CLUSTERING TOPIK PADA DATA SENTIMEN BPJS KESEHATAN MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION

TOPIC CLUSTERING ON SENTIMENT DATA OF BPJS KESEHATAN USING LATENT DIRICHLET ALLOCATION METHOD

Salsabilla Aliska Putri[1], Purba Daru Kusuma[2], Casi Setianingsih[3]

[1,2,3] Universitas Telkom, Bandung
1salsabillaap@student.telkomuniversity.ac.id, 2purbadarukusuma@telkomuniversity.ac.id, 3setiacasie@telkomuniversity.ac.id

Abstrak

Pemerintah Republik Indonesia dalam berupaya memberikan perlindungan kepada masyarakat melalui program jaminan sosial dengan menetapkan BPJS Kesehatan sebagai penyelenggara jaminan sosial di bidang kesehatan. Pelaksanaan program BPJS Kesehatan mendapatkan tanggapan positif, negatif, dan netral terkait kualitas layanan dan kebijakan nya oleh masyarakat melalui media sosial. Twitter sebagai salah satu media sosial untuk menyampaikan opini, kritik, dan saran pengguna nya terhadap BPJS Kesehatan. Banyaknya sentimen pengguna Twitter dapat menyulitkan dalam memahami topik pembahasan terkait kualitas layanan dan kebijakan dari BPJS Kesehatan. Penelitian ini bertujuan melakukan pengelompokan topik dari sentimen pengguna Twitter terkait BPJS Kesehatan dengan menggunakan metode Latent Dirichlet Allocation (LDA). Sehingga, dapat memudahkan dalam mengetahui topik pembicaraan yang sering dibahas oleh pengguna Twitter terkait BPJS Kesehatan. Pada tahap pengujian kinerja LDA diperoleh perplexity 6,0907 dengan nilai alpha sebesar 0.01, nilai beta sebesar 0.1, pada iterasi ke-170, dan jumlah topik 2 untuk sentimen positif. Kemudian, nilai perplexity 6,7364 dengan nilai alpha sebesar 0.001, nilai beta sebesar 0.1, pada iterasi ke-180, dan jumlah topik 2 untuk sentimen negatif. Sedangkan, untuk sentimen netral diperoleh nilai perplexity 6,2094 dengan nilai alpha sebesar 0.001, nilai beta sebesar 1, pada iterasi ke-160, dan jumlah topik 2.

Kata kunci : topic modelling, latent dirichlet allocation, BPJS Kesehatan.

Abstract

The Government of the Republic of Indonesia in an effort to provide protection to the community through social security programs by establishing BPJS Kesehatan (Health & Social Security Agency) as the provider of social security in the health sector. Its implementation received positive, negative, and neutral responses by the public through social media such as Twitter regarding the quality of its services and policies. Twitter is a platform to convey the opinions, criticisms, and suggestions of Twitter users on BPJS Kesehatan. Many sentiments of them can make it difficult to understand the discussion topics related to the quality of services and policies from BPJS Kesehatan. This study aims to group topics from Twitter user sentiment related to BPJS Kesehatan using the Latent Dirichlet Allocation (LDA) method. Hence, it can make it easier to find out the topics of conversation often discussed by Twitter users related to BPJS Kesehatan. At the performance testing on LDA stage obtained perplexity 6,0907 with the alpha value of 0.01, the beta value of 0.1, at iteration 170, and 2 topic for positive sentiment. Thus, the value of perplexity 6,7364 with the alpha value of 0.001, the beta value of 0.1, at iteration 180, and 2 topics for negative sentiment, whereas the sentiment of the neutral values obtained perplexity 6,2094 with the alpha value of 0.001, the beta value is equal to 1, at iteration 160, and 2 topics.

Keywords: topic modelling, latent dirichlet allocation, BPJS Kesehatan.

1. PENDAHULUAN

Pemerintah Indonesia terus mengembangkan program dan layanan yang diselenggarakan dalam jaminan sosial sebagai perlindungan terhadap masyarakat. Pemerintah menetapkan UU Nomor

24 Tahun 2011 mengenai Badan Penyelenggara Jaminan Sosial (BPJS) dengan menunjuk PT Askes (Persero) sebagai penyelenggara program jaminan sosial di bidang kesehatan, sehingga PT Askes (Persero) pun mengubah nama nya menjadi BPJS Kesehatan [1]. Melalui program Jaminan Kesehatan Nasional-Kartu Indonesia Sehat (JKN-KIS) yang diselenggarakan oleh BPJS Kesehatan, negara hadir di tengah kita untuk memastikan seluruh masyarakat Indonesia terlindungi oleh jaminan kesehatan yang komprehensif, adil, dan merata [1]. Dalam pelaksanaan nya, layanan dan kebijakan dari BPJS Kesehatan mendapatkan tanggapan positif, negatif, dan netral oleh masyarakat yang disampaikan melalui media sosial. Media sosial merupakan perantara penyaluran informasi yang terhubung dengan internet sehingga memungkinkan penggunanya dapat berinteraksi untuk saling bertukar informasi. Twitter menjadi salah satu media sosial yang populer di kalangan masyarakat untuk menuliskan opini nya dalam bentuk teks. Tercatat pengguna aktif harian twitter pada kuartal keempat tahun 2020 sebanyak 192 juta pengguna [2]. Banyaknya opini, kritik, dan saran dari pengguna twitter terkait layanan dan kebijakan BPJS Kesehatan dapat menyulitkan masyarakat dalam memahami topik pembahasan nya. Sehingga, perlu dilakukan clustering topik dengan pemodelan topik untuk mempermudah dalam memahami sentimen yang diberikan oleh pengguna twitter terkait BPJS Kesehatan.

Salah satu metode yang dapat digunakan dalam pemodelan topik yaitu *latent dirichlet allocation* (LDA). *Latent dirichlet allocation* (LDA) memiliki performa yang lebih unggul dibandingkan metode pemodelan topik yang lain serta dapat diimplementasikan untuk mengidentifikasi topik dalam jurnal ilmiah, klasifikasi, dan pengelompokan [3]. Pada penelitian ini dilakukan pemodelan topik yang ada pada kelompok sentimen pengguna twitter terhadap layanan dan kebijakan BPJS Kesehatan menggunakan metode pemodelan topik LDA untuk memudahkan dalam memahami topik dalam kelompok sentimen serta dapat digunakan untuk mengevaluasi kualitas layanan sekaligus sebagai bahan pertimbangan kebijakan dari BPJS Kesehatan.

2. DASAR TEORI

2.1 Media Sosial

Media sosial menurut Kaplan dan Haenin didefinisikan sebagai “sebuah kelompok aplikasi berbasis internet yang berdasar pada ideologi dan teknologi Web 2.0 yang memungkinkan penciptaan dan pertukaran konten berupa teks, gambar, video, dan sebagainya” [4]. Media sosial sebagai perantara yang memungkinkan para penggunanya dapat berdiskusi, saling berkomentar, berbagi, dan berkomunikasi dalam komunitas yang luas secara daring di era digital ini.

2.2 Text Mining

Text mining merupakan suatu istilah untuk menambang data dalam bentuk teks dengan tujuan untuk menemukan kata-kata yang dapat mewakili isi dari data sehingga dapat diketahui hubungan antar data yang lain [5]. *Text mining* mengolah data tekstual untuk dianalisis dan diproses menjadi suatu informasi dengan beberapa metode *classification*, *clustering*, dan *information retrieval* [6]. Dapat didefinisikan bahwa *text mining* sebagai proses mengolah koleksi data teks dari waktu ke waktu menggunakan beberapa metode analisis yang memiliki tujuan untuk menemukan informasi yang bermanfaat dari sumber data dan mengetahui hubungan antar data.

2.3 Preprocessing

Preprocessing merupakan tahap pertama dalam pengolahan dari data yang akan akan digunakan. *Preprocessing* memiliki beberapa tahapan yang dapat digunakan dengan tujuan untuk membersihkan data yang tidak relevan untuk digunakan dan mempersiapkan data yang tidak terstruktur menjadi data terstruktur yang dapat digunakan untuk proses berikutnya. Pada proses *text*

mining, tahap *preprocessing* sangat penting untuk dilakukan karena pada dasar nya data yang akan digunakan mempunyai dimensi yang tinggi, terpadat derau (*noise*) dan struktur data yang kurang baik [7].

2.4 Doc2bow

Dalam *Natural Language Processing* (NLP) dan *Information Retrieval* (IR) menggunakan representasi sederhana dari *bag-of-words*. *Bag-of-words* akan merepresentasikan kalimat atau dokumen sebagai kantung (*bag*) multiset dari kata – kata yang terdapat di dalamnya. *Bag-of-words* (BoW) merupakan model yang akan mempelajari kosakata dari keseluruhan dokumen dan memodelkan setiap dokumen dengan menghitung berapa kali setiap kata itu muncul [8]. Salah satu fungsi yang dapat digunakan untuk membangun BoW adalah doc2bow [9]. Fungsi doc2bow melakukan perhitungan jumlah setiap kata unik yang muncul di dokumen kemudian mengkonversi nya menjadi berformat array dan mengembalikan nilai nya menjadi sebuah vektor. Vektor *bag-of-words* dapat direpresentasikan dengan persamaan berikut:

$$d_i = (w_1, w_2, w_3 \dots w_n) \quad (1)$$

dimana:

d_i : distribusi jumlah kata ‘w’ dalam dokumen ‘d’

w : kata dalam dokumen

2.5 Latent Dirichlet Allocation

“Topic” merupakan bagian dari kumpulan kosakata yang bersifat tetap. Setiap dokumen mempunyai proporsi yang berbeda dari topik – topik yang dibahas sesuai dengan kata – kata yang ada di dalamnya [10]. Suatu dokumen mempunyai distribusi probabilitas topik yang terdiri dari distribusi kata – kata tertentu. Manusia memahami kumpulan dokumen sebagai objek yang dapat diamati, sedangkan topik merupakan bagian tersembunyi yang ada dalam suatu dokumen. *Topic modelling* menjadi salah satu metode yang dapat digunakan untuk menemukan kelompok topik tersembunyi yang ada pada dokumen dan bertujuan untuk merepresentasikan informasi yang terkandung di dalam dokumen tersebut. *Preprocessing* merupakan tahap pertama dalam pengolahan dari data yang akan akan digunakan.

Latent dirichlet allocation (LDA) merupakan salah satu metode yang paling banyak digunakan untuk pemodelan topik. *Latent dirichlet allocation* (LDA) merupakan unsupervised learning dan model probabilitas generatif yang merepresentasikan dokumen sebagai campuran acak atas topik laten dimana topik tertentu dicirikan oleh distribusi kata [11]. Latent dirichlet allocation (LDA) mengasumsikan bahwa dokumen dianggap sebagai “*bag-of-word*s” atau kantung kata – kata [12].

2.6 Perplexity

Pada penelitian yang menggunakan metode *latent dirichlet allocation* (LDA) ini dapat dilakukan evaluasi kinerja nya dengan melakukan perhitungan nilai *perplexity*. *Perplexity* merupakan pengukuran untuk mengevaluasi ketepatan informasi dari pemodelan topik dalam dokumen. Perhitungan nilai *perplexity* yaitu dengan menentukan log teks dari dokumen yang tersembunyi. Semakin rendah nilai *perplexity* maka semakin baik model yang dihasilkan dalam pemodelan topik [10]. Perhitungan nilai *perplexity* dirumuskan sebagai berikut [10]:

$$Perplexity(D_{test}) = exp\{\frac{-\sum_d \log p\ (w_d)}{\sum_d N_d}\} \quad (2)$$

dimana:

$Perplexity(D_{test})$: tingkat kebingungan dalam suatu dokumen
$p\ (w_d)$: peluang keseluruhan jumlah kata
N_d : keseluruhan jumlah kata dalam dokumen

3. PEMBAHASAN

3.1 Gambaran Umum Sistem

Proses sistem yang dibuat dimulai dengan melakukan preprocessing terhadap koleksi data melalui tiga proses yaitu, *stop words*, *stemming*, dan *tokenizing*. Selanjutnya tahap doc2bow, pemodelan topik menggunakan LDA, dan mendapatkan kesimpulan dari *topic modelling* menggunakan LDA. Tahap terakhir, menampilkan hasil kesimpulan dari pemodelan topik menggunakan LDA ke dalam web.

3.2 Perancangan Sistem

Proses sistem dimulai dengan mengolah data sentimen menggunakan proses *preprocessing*. Pada tahap *preprocessing*, hasil yang dikeluarkan merupakan data yang telah dibersihkan dari kata – kata tidak perlu dan dianggap tidak mengandung informasi, serta data yang berisi teks tersebut telah diubah menjadi bentuk kata dasar nya. Selanjutnya, hasil dari preprocessing akan digunakan sebagai data dalam fungsi doc2bow. Fungsi doc2bow akan melakukan pengubahan kata menjadi bentuk angka dan menghitung frekuensi kemunculan kata. Mengubah dalam bentuk angka, dan menghitung frekuensi dari kemunculan kata. Langkah terakhir yaitu melakukan pemodelan topik menggunakan LDA dan hasil pemodelan topik akan ditampilkan dalam bentuk gambar.

3.3 Tahapan Sistem Pemodelan Topik

3.3.1 Text Preprocessing

Preprocessing menjadi langkah penting untuk menghasilkan *topic modelling* yang lebih baik. Berikut merupakan beberapa tahapan *text preprocessing* yang digunakan pada penelitian ini:

1. *Stop Words Removal*
   Stop words merupakan sekumpulan kata yang sangat umum dalam korpus sehingga dapat dianggap tidak informatif, dan *stop words* sering dihilangkan dalam preprocessing di *text mining*.
2. *Stemming*
   Tahap *stemming* ini dilakukan dengan menghapus imbuhan pada kata yang berada pada awalan maupun akhiran kata
3. *Tokenizing*
   Tokenizing merupakan tahapan untuk menghapus spasi yang berlebih dan memenggal kata yang menyusun suatu kalimat.

3.3.2 Doc2bow

Doc2bow dalam membuat kamus akan diawali dengan proses mengubah kumpulan kata yang menyusun kalimat dalam dokumen menjadi bentuk dasar dari kata tersebut dan memenggal masing – masing kata. Berikut ini merupakan contoh data masukan dari tahap pembuatan kamus dengan fungsi doc2bow:

Tabel 3.1 Kata / Token dalam Kamus

	Token					
	1	**2**	**3**	**4**	**5**	**6**
D1	layan	bpjs	kurang	baik	iur	naik
D2	naik	iur	bpjs	sehat	wajar	jangkau
D3	iur	bpjs	sehat	resmi	naik	

Pada tabel 3.1 diatas menunjukkan hasil dari pemenggalan kata untuk pembuatan kamus. Langkah berikutnya yaitu akan dilakukan pemberian identitas pada kata yang unik dan yang memiliki kesamaan kata akan diberi identitas yang sama. Berikut ini merupakan hasil dari mengubah kata dalam dokumen dengan identitas:

Tabel 3.2 Proses doc2bow

	Token					
	1	**2**	**3**	**4**	**5**	**6**
D1	1	2	3	4	5	6
D2	6	5	2	7	8	9
D3	5	2	7	10	6	

Pada tabel 3.2 diatas menunjukkan pemberian identitas terhadap kata berdasarkan pada keunikan kata, jika terdapat kesamaan kata maka akan diberikan identitas yang sama.

4. Hasil dan Pembahasan

Pemodelan topik telah dilakukan terhadap 2117 data sentimen yang terdiri dari 714 data positif, 700 data negatif, dan 703 data netral terkait BPJS Kesehatan. Berikut hasil dari pemodelan topik LDA:

Tabel 4.1 Kelompok Topik Positif

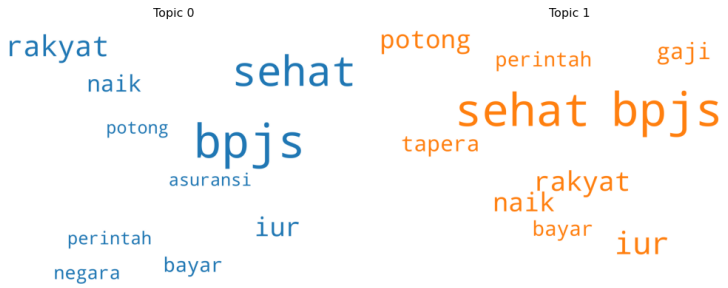
Positif			
Topik 0		**Topik 1**	
Kata	**Peluang**	**Kata**	**Peluang**
bpjs	0.093	bpjs	0.079
sehat	0.076	sehat	0.078
iur	0.034	iur	0.021
layan	0.021	serta	0.020
naik	0.019	bantu	0.019
serta	0.016	layan	0.019
perintah	0.014	gotong	0.016
masyarakat	0.013	royong	0.016
bayar	0.013	kelas	0.014
jkn	0.011	jamin	0.012
Kenaikan Iuran, Pelayanan, Peserta JKN-KIS		Prinsip BPJS Kesehatan, Bantuan Iuran, Kelas Jaminan Kesehatan	

Gambar 4.1 *Word Cloud* Kelompok Topik Positif

Pada pengujian diperoleh nilai *perplexity* 6,0907 pada parameter alpha = 0.01, beta = 0.1, iterasi = 600, dan jumlah topik = 2 sehingga akan terdapat dua topik pada sentimen positif. Berdasarkan data yang digunakan dan distribusi kata pada topik 0 dapat diambil kesimpulan bahwa topik 0 membahas mengenai kenaikan pembayaran iuran yang diberlakukan oleh pemerintah bagi peserta JKN-KIS adalah untuk memperbaiki pelayanan BPJS Kesehatan. Kemudian, untuk distribusi kata pada topik 1 diambil kesimpulan bahwa topik 1 membahas mengenai prinsip BPJS Kesehatan adalah gotong royong dengan pemerintah membantu iuran bagi yang tidak mampu untuk mendapatkan layanan dan jaminan kesehatan sesuai kelas nya.

Tabel 4.2 Kelompok Topik Negatif

Negatif			
Topik 0		**Topik 1**	
Kata	**Peluang**	**Kata**	**Peluang**
bpjs	0.081	bpjs	0.073
sehat	0.064	sehat	0.069
rakyat	0.024	iur	0.025
iur	0.020	rakyat	0.019
naik	0.015	naik	0.017
bayar	0.011	potong	0.017
negara	0.009	gaji	0.013
perintah	0.008	tapera	0.012
potong	0.008	bayar	0.011
asuransi	0.008	perintah	0.010
Pemotongan Gaji, Kenaikan Iuran, Pemerintah		Kenaikan Iuran, Pembayaran TAPERA, Pemerintah	



Gambar 4.2 *Word Cloud* Kelompok Topik Negatif

Pada pengujian diperoleh nilai *perplexity* 6,7364 pada parameter alpha = 0.001, beta = 0.1, iterasi = 300, dan jumlah topik = 2 sehingga akan terdapat dua topik pada sentimen negatif. Berdasarkan data yang digunakan dan distribusi kata pada topik 0 diambil kesimpulan bahwa topik 0 membahas mengenai protes para pekerja dan karyawan terkait pemotongan gaji dan kenaikan pembayaran iuran oleh pemerintah dan negara untuk asuransi BPJS Kesehatan. Kemudian, distribusi kata pada topik 1 diambil kesimpulan bahwa topik 1 membahas mengenai protes para pekerja terhadap tabungan perumahan rakyat (TAPERA) dengan melakukan pemotongan gaji oleh pemerintah.

Tabel 4.2 Kelompok Topik Netral

Netral			
Topik 0		**Topik 1**	
Kata	**Peluang**	**Kata**	**Peluang**
sehat	0.020	sehat	0.075
bpjs	0.016	bpjs	0.061
terima	0.010	serta	0.020
salam	0.008	layan	0.019
dm	0.008	iur	0.017
informasi	0.008	terima	0.015
kasih	0.007	kasih	0.011
data	0.007	jamin	0.011
jkn	0.006	salam	0.011
bantu	0.006	pandu	0.010
Informasi Data, dm (*direct message*), Admin Akun Twitter BPJS Kesehatan		Layanan Panduan, Informasi Iuran, Jaminan	

Gambar 4.3 *Word Cloud* Kelompok Topik Netral

Pada pengujian diperoleh nilai *perplexity* 6,2094 pada parameter alpha = 0.001, beta = 1, iterasi = 500, dan jumlah topik = 2 sehingga akan terdapat dua topik pada sentimen netral. Berdasarkan data yang digunakan dan distribusi kata pada topik 0 diambil kesimpulan bahwa topik 0 membahas permintaan informasi data peserta yang dibantu dalam pengecekan oleh admin akun twitter resmi BPJS Kesehatan melalui dm (direct message) dengan menunjukkan kartu jkn-kis. Kemudian, distribusi kata pada topik 1 diambil kesimpulan bahwa topik 1 membahas mengenai permintaan layanan panduan BPJS Kesehatan untuk peserta jaminan sosial dan pertanyaan seputar iuran pada akun twitter resmi BPJS Kesehatan.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan pengujian yang dilakukan pada penelitian tugas akhir ini, maka dapat disimpulkan bahwa:

1. Sistem *clustering* topik dari sentimen pengguna twitter terhadap layanan BPJS Kesehatan sudah berjalan dengan baik dengan hasil uji alpha atau fungsionalitas sebesar 100%
2. Sistem *clustering* topik dari sentimen pengguna twitter terhadap layanan BPJS Kesehatan dengan menggunakan pemodelan topik Latent Dirichlet Allocation (LDA) diperoleh perplexity 6,0907 dengan nilai alpha sebesar 0.01, nilai beta sebesar 0.1, pada iterasi ke-170, dan jumlah topik 2 untuk sentimen positif. Kemudian, nilai perplexity 6,7364 dengan nilai alpha sebesar 0.001, nilai beta sebesar 0.1, pada iterasi ke-180, dan jumlah topik 2 untuk sentimen negatif. Sedangkan, untuk sentimen netral diperoleh nilai perplexity 6,2094 dengan nilai alpha sebesar 0.001, nilai beta sebesar 1, pada iterasi ke-160, dan jumlah topik 2.

5.2 Saran

Berdasarkan hasil penelitian, analisis, dan pengujian pada penelitian tugas akhir ini maka saran yang dapat disampaikan untuk penelitian selanjutnya adalah untuk melakukan penelitian dengan menerapkan metode lainnya seperti *Latent Semantic Analysis* (LSA) untuk pemodelan topik kedepannya.

REFERENSI

[1] Humas, "BPJS Kesehatan," BPJS Kesehatan, 20 September 2018. [Online]. Available: https://www.bpjs-kesehatan.go.id/bpjs/pages/detail/2013/4. [Accessed 13 November 2020].

[2] K. Wagner, "Twitter Jumps Most in a Year After Sales Top Estimates," Bloomberg, 2021.

[3] J. T. L. T.-O. R. J. M. E. B. R. A. A. W. F. PATRICIO ZAMBRANO, "Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach," *SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE IN CYBERSECURITY*, vol. 7, pp. 142129 - 142146, 2019.

[4] M. H. Andreas M.Kaplan, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59 - 68, 2010.

[5] J. S. Ronen Feldman, The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data, Cambridge: Cambridge University Press, 2009.

[6] E. M. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study," *International Journal of Computer Applications (0975 – 8887)*, vol. 112, no. 5, pp. 44 - 48, 2015.

[7] I. A. Ph.D., "Text Mining dan Knowledge Discovery," *Kolokium bersama komunitas datamining Indonesia & soft-computing Indonesia*, Vols. 1 - 9, 2006.

[8] ,. P. R. a. R. Deepu S, "A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction," *International Journal of Advanced Networking & Applications (IJANA)* , pp. 320 - 323, 2016.

[9] J. N. a. P. Ircing, "Unsupervised Document Classification and Topic Detection," *Springer International Publishing AG*, pp. 748 - 756, 2017.

[10] A. Y. N. a. M. I. J. David M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3*, pp. 993 - 1022, 2003.

[11] S. T. P. D. P. T. D. S. M. Yaswanth Kalepalli, "Effective Comparison of LDA with LSA for Topic Modelling," *Proceedings of the International Conference on Intelligent Computing and Control Systems*, pp. 1245 - 1250, 2020.

[12] N. I. a. K. F. Siti Qomariyah, "Topic Modeling Twitter Data Using Latent Dirichlet Allocation and Latent Semantic Analysis," *AIP Conference Proceedings*, vol. 2194, no. 1, pp. 1 - 7, 2019.