influence Measurement (TSIM) dengan bantuan pemodelan Latent Dirichlet Allocation (LDA) dalam meranking setiap pengguna pada suatu topik dan nilai uji performansi metode TSIM serta evaluasi pembagian topik terbaik di metode LDA pada penelitian pengguna Twitter Portal Berita di Indonesia

## Organisasi Tulisan

Berikut ini adalah beberapa bab yang akan dilakukan pada penelitian Tugas akhir ini:

1. Pendahuluan terdiri dari latar belakang, topik dan batasannya, dan tujuan.
2. Studi terkait bagian ini menjelaskan teori/studi literatur yang mendukung pengerjaan Tugas Akhir.
3. Sistem yang dibangun bagian ini akan mendeskripsikan rancangan dan sistem yang akan dibangun pada pengerjaan Tugas Akhir.
4. Evaluasi akan menjelaskan hasil dari penelitian Tugas Akhir ini.
5. Kesimpulan memuat tentang kesimpulan dan saran dari analisis dan hasil pengujian Tugas Akhir ini.

## 2. Studi Terkait

### 2.1 Penelitian Terkait

Penelitian terdahulu merupakan salah satu referensi penulis untuk membantu dalam proses penelitian ini, baik dari jenis penelitian atau teori yang digunakan dan metode yang dilakukan pada penelitian terdahulu, agar dapat membandingkan dengan penelitian yang dilakukan oleh penulis.

Pada penelitian *Asso Hamzehei et al.* tersebut crawling data dengan mentargetkan "Machine Learning" [1] dengan media sosial Twitter, sedangkan pada penelitian TA ini memanfaatkan 15 pengguna Portal Berita Indonesia di media sosial Twitter. Pada penelitian lain yaitu penelitian *Varun. R.E, Indrajit Bhattacharya, dan Vinayaka Pandit* juga memanfaatkan Topic-based Social Influence untuk menentukan pengguna yang berpengaruh dengan melakukan 5 indikator yaitu *follower strength, activity, engagement, authority or network centrality, dan Response rate or timelines* [5], sedangkan dengan penelitian yang dilakukan pada TA ini hanya memanfaatkan 4 indikator sesuai dengan metode TSIM yaitu *follower scale, topic activity, topic-based attractiveness, dan network centrality*.

### 2.2 Topic Modelling

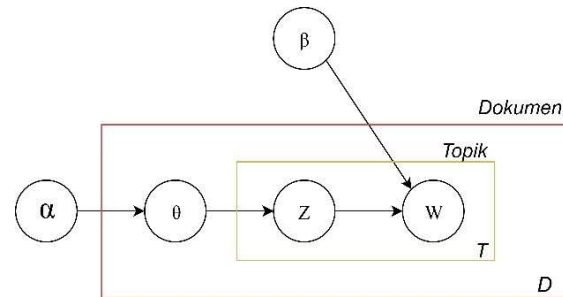
*Topic Modelling* atau pemodelan topik secara umum adalah proses pengelompokan data teks berdasarkan suatu topik tertentu. Pemodelan topik menurut Bleu terdiri dari beberapa entitas yaitu "kata", "dokumen", dan "corpora". "Kata" adalah unit dasar dari data diskrit dalam suatu dokumen, yang didefinisikan sebagai item dari kosa kata yang diberi indeks untuk setiap kata unik pada dokumen. "Dokumen" merupakan susunan dari N kata. Corpus merupakan kumpulan M dokumen, sedangkan corpora merupakan bentuk jamak dari corpus. Sementara "topik" adalah distribusi dari beberapa kosakata yang tetap. Secara singkatnya, setiap dokumen dalam corpus mengandung proposisi tersendiri dari topik-topik yang dibahas sesuai kata-kata di dalamnya [3].

Sebuah topik terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing adalah dasar dari pemodelan topik. Dengan keterbatasan manusia dalam pemahaman suatu topik, distribusi topik per-dokumen, dan penggolongan setiap kata, maka adanya pemodelan topik ini bertujuan untuk menemukan topik dan kata-kata yang terdapat pada topik tersebut [2].

### 2.3 Latent Dirichlet Allocation

*Latent Dirichlet Allocation* merupakan metode yang membantu dalam proses pengelompokan atau *clustering*. LDA bermanfaat jika digunakan dalam meringkas, melakukan klusterisasi, menghubungkan atau memproses data, karena LDA menghasilkan daftar topik yang diberi bobot pada setiap dokumen [2]. Adapun metode distribusi yang digunakan untuk mendapatkan distribusi topik per-dokumen disebut distribusi Dirichlet, lalu dari hasil Dirichlet akan digunakan untuk mengalokasikan kata-kata pada dokumen untuk topik yang berbeda. Dalam LDA, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen, penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi yang tidak dapat diamati secara manual oleh manusia, oleh sebab itu algoritma ini dinamakan *Latent Dirichlet Allocation* [2].

Terdapat tiga tingkatan pada LDA Modeling. Pada  $D = [d_1, d_2, d_3, \dots, d_q]$  dokumen terdapat parameter  $\alpha$  dan  $\beta$  yang merupakan parameter distribusi topik ( $T = [t_1, t_2, t_3, \dots, t_j]$ ) pada tingkatan corpus. Parameter  $\alpha$  digunakan untuk menentukan distribusi topik dalam dokumen, semakin banyak topik yang dibahas di suatu dokumen, maka semakin tinggi nilai alpha dalam suatu dokumen. Parameter  $\beta$  berfungsi untuk menentukan distribusi di dalam topik, semakin sedikit kata-kata yang terdapat dalam topik atau mengandung kata-kata yang lebih spesifik, maka



**Gambar 1.** Visualisasi Pemodelan Topic dengan metode LDA [3]

nilai beta semakin kecil, begitu sebaliknya. Variabel  $\vartheta_D$  adalah variabel yang berada di tingkat dokumen ( $D$ ). Variabel  $\vartheta$  merepresentasikan distribusi topik untuk dokumen tertentu. Semakin tinggi nilai  $\vartheta$ , maka semakin banyak topik yang ada di dalam dokumen, sedangkan semakin kecil nilai  $\vartheta$ , maka dapat dikatakan dokumen tersebut semakin spesifik pada topik tertentu. Variabel  $Z$  dan  $W_m$  adalah variabel tingkat kata. Variabel  $Z$  merepresentasikan topik dari kata tertentu pada sebuah dokumen sedangkan variabel  $W$  merepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen [3].

LDA dapat digunakan di berbagai bidang, diantaranya adalah untuk menganalisis trend pada media sosial [10] dan terbukti dapat bekerja dengan baik untuk menganalisis dokumen panjang seperti Wikipedia atau dokumen pendek seperti tweet [16].

LDA pada penelitian ini hanya untuk membantu pengelompokan setiap tweet, agar tweet tersebut dikelompokkan sesuai dengan topik yang dibentuk.

## 2.4 Topic-based Social Influence Measurement

*Topic-based Social Influence Measurement* adalah metode berbasis topik yang bertujuan untuk mengukur pengaruh pengguna terhadap suatu topik yang spesifik. Metode ini dilakukan dengan mengidentifikasi topik berdasarkan konten yang dibuat pengguna, dan menghitung pengaruh pengguna untuk topik yang terdeteksi. Metode ini dihitung berdasarkan empat indikator. Indikator-indikator yang dimaksud adalah *Followers Scale*, *Topic Activity*, *Topic-Based Attractiveness*, dan *Network Centrality*.

### a. Followers Scale

*Followers Scale* ( $F_f$ ) adalah perhitungan jumlah *followers* pengguna. *Followers* dapat menjadi salah satu indikator untuk pengaruh pengguna terhadap suatu topik yang spesifik, walaupun jumlah *followers* yang besar tidak selalu memberikan pengaruh yang tinggi terhadap penyebaran suatu topik.

### b. Topic Activity

*Topic Activity* ( $F_a$ ) adalah perhitungan keterkaitan aktivitas pengguna terhadap suatu topik.  $F_a$  dihitung berdasarkan jumlah tweets pengguna  $u_i$  terkait topik  $t_j$ .

$$F_a = \sum_{d_i \in D_t} \delta(d_i) \quad (1)$$

$\delta(d_i)$  bernilai 1 jika tweets ke- $i$  ( $d_i$ ) dari pengguna  $u_i$  termasuk kedalam topik  $t_j$ .  $\delta(d_i)$  bernilai 0 jika tweets ke- $i$  ( $d_i$ ) dari pengguna  $u_i$  tidak termasuk kedalam topik  $t_j$ . Dengan kata lain,  $F_a$  adalah berapa banyak tweets pengguna  $u_i$  yang termasuk kedalam topik  $t_j$ .

### c. Topic-Based Attractiveness

*Topic-Based Attractiveness* ( $F_e$ ) adalah perhitungan seberapa tertariknya pengguna lain terhadap tweets yang dibuat oleh pengguna  $u_i$ .  $F_e$  dihitung berdasarkan jumlah *retweet* ( $R_i$ ) dan *mention* ( $M_i$ ) pengguna lain terhadap tweets ke- $i$  ( $d_i$ ) milik pengguna  $u_i$  mengenai topik  $t_j$ .

$$F_e = \sum_{d_i \in D_t} ((\delta(R_i) + \delta(M_i))) \quad (2)$$

### d. Network Centrality

*Network Centrality* ( $F_c$ ) adalah perhitungan pengaruh pengguna terhadap suatu topik dalam suatu graf.  $F_c$  dihitung menggunakan PageRank yang telah dijelaskan sebelumnya.  $G(D_t)$  adalah graf yang sesuai dengan

pengguna  $u_i$  pada dokumen  $D_t$  tentang topik  $t$ .  $PR(u_i, G(D_t))$  adalah nilai PageRank yang diperoleh oleh pengguna  $u_i$  dalam graf  $G(D_t)$ .

$$F_c = PR(u_i, G(D_t)) \tag{3}$$

PageRank adalah algoritma yang biasa digunakan untuk melakukan pemeringkatan suatu halaman web [11]. Pada suatu *graph* memiliki beberapa node.  $In(N_i)$  adalah seluruh node yang mengarah ke  $N_i$ , sedangkan  $Out(N_i)$  adalah node yang mengarah dari  $N_i$ . Sehingga,  $|In(N_i)|$  berada didalam derajat (*in-degree*) node  $N$  dan  $|Out(N_i)|$  berada diluar derajat (*out-degree*) node  $N$ .

Untuk mendapatkan nilai PageRank, ditentukan rumus sebagai berikut [14].

$$S(N_i) = (1 - d) + d * \sum_{j \in In(N_i)} \frac{1}{|Out(N_j)|} S(N_j) \tag{4}$$

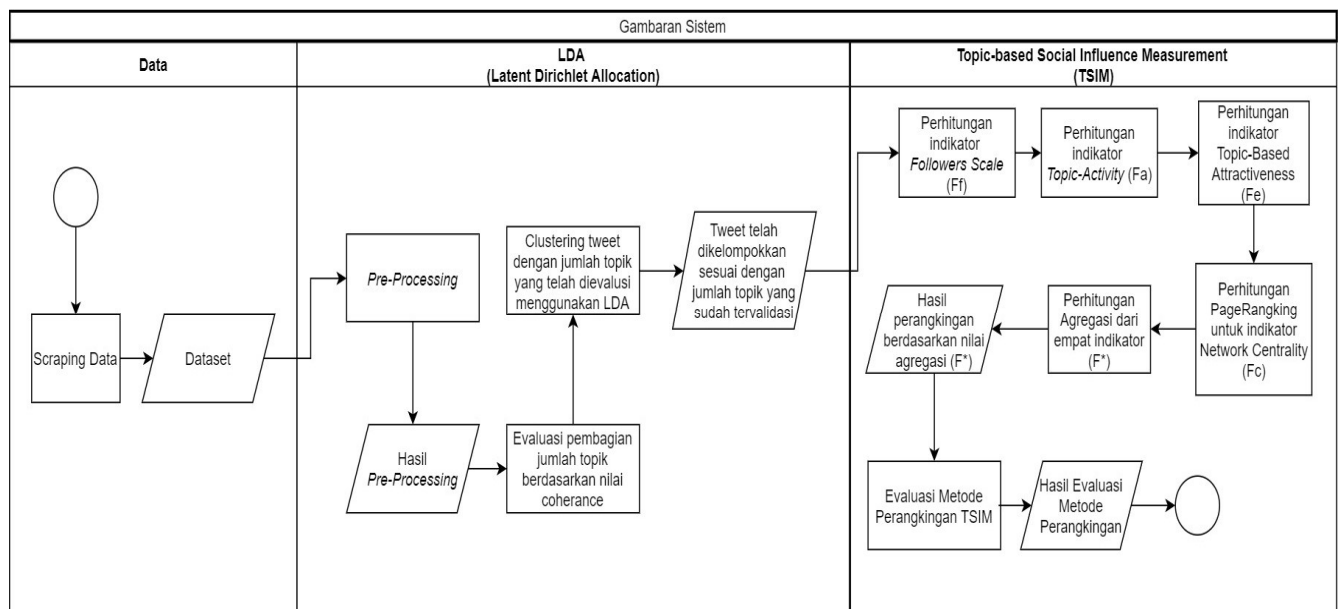
Dengan  $d$  adalah *damping factor* yang merupakan bilangan antara 0 sampai 1, dan biasanya ditetapkan menjadi 0,85 [13]. Metode peringkat ini dilakukan berulang kali, dimana nilai inisiasi untuk setiap node ditetapkan, yaitu 1. Setelah node tersebut diurutkan berdasarkan nilai PageRank dari persamaan diatas, maka pengguna pada peringkat pertama disebut sebagai pembuat topik.

Untuk menentukan peringkat pengguna yang paling berpengaruh terhadap suatu topik, dilakukan penghitungan rata-rata terhadap keempat indikator tersebut ( $F_f, F_a, F_e, dan F_c$ ) yang selanjutnya akan disebut dengan  $F^*$ .

### 3. Sistem yang Dibangun

#### 3.1 Gambaran Umum Sistem

Gambaran umum sistem akan menjelaskan proses penelitian ini secara umum. Proses penelitian ini akan menghasilkan peringkat pengguna yang paling berpengaruh berdasarkan suatu topik. Dapat dilihat pada gambar 2



Gambar 2. Gambaran Umum Sistem

#### 3.2 Tahapan Sistem

##### 3.2.1 Scraping Data Twitter

Pengguna dipilih dari media sosial Twitter, dimana proses pemilihan pengguna ini dilakukan dengan membuka sebuah pengguna twitter portal berita di Indonesia dan pada penelitian ini dipilih 15 pengguna Portal Berita Indo-

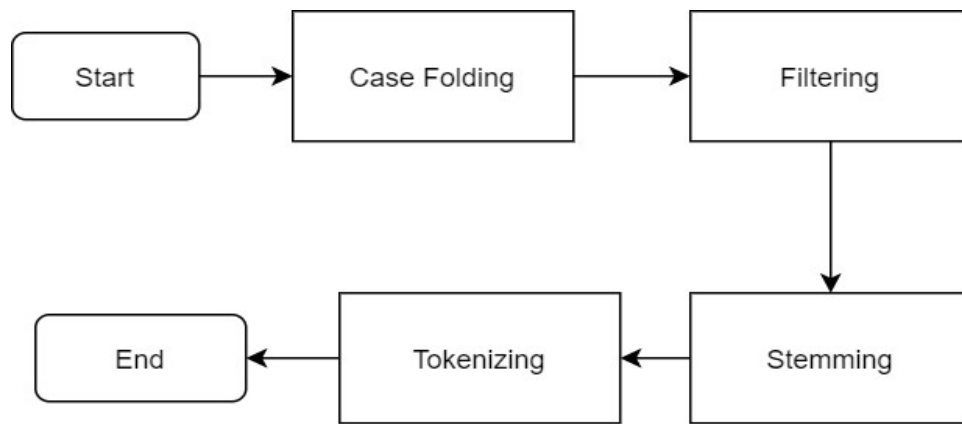
nesia yang sudah terverifikasi (*verified account*) untuk diteliti. pengguna Twitter dipilih berdasarkan 40 *Top Site* di Indonesia pada link <https://www.alexacom/topsites/countries/ID>.

1. @okezonenews
2. @TribunNews
3. @detikcom .
4. @kompascom
5. @SINDOnews
6. @liputan6dotcom
7. @merdekadotcom
8. @suaradotcom
9. @kumparan
10. @idntimes
11. @KapanLagicom
12. @CNNIndonesia
13. @kompasiana
14. @pikiran rakyat
15. @tempodotco

Alexa memiliki cara perhitungan sendiri untuk menentukan peringkat *site* yaitu perhitungan dengan kombinasi antara *daily visitor* (pengunjung *site* tersebut) dengan *page view* (seberapa banyak *site* tersebut dilihat. Tabel 1 akan menunjukkan contoh 20 sampel *site* teratas berdasarkan Alexa.

**Tabel 1. Sampel 20 *site* teratas berdasarkan Alexa**

Site	Daily time on site	Daily page view per visitor	% of traffic from search	Total Sites Linking In
Okezone.com	04:20	4.36	7.80%	17,007
Google.com	14:09	15.39	0.40%	1,982,688
Tribunnews.com	03:53	1.98	62.00%	29,454
Youtube.com	14:02	7.87	15.80%	1,523,322
Grid.id	04:04	1.92	55.60%	8,535
Detik.com	06:29	4.02	29.80%	45,073
Kompas.com	04:26	2.11	45.70%	46,900
Sindonews.com	03:32	2.25	41.50%	5,064
Liputan6.com	04:44	2.08	57.40%	21,312
Tokopedia.com	12:18	7.18	23.20%	54,658
Bukalapak.com	07:18	4.61	30.80%	26,421
Brilio.net	02:33	1.42	61.70%	2,681
Merdeka.com	03:12	1.62	53.30%	11,154
Suara.com	03:30	2.14	45.40%	4,707
Kumparan.com	02:54	1.98	50.30%	5,826
Idntimes.com	05:10	4.02	56.90%	5,051
Google.co.id	05:34	6.1	19.20%	29,608
Yahoo.com	04:40	4.4	7.80%	453,735
Kapanlagi.com	03:22	2.56	59.50%	11,007
Blogspot.com	03:52	2.9	35.50%	14,400



**Gambar 3. Diagram proses *Pre-Processing***

### 3.2.2 *Pre-Processing*

Proses *Pre-processing* merupakan proses yang membantu dalam pemodelan LDA. Pada penelitian ini dilakukan beberapa tahap yang diperlukan diantaranya *case folding*, *tokenizing*, *filtering*, dan *stemming*, bisa dilihat pada gambar 3

1. *Case Folding* merupakan proses penyamaan dimana tweet yang bercampur dengan huruf besar akan dirubah keseluruhan menjadi huruf kecil. Proses ini hanya menerima huruf 'a' hingga 'z'. Menghilangkan karakter seperti tanda baca dan angka dan dianggap *delimiter* [17]. Pada tabel 2 menunjukkan penggunaan *case folding* pada penelitian ini.

**Tabel 2. Contoh proses *Case Folding***

Tweet	<i>Case Folding</i>
Pertahanan pertama kita. Karena kita tidak boleh karantina wilayah, jadi kita karantina wajah, kata Anies.Â â	pertahanan pertama kita. karena kita tidak boleh karantina wilayah, jadi kita karantina wajah, kata anies.
Maskapai pelat merah PT Garuda Indonesia (Persero) Tbk kembali melayani operasional penerbangan pada Kamis, 7 Mei 2020 pukul 00.01 Â â	maskapai pelat merah pt garuda indonesia persero tbk kembali melayani operasional penerbangan pada kamis mei pukul

2. *Filtering* merupakan tahap pengambilan kata yang penting. Pada proses filtering, penelitian ini memanfaatkan kamus Sastrawi. Kata yang tidak sesuai dengan kamus Sastrawi akan dihapuskan (membuang kata-kata yang kurang penting). Berikut adalah contoh dari proses *Filtering*. Contoh proses Filtering bisa dilihat di tabel 3

**Tabel 3. Contoh proses *Filtering***

Tweet	<i>Filtering</i>
pertahanan pertama kita. karena kita tidak boleh karantina wilayah, jadi kita karantina wajah, kata anies.	pertahanan pertama tidak karantina wilayah jadi karantina wajah kata anies
maskapai pelat merah pt garuda indonesia persero tbk kembali melayani operasional penerbangan pada kamis mei pukul	maskapai pelat merah pt garuda indonesia persero tbk melayani operasional penerbangan kamis mei pukul

3. *Stemming* merupakan proses perubahan kata ke bentuk dasar dari setiap kata hasil filtering. Menghilangkan kata di-, ke-, me-. Dengan memanfaatkan *Stopword* dapat menghilangkan kata-kata yang tidak diskriptif dengan pendekatan *bag-of-words*. Pada tabel A menunjukkan hasil dari proses stemming.

Tabel 4. Contoh proses *Stemming*

Tweet	Stemming
pertahanan pertama tidak karantina wilayah jadi karantina wajah kata anies	tahan pertama karantina wilayah jadi karantina wajah kata anies
maskapai pelat merah pt garuda indonesia persero tbk kembali melayani operasional penerbangan pada kamis mei pukul	maskapai pelat merah pt garuda indonesia persero tbk layan operasional terbang kamis mei pukul

4. *Tokenizing* proses pemotongan *string* dalam suatu kalimat menjadi kata per kata. Pada proses ini juga dapat menampilkan berapa kali kata tersebut muncul dalam suatu dokumen. Pada tabel A dapat dilihat hasil dari *tokenizing*.

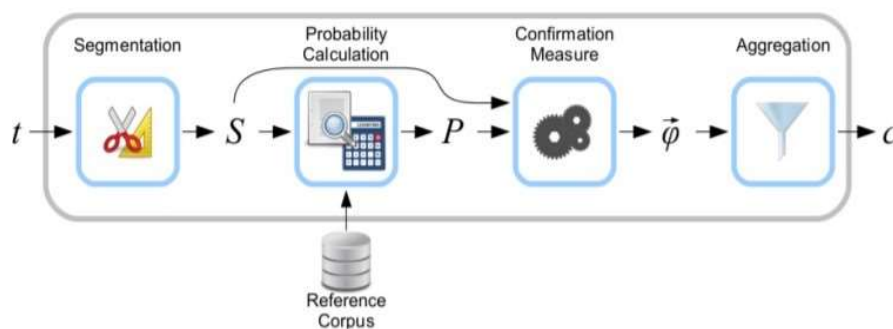
Tabel 5. Contoh proses *Tokenizing*

Tweet	Tokenizing
tahan pertama karantina wilayah jadi karantina wajah kata anies	"tahan ", "pertama", "karantina", "wilayah", "jadi", "karantina", "wajah ", "kata", "anies"
maskapai pelat merah pt garuda indonesia persero tbk layan operasional terbang kamis mei pukul	"maskapai", "pelat", "merah", "pt", "garuda", "indonesia", "persero", "tbk", "layan", "operasional", "terbang", "kamis", "mei", "pukul"

### 3.2.3 Evaluasi Pemodelan Topik

Pembagian jumlah topik pada proses LDA merupakan hal yang penting untuk dilakukan dalam proses evaluasi pemodelan topik. Hasil dari penentuan jumlah topik dapat membantu dalam menunjukkan optimal dan validnya model tersebut agar dapat dipahami oleh manusia. Teknik evaluasi untuk pemodelan topik, salah satunya adalah *Coherence score*.

*Coherence score* merupakan acuan dalam menentukan pembagian jumlah topik terbaik secara optimal. Topik *coherence* memiliki metrik yang berhubungan dengan interpretasi manusia [11]. Semakin tinggi nilai *coherence*, maka semakin bagus untuk diinterpretasikan oleh manusia [15].



Gambar 4. Tahap perhitungan *coherence score* [15]

Pada gambar 4, dapat dilihat proses untuk mendapatkan nilai *coherence*. Langkah pertama, untuk kumpulan kata  $t$  dilakukan segmentasi agar menjadi sekumpulan kata ( $S$ ). Lalu, menghitung probabilitas ( $P$ ) untuk setiap kata, dihitung berdasarkan frekuensi kata dalam korpus. Kemudian hasil dari total  $P$  dan  $S$  akan diproses di *confirmation measure* untuk menghitung kesamaan dari pasangan  $S$  yang menghasilkan nilai  $\phi$ . Terakhir semua nilai  $\phi = \phi_1, \dots, \phi_f$  dirata-rata untuk mendapatkan nilai *coherence*.

Terdapat banyak teknik yang digunakan untuk mendapatkan nilai *coherence*. Pada penelitian ini menggunakan cara  $C_v$ , karena terbukti memiliki hasil yang baik untuk diinterpretasikan ke manusia [11].



### 3.2.4 Pemodelan Topik

Proses selanjutnya yaitu menggunakan pemodelan topik dimana proses tersebut dilakukan untuk mengelompokkan dokumen-dokumen menjadi beberapa topik dari data *tweets* di setiap pengguna. Pada penelitian ini menggunakan pemodelan topik LDA dengan memanfaatkan *library gensim*. Hasil dari proses LDA pada penelitian ini dapat membantu untuk pengelompokkan tweet.

#### 1. Pemodelan topik LDA dan pembobotan kata

Setelah dilakukan proses *Pre-Processing* akan dilakukan pembuatan *dictionary*. Pada tahap ini, *dictionary* untuk memetakan setiap kata yang telah ditokenisasi ke bentuk id unik.

Selanjutnya, *term weighting* untuk mengetahui bobot setiap kata unik dan memfilter kata yang penting dari banyak dokumen. Pada penelitian ini memiliki batasan untuk setiap kata yang muncul, yaitu *no above* = 50% dan *no below* = 2. Artinya, penghapusan *term* yang sering muncul lebih dari 50% dari jumlah dokumen, dan *term* yang jarang muncul kurang dari 2 kali kemunculan pada suatu dokumen.

#### 2. Proses *Bag Of Words* (BoW)

Setelah proses *dictionary* dan pembobotan kata atau *term weighting* selesai, langkah berikutnya adalah pembuatan *Bag of Word* dengan memanfaatkan fungsi *doc2bow*. *BoW* untuk menghitung jumlah kemunculan suatu kata unik pada suatu dokumen yang memetakan antara *word.id* dan *word.frequency*.

**Tabel 6. Bentuk *Bag of Words***

[(15, 1), (16, 1), (17, 2), (18, 1), (19, 1), (20, 1),...]

Pada tabel 6 diberikan contoh dari *bag of words*. Misalnya [(17,2)] dapat diartikan bahwa id kata ke 17 muncul dua kali pada dokumen pertama. Serta, [(15,1)] mengartikan bahwa id kata ke 15 muncul satu kali, dan seterusnya.

#### 3. Perhitungan evaluasi *coherence score*

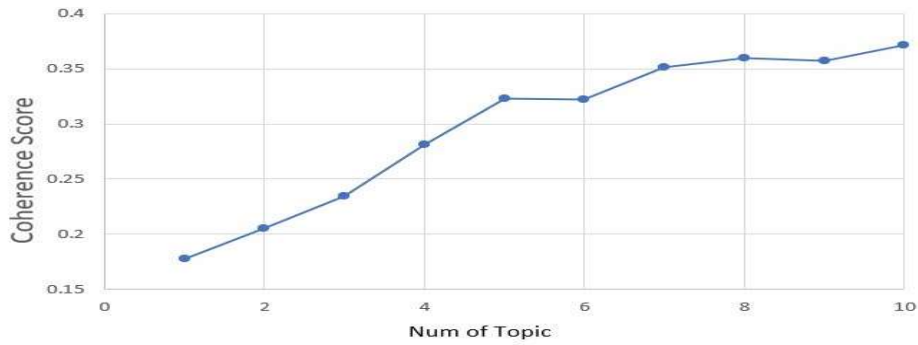
Pada penelitian ini untuk pemilihan jumlah topik memanfaatkan nilai *coherence* pada setiap jumlah pembagian topik di pemodelan LDA. Pada tahap perhitungan nilai *coherence* ini dilakukan untuk pemilihan pembagian jumlah topik dari 1 hingga 10 topik. Jumlah pembagian topik didasari oleh pembahasan topik yang berbeda-beda pada setiap dokumen dan macam-macam tema berita. Nilai *coherence* tertinggi merupakan pembagian jumlah topik terbaik.

#### 4. Analisis evaluasi *coherence score*

Evaluasi pada pemodelan topik dengan menghitung nilai *coherence* nya. Proses evaluasi ini dilakukan dengan menjalankan perhitungan *coherence* sebanyak 10 kali sebagai nilai validasi. Lalu, menghitung rata-rata setiap nilai *coherence* pada pemilihan pembagian topik 1 hingga 10 dan dapat dilihat nilai *coherence* yang optimal. Maka, pemilihan topik yang optimal dapat dijadikan sebagai parameter pada pemodelan topik LDA.

**Tabel 7. Rata-rata dari menjalankan 10 kali perhitungan *coherence***

Num Topic	Coherence Score
1	0.177652
2	0.2054439
3	0.234694
4	0.2816207
5	0.3228073
6	0.3221983
7	0.3510971
8	0.3594608
9	0.3571951
10	0.3711497



Gambar 5. Grafik coherence Score

Pada proses mengevaluasi hasil clustering, penelitian ini memanfaatkan nilai rata-rata coherence tertinggi, karena nilai coherence tertinggi menunjukkan pembagian topik terbaik. Evaluasi pemodelan topik pada penelitian ini, nilai coherence tertinggi dapat dilihat pada gambar 5.

Coherence score rata-rata tertinggi yaitu 0.3711497, yang berarti semakin tinggi coherence score, maka semakin bagus hasil interpretasi topik modeling. Berdasarkan coherence score tertinggi, pembagian topik tersebut nantinya akan dijadikan acuan pada penelitian ini, yaitu pembagian menjadi 10 topik, dapat dilihat pada tabel 8. Pada penelitian ini dapat dilihat pembentukan model LDA pada tabel 9

5. Pemberian nama pada setiap topik

Lalu, untuk pemberian nama pada setiap topik, penelitian ini memanfaatkan Wordcloud yang dapat membantu pemberian nama pada setiap topik. Pada gambar 6 dapat dilihat bahwa Wordcloud membantu pembentukan nama pada topik 0, yaitu membahas mengenai kesehatan pada keseluruhan tweet yang tergolong pada topik 0 dan topik 2 membahas mengenai berita cuaca.

Tabel 8. Nama-nama setiap topik berdasarkan Wordcloud

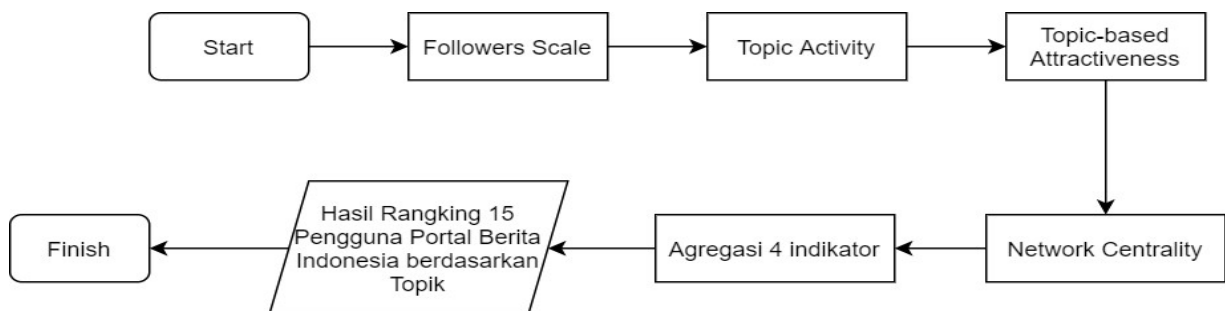
Topik	Nama Topik
0	Topik Kesehatan
1	Topik Tragedi
2	Topik Cuaca
3	Topik Olahraga
4	Topik Entertainmen
5	Topik Internasional
6	Topik Kriminal
7	Topik Hukum
8	Topik Sosial
9	Topik Daily

Tabel 9. Hasil Pembentukan Model LDA

Topik	Words
0	0.040* <b>”covid”</b> + 0.035* <b>”corona”</b> + 0.022* <b>”positif”</b> + 0.016* <b>”orang”</b> + 0.016* <b>”kasus”</b> + 0.014* <b>”pasien”</b> + 0.013* <b>”mei”</b> + 0.012* <b>”virus”</b> + 0.010* <b>”tambah”</b> + 0.009* <b>”sembuh”</b>
1	0.012* <b>”kempot”</b> + 0.012* <b>”didi”</b> + 0.009* <b>”satu”</b> + 0.009* <b>”jadi”</b> + 0.009* <b>”anak”</b> + 0.009* <b>”tahun”</b> + 0.008* <b>”dunia”</b> + 0.008* <b>”tinggal”</b> + 0.008* <b>”orang”</b> + 0.007* <b>”tak”</b>
2	0.017* <b>”kota”</b> + 0.012* <b>”hari”</b> + 0.012* <b>”sosial”</b> + 0.011* <b>”lebaran”</b> + 0.010* <b>”rp”</b> + 0.010* <b>”jokowi”</b> + 0.010* <b>”bantu”</b> + 0.009* <b>”harga”</b> + 0.009* <b>”cuaca”</b> + 0.008* <b>”bansos”</b>
3	0.012* <b>”perlu”</b> + 0.010* <b>”jadi”</b> + 0.010* <b>”yang”</b> + 0.009* <b>”makan”</b> + 0.009* <b>”tak”</b> + 0.009* <b>”badan”</b> + 0.008* <b>”nostalgia”</b> + 0.008* <b>”main”</b> + 0.008* <b>”puasa”</b> + 0.007* <b>”tongkok”</b>
4	0.011* <b>”the”</b> + 0.011* <b>”jadi”</b> + 0.010* <b>”persib”</b> + 0.009* <b>”puisi”</b> + 0.009* <b>”ajar”</b> + 0.009* <b>”rumah”</b> + 0.008* <b>”bikin”</b> + 0.008* <b>”pakai”</b> + 0.007* <b>”tak”</b> + 0.007* <b>”latih”</b>
5	0.016* <b>”psbb”</b> + 0.014* <b>”covid”</b> + 0.012* <b>”tugas”</b> + 0.011* <b>”laku”</b> + 0.010* <b>”kapal”</b> + 0.009* <b>”abk”</b> + 0.009* <b>”usaha”</b> + 0.009* <b>”perintah”</b> + 0.009* <b>”kerja”</b> + 0.008* <b>”atur”</b>
6	0.034* <b>”corona”</b> + 0.020* <b>”covid”</b> + 0.020* <b>”pandemi”</b> + 0.020* <b>”virus”</b> + 0.015* <b>”tengah”</b> + 0.009* <b>”jadi”</b> + 0.008* <b>”indonesia”</b> + 0.008* <b>”sebut”</b> + 0.008* <b>”tak”</b> + 0.008* <b>”roy”</b>
7	0.021* <b>”ferdian”</b> + 0.018* <b>”paleka”</b> + 0.015* <b>”polisi”</b> + 0.010* <b>”tahan”</b> + 0.009* <b>”online”</b> + 0.009* <b>”tangkap”</b> + 0.008* <b>”ikut”</b> + 0.007* <b>”jabar”</b> + 0.007* <b>”uu”</b> + 0.006* <b>”kamu”</b>
8	0.022* <b>”ramadan”</b> + 0.018* <b>”liga”</b> + 0.017* <b>”rp”</b> + 0.012* <b>”juta”</b> + 0.011* <b>”inggris”</b> + 0.010* <b>”kamu”</b> + 0.010* <b>”jadi”</b> + 0.008* <b>”puasa”</b> + 0.008* <b>”potret”</b> + 0.008* <b>”main”</b>
9	0.019* <b>”mei”</b> + 0.015* <b>”buka”</b> + 0.014* <b>”jadwal”</b> + 0.012* <b>”ramadhan”</b> + 0.012* <b>”puasa”</b> + 0.011* <b>”hari”</b> + 0.010* <b>”tutup”</b> + 0.009* <b>”jangan”</b> + 0.009* <b>”minggu”</b> + 0.008* <b>”sarinah”</b>



Gambar 6. Wordcloud Topik 0 dan Topik 2



Gambar 7. Proses Perhitungan TSIM

### 3.2.5 Perhitungan TSIM

Dilakukan proses perhitungan pengaruh pengguna berdasarkan topik yang telah dibentuk dalam proses LDA. Perhitungan TSIM terdiri dari empat indikator yaitu  $F_f$ ,  $F_a$ ,  $F_e$ ,  $F_c$  dan dilakukan normalisasi nilai pada setiap indikator.  $F^*$  merupakan rata-rata dari empat indikator tersebut. Dapat dilihat proses dari perhitungan TSIM di gambar 7

1. *Followers Scale* ( $F_f$ ) merupakan tahap utama dalam proses perhitungan TSIM yang memanfaatkan followers pengguna Twitter dari 15 Portal Berita. Dapat dilihat jumlah followers dari sample setiap pengguna di tabel 10.

**Tabel 10. Sampel *Followers Scale* ( $F_f$ )**

Portal Berita	Followers
KapanLagicom	2096892
CNNIndonesia	993594
kompascom	7716407
detikcom	15993396

2. *Topic Activity* ( $F_a$ ) merupakan proses perhitungan jumlah *tweets* yang termasuk pada suatu topik. Proses dilakukan menggunakan rumus yang mengacu pada persamaan (2). Hasil proses perhitungan *Topic Activity* pada topik 0 (Topik Berita Kesehatan) dapat dilihat pada sampel di tabel 11.

**Tabel 11. Sampel *Topic Activity* ( $F_a$ ) pada topik 0 (Topik Berita Kesehatan)**

Portal Berita	Topic Activity
detikcom	0.986206897
CNNIndonesia	0.889655172
kumparan	0.868965517
kompascom	0.455172414
tempodotco	0.84137931

3. *Topic-based Attractiveness* ( $F_e$ ) proses perhitungan berdasarkan jumlah retweet dan mention pengguna terhadap tweet di setiap topik. Pada tabel 12 merupakan sampel dari proses *Topic-based Attractiveness* pada topik 0 (Topik Berita Kesehatan). Proses dilakukan menggunakan rumus yang mengacu pada persamaan (3).

**Tabel 12. Sampel *Topic-based Attractiveness* ( $F_e$ ) pada topik 0 (Topik Berita Kesehatan)**

Portal	<i>Topic-based Attractiveness</i>
detikcom	0.469294536
CNNIndonesia	1
kumparan	0.087238482
kompascom	0.073929961
tempodotco	0.053290475

4. *Network Centrality* ( $F_c$ ) proses ini memanfaatkan nilai dari perhitungan PageRank, dapat dilihat pada persamaan (4). Nilai PageRank dapat diperoleh dari pengguna dalam suatu graf. Hasil dari proses *Network Centrality* berdasarkan topik 0 (Topik Berita Kesehatan) dapat dilihat sampel pada tabel 13.

**Tabel 13. Sampel *Network Centrality* ( $F_c$ ) pada topik 0 (Topik Berita Kesehatan)**

<b>Portal Berita</b>	<b>Network Centrality</b>
detikcom	1
CNNIndonesia	0.984235342
kumparan	0.366098134
kompascom	0.22056719
tempodotco	0.181739799

5. *Agregating* ( $F^*$ ) merupakan proses terakhir pada perhitungan TSIM, yang menggabungkan nilai empat indikator  $F_F$ ,  $F_a$ ,  $F_e$ , dan  $F_c$ . Agregasi pada penelitian ini menjumlahkan seluruh indikator. Hasil dari agregasi menentukan ranking setiap pengguna berdasarkan topik. Pada tabel A dapat dilihat sampel  $F^*$  berdasarkan topik 0 (Topik Berita Kesehatan).

**Tabel 14. Sampel *Agregating* ( $F^*$ ) pada topik 0 (Topik Berita Kesehatan)**

<b>Portal Berita</b>	<b>Agregasi (<math>F^*</math>)</b>
detikcom	0.863875
CNNIndonesia	0.733631
kumparan	0.337458
kompascom	0.30783
tempodotco	0.295579

### 3.3 Evaluasi Performansi

Setelah 15 pengguna Portal Berita Indonesia berhasil diranking dengan nilai agregating menggunakan metode TSIM berdasarkan topik yang telah dibentuk, maka dilakukan evaluasi performansi pada metode TSIM untuk melihat tingkat akurasi dari model yang dilakukan dalam penelitian. Proses ini bertujuan untuk melihat *error* atau kesalahan hasil ranking pada metode TSIM dengan data pembandingan. Pada penelitian ini memanfaatkan metode evaluasi yaitu *MAE* (*Mean Absolute Error*)

*MAE* (*Mean Absolute Error*) merupakan metode untuk mengukur error yang dihasilkan oleh sistem. Nilai didapatkan dari hasil pengurangan nilai aktual dan nilai peramalan, lalu dirata-rata untuk mendapatkan hasil akhir. Dalam pengukuran ini model prediksi memiliki rentang [0-1] jika semakin mendekati angka 1 maka hasil penelitian tidak akurat [18]. Berikut adalah rumus dari MAE:

$$MAE = \frac{1}{N} \sum_{i=1}^n (f_i - y_i) \quad (5)$$

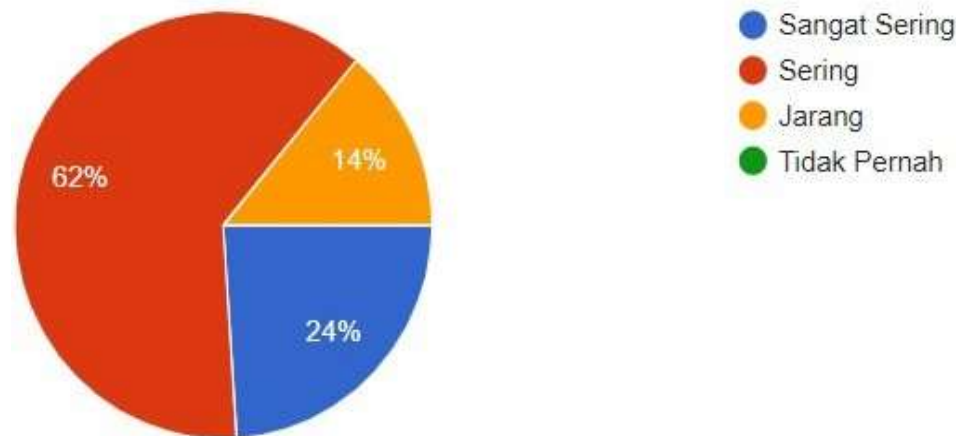
## 4. Evaluasi

Setelah menemukan ranking setiap pengguna yang paling berpengaruh berdasarkan topik menggunakan metode TSIM, dilakukan tahap evaluasi sistem. Evaluasi Sistem ini dapat mengukur kesalahan peringkat pengguna yang berpengaruh terhadap topik dengan metode MAE pada setiap topik, untuk mengetahui kesalahan prediksi setiap topik.

### 4.1 Hasil Pengujian

Hasil pengujian dilakukan dengan *survey* mengenai pendapat masyarakat secara umum untuk meranking Portal Berita di Indonesia dengan beberapa topik yang telah disediakan di kuisiner dengan menggunakan *google form*. Total responden 50 orang. Responden yang mengisi kuisiner 24% sangat sering membaca berita yaitu dari satu minggu, 6 hingga 7 hari selalu mengikuti berita atau responden memiliki kemauan yang tinggi untuk mengetahui berita terkini. Responden 62% sering membaca berita yaitu dari satu minggu, 4 hingga 5 hari sering membaca berita dan juga mengikuti akun portal berita di media sosial. Responden yang jarang membaca berita yaitu dari satu minggu hanya 2 hingga 3 hari responden membaca berita dan kurangnya minat untuk mengetahui berita terkini, yaitu 14%. Maka, penelitian ini memanfaatkan kuisiner ini sebagai data pembandingan, karena tidak ada masyarakat yang tidak membaca berita atau tidak ada responden yang tidak memiliki minat untuk mengetahui berita atau informasi yang terjadi di Indonesia pada hasil kuisiner di penelitian ini. Responden mengisi kuisiner

dengan memilih lima pengguna twitter yang menurut mereka paling berpengaruh pada masing-masing topik. *Survey* ini bertujuan untuk membandingkan setiap hasil ranking yang dinilai oleh responden dengan hasil ranking dari model. Hasil pengujian kesalahan prediksi setiap topik pada peringkat pengguna dengan metode TSIM menggunakan MAE pada rumus [5], karena nilai MAE adalah nilai rata-rata yang absolute dibandingkan dengan metode perhitungan error yang lain seperti RMSE, MAPE, dll.



Gambar 8. Diagram Responden

Pada proses mengevaluasi hasil perankingan, penelitian ini memanfaatkan nilai dari perhitungan MAE. Namun, karena memiliki hasil rentang nilai yang berbeda dari data pembandingan dan data yang dihasilkan oleh model, maka dilakukan perhitungan yang menyamakan rentang nilai dari kedua data tersebut. Hasil pengujian pada penelitian ini adalah sebagai berikut.

1. Perhitungan pada data model dilakukan dengan menjumlahkan keseluruhan agregasi dari 15 pengguna di setiap topik pada data model. Lalu, nilai agregasi setiap pengguna Portal Berita dibagi dengan total agregasi, perhitungan dilakukan pada setiap topik. Maka, menghasilkan nilai baru untuk setiap pengguna Portal Berita di setiap topik pada tabel [15]. Dapat dilihat pada tabel [15]

Portal Berita	Agregasi
detikcom	0.863875358
CNNIndonesia	0.733630877
kumparan	0.337458131
kompascom	0.307830174
tempodotco	0.295579478
pikiran_rakyat	0.266230075
merdekadotcom	0.262738514
SINDOnews	0.206569796
liputan6dotcom	0.202651031
okezonenews	0.184931431
suaradotcom	0.181034483
tribunnews	0.165532288
kompasiana	0.105485842
idntimes	0.086594032
KapanLagicom	0.063994422
<b>Total</b>	<b>4.264135934</b>

(a) Nilai model yang dihasilkan oleh sistem

Portal Berita	Model
detikcom	0.202590952
CNNIndonesia	0.172046785
kumparan	0.07913869
kompascom	0.072190516
tempodotco	0.069317555
pikiran_rakyat	0.062434706
merdekadotcom	0.061615886
SINDOnews	0.04844353
liputan6dotcom	0.047524524
okezonenews	0.043369028
suaradotcom	0.042455139
tribunnews	0.038819655
kompasiana	0.024737917
idntimes	0.020307521
KapanLagicom	0.015007594

(b) Nilai model setelah dilakukan perhitungan

Tabel 15. Perhitungan dari Nilai Model pada Topik 0 (Topik Kesehatan)

2. Perhitungan data responden atau data pembandingan dilakukan dengan menjumlahkan jumlah responden dari masing-masing portal berita. Lalu, proses pembagian dari jumlah pemilihan responden di setiap pengguna Portal Berita dengan jumlah keseluruhan pemilihan responden di setiap pengguna Portal Berita pada setiap topik. Maka, dapat menghasilkan nilai baru untuk setiap pengguna Portal Berita di setiap topik pada tabel 16b. Dapat dilihat pada tabel 16a.

Portal Berita	Jumlah Responden
detikcom	41
CNNIndonesia	38
kumparan	15
kompascom	31
tempodotco	7
pikiran_rakyat	6
merdekadotcom	3
SINDOnews	8
liputan6dotcom	21
okezonenews	14
suaradotcom	6
tribunnews	18
kompasiana	12
idntimes	19
KapanLagicom	11
<b>Total</b>	<b>250</b>

(a) Nilai jumlah responden pada setiap pengguna Portal Berita

Portal Berita	Nilai Responden
detikcom	0.164
CNNIndonesia	0.152
kumparan	0.06
kompascom	0.124
tempodotco	0.028
pikiran_rakyat	0.024
merdekadotcom	0.012
SINDOnews	0.032
liputan6dotcom	0.084
okezonenews	0.056
suaradotcom	0.024
tribunnews	0.072
kompasiana	0.048
idntimes	0.076
KapanLagicom	0.044

(b) Nilai responden pada setiap pengguna Portal Berita setelah dilakukan perhitungan

**Tabel 16.** Perhitungan dari Nilai Responden pada Topik 0 (Topik Kesehatan)

3. Tahap terakhir dilakukan perhitungan MAE dengan nilai baru dari kedua data pada tabel 5b dan tabel 6b. Dapat dilihat pada tabel 7 telah dilakukan proses perhitungan MAE untuk topik 0 (Topik Kesehatan). Pada tabel 8 dapat dilihat keseluruhan nilai MAE pada setiap topik.

**Tabel 17.** Hasil perhitungan MAE pada topik 0 (Topik Kesehatan).

Portal Berita	Model	Responden	Abs Error
detikcom	0.202590952	0.164	0.038591
CNNIndonesia	0.172046785	0.152	0.020047
kumparan	0.07913869	0.06	0.019139
kompascom	0.072190516	0.124	0.051809
tempodotco	0.069317555	0.028	0.041318
pikiran_rakyat	0.062434706	0.024	0.038435
merdekadotcom	0.061615886	0.012	0.049616
SINDOnews	0.04844353	0.032	0.016444
liputan6dotcom	0.047524524	0.084	0.036475
okezonenews	0.043369028	0.056	0.012631
suaradotcom	0.042455139	0.024	0.018455
tribunnews	0.038819655	0.072	0.03318
kompasiana	0.024737917	0.048	0.023262
idntimes	0.020307521	0.076	0.055692
KapanLagicom	0.015007594	0.044	0.028992
<b>MAE</b>			<b>0.032272</b>

Tabel 18. Hasil perhitungan MAE pada setiap topik.

Topik	MAE
Topik Tragedi	0.034085989
Topik Olahraga	0.021742168
Topik Kesehatan	0.032272432
Topik Daily	0.02845627
Topik Kriminal	0.027272408
Topik Cuaca	0.02510168
Topik Internasional	0.022975473
Topik Social	0.022305634
Topik Entertainment	0.018911518
Topik Hukum	0.0165406

#### 4.2 Analisis Hasil Pengujian

Analisis hasil pengujian pada penelitian ini menggunakan diagram batang karena data tersebut membentuk suatu kategori, sehingga lebih mudah dipahami jika diimplementasikan ke diagram batang. Hasil pengujian ini memprediksi kesalahan perangkingan yang dihasilkan oleh sistem dengan membandingkan hasil perangkingan dari kuisisioner.

Diagram dan tabel dari hasil pengujian prediksi kesalahan pada penelitian ini digambarkan melalui sebuah grafik. Hasil uji prediksi kesalahan sistem dengan menggunakan metode TSIM dapat dilihat pada diagram di gambar 9.

Nilai dari hasil perhitungan agregasi pada metode TSIM yang telah di normalisasi memiliki rentang 0-1. Nilai pada perhitungan TSIM dapat dikatakan pengguna tersebut berpengaruh pada topik tersebut. Lalu, untuk rentang nilai yang dimiliki oleh MAE adalah 0-1. Artinya, setiap nilai MAE yang mendekati angka 1 semakin menandakan bahwa ada kesalahan pada pada model [9]. Nilai MAE yang dihasilkan pada penelitian ini untuk setiap topik keseluruhan yaitu mendekati 0. Maka, hasil prediksi dari model dapat dikatakan baik, yaitu hampir tidak ada kesalahan pada model prediksi tersebut. Jika hasil prediksi ini semakin mendekati nilai 1 maka performansi model tersebut semakin buruk.



Gambar 9. Diagram hasil uji performansi.

Pada uji performansi metode TSIM dilakukan dengan menggunakan MAE. Nilai MAE terendah ditunjukkan pada topik hukum, yaitu 0.0165406. Nilai MAE tertinggi ditunjukkan oleh topik tragedi, yaitu 0.034085989. Namun, secara keseluruhan nilai setiap topik pada perhitungan MAE mendekati 0, yang berarti uji performansi pada metode TSIM di penelitian ini baik.



## 5. Kesimpulan

Penelitian ini bertujuan untuk mengetahui peringkat pengguna Portal Berita yang paling berpengaruh terhadap penyebaran berita atau informasi mengenai suatu topik dengan menggunakan metode TSIM. Dataset yang digunakan adalah data *tweet* sebanyak 11.761 *tweets* dengan pengguna Portal Berita Indonesia berdasarkan *top site* dari *website* Alexa.

Perhitungan dengan metode TSIM dilakukan dengan menggunakan empat indikator, yaitu *followers scale*, *topic activity*, *topic-based attractiveness*, dan *network centrality*. Indikator tersebut mempengaruhi peringkat pengguna Portal Berita. Dengan hal ini, dapat diketahui bahwa metode TSIM dapat membantu dalam perhitungan peringkat pada penelitian ini, yaitu menentukan peringkat 15 pengguna Portal Berita di Indonesia.

Pembagian jumlah topik terbaik menggunakan pemodelan topik LDA ditentukan oleh nilai coherence tertinggi. Nilai coherence tertinggi didapat pada pembagian 10 topik, yaitu sebesar 0.3711497. Perbedaan nilai-nilai coherence pada setiap topik dipengaruhi oleh jumlah topik yang akan dibagi. Semakin banyak pembagian topik, maka semakin banyak pembahasan pada dokumen.

Sedangkan, untuk hasil perankingan dari metode TSIM yaitu pengguna dengan ranking tertinggi di tiap topik-nya adalah detikcom dan CNNIndonesia. Hal ini dipengaruhi oleh keempat indikator perhitungan metode TSIM, yaitu indikator *followers scale*, *topic activity*, *topic-based attractiveness*, dan *network centrality*. Selain itu, hal ini juga dipengaruhi oleh tingkat popularitas masing-masing portal berita berdasarkan hasil wawancara yang dilakukan dengan para responden.

Peringkat yang dihasilkan oleh sistem menunjukkan bahwa hampir keseluruhan pengguna dengan peringkat tertinggi pada setiap topik adalah detikcom dan CNNIndonesia. Jika dilihat dari jumlah followers, jumlah followers dari pengguna CNNIndonesia bukan merupakan jumlah followers tertinggi, namun CNNIndonesia menempati peringkat kedua pada hasil perankingan menggunakan metode TSIM. Hal ini dikarenakan metode TSIM menggunakan empat indikator dalam menghasilkan peringkat pengguna, selain itu tingkat popularitas pengguna yang tinggi di mata responden juga mempengaruhi hasil uji performansi sistem. Selain itu, hal yang sangat berpengaruh pada uji prediksi perankingan ini adalah perbedaan antara peringkat yang diberikan oleh responden dan hasil peringkat dihasilkan oleh sistem. Uji performansi MAE terendah pada sistem ini ditunjukkan pada topik hukum, yaitu 0.0165406 dan rata-rata pada uji performansi MAE pada penelitian ini memiliki nilai mendekati 0. Artinya, metode TSIM pada penelitian ini memiliki performansi yang baik.

Saran untuk penelitian selanjutnya adalah sebaiknya dilakukan perbandingan antara dua metode pemodelan topik untuk membantu proses perankingan dengan metode TSIM, agar dapat diketahui pemodelan topik yang baik selain menggunakan pemodelan topik LDA.

## Daftar Pustaka

- [1] D. K. Asso Hamzehei, Shanqing Jiang and R. K. Wong. Tsim: Topic-based social influence measurement for social networks. 2016.
- [2] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [4] P. De Maeseneer and A.-P. I. for Broadcasting Development. *Here's the News: A Radio News Manual*. Asia-Pacific Institute for Broadcasting Development.
- [5] V. Embar, I. Bhattacharya, V. Pandit, and R. Vacul 'in. Online topic-based social influence analysis for the wimbledon championships. pages 1759–1768, 08 2015.
- [6] M. Furini and M. Montanero. Trank: Ranking twitter users according to specific topics. 01 2015.
- [7] L. Indonesia. *Masyarakat Indonesia*. Lembaga Ilmu Pengetahuan Indonesia., 1968.
- [8] G. Koespradono. *MEREKAYASA FAKTA MENJADI BERITA: Kreatif Menulis Efektif Menggunakan Bahasa Indonesia Jurnalistik*. Gantyo Koespradono, 2017.
- [9] A. Kulkarni and S. Satapathy. *Optimization in Machine Learning and Applications*. Algorithms for Intelligent Systems. Springer Singapore, 2019.
- [10] J. H. Lau, N. Collier, and T. Baldwin. On-line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of COLING 2012*, pages 1519–1534, Mumbai, India, Dec. 2012. The COLING 2012 Organizing Committee.

- [11]D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.
- [12]A. Nurmalasari. *Mini Book Master Bahasa Indonesia & Inggris SMA Kelas X, XI, & XII: Belajar Bahasa Indonesia & Inggris menjadi lebih mudah, lebih asyik, & gak bikin pusing*. WahyuMedia, 2012.
- [13]L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.
- [14]S. Priyanta and I. N. P. Trisna. Social network analysis of twitter to identify issuer of topic using pagerank. *International Journal of Advanced Computer Science and Applications*, 10(1), 2019.
- [15]M. R öder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.
- [16]Z. Tong and H. Zhang. A text mining research based on lda topic modelling.
- [17]C. Triwati. Metode pembobotan statistical concept based untuk klastering dan kategorisasi dokumen berbahasa indonesia. *Institut Teknologi Telkom Bandung*, 2009.
- [18]H. Widiatmoko. Studi mengenai karakteristik hubungan variabilitas cuaca musiman dengan perkembangan penyakit demam berdarah dengue (dbd): Penelitian di wilayah jakarta timur. pages 113–115, 2008.