ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

Pelabelan Data Dengan Latent Dirichlet Allocation dan K-Means Clustering pada Data Twitter Menggunakan Bahasa Indonesia

Data Labeling using Latent Dirichlet Allocation and K-Means Clustering on Indonesian-Based Twitter

Patrick Adolf Telnoni[1], Suryatiningsih[2], Ely Rosely[3]

[1,2,3]Diploma Sistem Informasi, Fakultas Ilmu Terapan, Universitas Telkom

[1]patrick.telnoni@tass.telkomuniversity.ac.id, [2]suryatiningsih@tass.telkomuniversity.ac.id, [3]ely.rosely@tass.telkomuniversity.ac.id

Abstrak

Media sosial telah cukup lama menjadi sumber utama untuk melakukan sentimen analisis terhadap suatu trend. Hal ini juga terjadi pada media sosial Twitter yang telah ada lebih dari satu dekade. Data yang bervolume besar pada twitter sangat bermanfaat untuk mencari sentimen. Akan tetapi data pada twitter umumnya merupakan data yang tidak berkategori. Data yang tidak berkategori tersebut dapat diberikan label menggunakan metode *unsupervised learning*, akan tetapi metode ini memberikan akurasi yang buruk. Untuk memperoleh akurasi yang baik dari metode *unsupervised learning*, perlu digunakan teknik *Deep Learning* seperti *Generative Adversarial Networks* (GAN), namun teknik ini membutuhkan komputasi yang sangat tinggi. Paper ini memberikan solusi pelabelan data dengan teknik mudah dankomputasi yang ringan menggunakan data dari twitter pada bahasa Indonesia. Teknik yang digunakan pada paper ini adalah k-means clustering dan Latent Dirichlet Allocation (LDA). Hasil dari percobaan ini menunjukan bahwa dari 3 klaster label data yang digunakan memiliki sebaran data yang tidak seimbang, di mana klaster 1 berisi 200 data, klaster 2 berisi 50 data, dan klaster 3 berisi 1200 data. Dari model LDA yang dibangun, diperoleh *log likelihood* sebesar -24842.65 dan *perplexity* sebesar 1859.5279290335732.

Kata kunci : text labeling, text summarization, kmeans clustering, latent dirichlet allocation

Abstract

Social media has been the primary source for sentiment analysis, especially when a topic surges. Twitter, which has been around for more than a decade, the voluminous data will become a deadly weapon to find the sentiment. However, determining label for uncategorized voluminous text data is problem often occurred in upon doing sentiment analysis. Performing scripting or even unsupervised learning to this ungrouped data such as K-Means clustering will give poor quality of the label, while using Deep Learning technique such as Generative Adversarial Networks (GAN) will consume high computing process. Hence, this paper give alternative to do data labeling, particularly in sentiment analysis purpose without using excessive computing power. This paper will present combining k-means clustering and latent dirichlet allocation to extract topic from twitter data in indonesian language. The result from this research is that while it can extract topic from each cluster of text data, the distribution in each cluster is highly unbalance. The result from the experiment shows that first cluster contain 200 data, second cluster contain 50 data, and third cluster contain 1200 data. From the proposed model, this research aquired log likelihood score at -24842.65 and perplexity score at 1859.5279290335732.

Keywords: text labeling, text summarization, kmeans clustering, latent dirichlet allocation

1. PENDAHULUAN

Pelabelan data merupakan proses untuk mengkategorikan kumpulan data ke dalam kelompok tertentu. Proses ini merupakan fase awal ketika melakukan pekerjaan yang berkaitan dengan *machine learning*.

Terdapat beberapa metode untuk melakukan pelabelan data. Yang pertama adalah dengan memberikan data secara manual, baik menggunakan tim internal, menyewa pihak ketiga, maupun

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

melakukan *crowdsourcing*. Bagaimanapun, metode ini memakan waktu lama dan membutuhkan banyak biaya pada kasus tertentu.

Alternatif lainnya adalah membuat script sederhana untuk melakukan pelabelan, namun akan diperoleh dataset dengan kualitas yang rendah. Cara lain untuk melakukan pelabelan data dengan akurasi yang lebih tinggi dapat diperoleh dengan menggunakan teknik *deep learning* seperti *Generative Adversarial Networks* (GANs) [1], tetapi metode ini membutuhkan komputasi yang sangat tinggi [2].

Text summarization adalah teknik dalam Natural Language Processing (NLP) untuk merangkum teks dalam jumlah besar. Tujuan dari teknik ini adalah untuk memperoleh ide utama dari dokumen yang dirangkum. Text summarization juga dapat digunakan untuk mengekstrak keyword dari sebuah dokumen, sehingga membantu praktisi NLP dalam melakukan pelabelan data. Salah satu metode text summarization yang unsupervised adalah Latent Dirichlet Allocation (LDA). LDA terkenal karena kesederhanaan dan modularitasinya [3]. LDA awalnya diperkenalkan pada tahun 2003 dan banyak dikembangkan setelahnya ekstension untuk algoritma ini.

Penelitian yang serupa dengan dengan artikel ini menggunakan LDA dan KNN untuk melakukan kategorisasi teks [4]. Pada penelitian tersebut digunakan data teks corpus Cina dari Universitas portal berita menggunakan 870 data training dan 787 data tes. Hasilnya, LDA memberikan akurasi 93,3% dibandingkan dengan Lateng Semantic Index (LSI) dengan 90% akurasi. Terdapat juga penelitian serupa dengan yang ditulis artikel ini menggunakan K-Means Clustering dan LDA [5]. Pada penelitian tersebut, digunakan data dari portal berita. Berbeda dengan penelitian tersebut, pada penlitian ini digunakan data dari Twitter yang memiliki karakteristik berbeda dari teks pada portal berita.

Usaha untuk melakukan pelabelan data juga dilakukan dengan mensilangkan data twitter dengan data dari reddit menggunakan subforum / subreddit yang juga menggunakan LDA [6]. Metode ini termasuk metode yang mudah dilakukan, mengingat reddit dan twitter memiliki karakteristik yang mirip. Hanya saja, pada penelitian ini, domain reddit.com diblokir oleh hampir semua provider internet. Selain itu, basis pengguna reddit di Indonesia tidak sebesar pengguna Twitter. Pada penelitian tersebut, diperoleh presisi rata-rata sebesar 75.62%.

Penelitian lain yang serupa dengan ide pada artikel ini juga dilakukan pada bahasa India [7]. Namun, mengingat bahasa yang digunakan adalah bahasa India, menjadi tidak kompatibel dengan bahasa Indonesia. Berdasarkan banyaknya adopsi LDA dan akurasi yang diperoleh pada penelitian [4, 5, 6], maka penelitian ini menggunakan LDA dikombinasikan dengan kmeans untuk menentukan topik dari tiap klaster.

2. METODOLOGI

2.1 Dataset

Dataset yang digunakan diambil dari Twitter secara langsung. Karena Twitter memiliki banyak kata-kata yang tidak baku, khususnya dalam bahasa Indonesia, maka pengambilan data ditambahkan filter pada query API Twitter. Filter yang ditambahkan adalah, data yang diambil hanya dari akun resmi, sehingga diperoleh bahasa yang lebih formal. Data yang sifatnya *retweet* dikeluarkan dari *query*, sehingga tidak diperoleh data duplikat.

Mengambil data dari Twitter mempunyai kesulitan tersendiri. Hal ini karena API Twitter memeliki beberapa batasan. Pertama, Twitter tidak dapat memberikan data lebih dari 1 minggu terakhir. Kedua, setiap request penarikan data, dibatasi hanya sekitar 300 baris data yang dikembalikan. Menggunakan *scheduled task* untuk mengambil data melalui API juga berpotensi mengembalikan data yang sama. Oleh karena itu, dibuat dua buah *crawler* dalam penelitian ini, yaitu crawler melalui API dan crawler yang mengambil data menggunakan *web scrapping*. Dengan teknik *web scrapping*, bisa diperoleh data yang berada lebih dari 1 minggu sebelumnya. Dari dua *crawler* tersebut, diperolah kurang lebih 1300 data.

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

Data teks yang diperoleh kemudia dibersihkan dari karakter yang tidak perlu. Karakter-karakter yang dihapus antara lain RT, @ yang mengandung *username*, url yang mengantung karakter http dan https. Setelah menghapus karakter noise, seluruh data dikonversi ke dalam karakter huruf kecil untuk menghindari operasi yang sifatnya *case sensitive*.

Khusus untuk data dari *crawler web scrapping*, terdapat *pre-processing* lanjutan. Karena data dari *crawler* ini bercampur dengan usernamem sehingga perlu dilakukan proses splitting untuk mendapat teks tweetnya saja. Setelah teks asli diperoleh, tiap teks akan diperiksa dan dirubah ke dalam bentuk formal, jika terdapat kosa kata non formal.

2.2 Representasi Teks

Sebelum dimasukan ke dalam algoritma classifier, setiap teks pada dataset dirubah ke dalam bentuk vektor numerik menggunakan teknik *bag of words*. Pada teknik ini, frekuensi kemunculan sebuah kata akan dihitung, tanpa memperdulikan urutan dari kata tersebut. Secara detail, teknik *bag of words* yang digunakan adakan *Term Frequency - Inverse Document Frequency* atau yang biasa disingkat TF-IDF Pada TF-IDF, dilakukan juga pengeleminasian kata-kata umum atau yang biasa disebut *stopwords*. *Stopwords* yang digunakan adalah *stopwords* dari bahasa Indonesia. Setelah itu, *unigram* dan *bigram* akan diambil untuk tiap kategori. *Unigram* adalah sebuah padanan kata yang paling sering muncul dari sebuah kategori, sedangkan *bigram* adalah dua kata yang paling sering muncul dari sebuah kategori.

2.3 K Means Clustering

K-Means clustering adalah algoritma *machine learning* yang sifatnya adalah *unsupervised*. Pada algoritma ini, digunakan klaster untuk mengelompokan N obyek menjadi klaster sejumlah K, di mana K dan N adalah bilangan bulat positif. Dalam algoritma ini, proses pengelompokan dilakukan dengan menghitung *centroid* dari koordinat yang disediakan. *Centroid* merupakan titik tengah yang dipilih sebanyak jumlah K yang ditentukan. Umumnya *centroid* dipilih secara random [8].

Setiap data koordinat dikalkulasikan kedekatannya dengan tiap *centroid* sampai seluruh data yang ada mendapat satu klaster. Kedekatan ini dihitung menggunakan *eucledian distance*. *Eucledian distance* yang paling kecil dengan sebuah *centroid* itulah yang akan digunakan untuk memasukan sebuah data ke dalam klaster. Persamaan 1 menunjukan rumus untuk *eucledian distance*. Pada persamaan 1, *x* adalah koordinat ke-*i* dari data pertama dan *w* adalah koordinat ke-*i* dari data kedua.

$$||p|| = \sqrt{\sum_{i=1}^{n}(x_i - w_i)^2} \quad (1)$$

Setelah tiap klaster terisi dengan data, dikalkulasikan *centroid* yang baru dengan menghitung rata-rata tiap komponen dari masing-masing data. Dari *centroid* baru yang diperoleh, tiap data yang ada dievaluasi kembali kedekatannya dengan *centroid* baru. Pada langkah ini, dapat terjadi perpindahan klaster dari sebuah data. Hal ini akan membuat klaster baru memperoleh anggota baru, dan melakukan perhitungan *centroid* yang baru. Perhitungan *centroid* yang baru ditunjukan oleh persamaan 2. Pada persamaan 2, *N* adalah jumlah data dalam klaster, sedangkan *x* adalah data yang berada dalam klaster tersebut, dari data pertama hingga data ke-*N*.

$$C_i = \frac{1}{Ni}\sum_{x \in x_i} x, i \; = \; 1, 2, \ldots k \quad (2)$$

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

Setelah *centroid* baru terbentuk dan tidak ada perpindahan data ke klaster yang lain, algoritma ini berhenti melakukan iterasi dan menghasilkan klaster final.

2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) adalah salah satu *topic modelling* yang sangat populer. Ide utama dari algoritma ini adalah sebuah dokumen terdiri dari lebih dari satu topik, di mana tiap topik memiliki katakata yang berkolerasi. Contohnya, artikel mengenai COVID-19 dapat berkorelasi dengan kesehatan, politik, dan ekonomi. Topik mengenai kesehatan berkorelasi dengan keywords seperti obat, gejala, sementara politik berhubungan dengan keyword presiden, lockdown, serta ekonomi berkorelasi dengan kata resesi. LDA mencari topik beserta kata-kata komposisinya [9][10].

Secara detail, LDA mereprentasikan k topik ke dalam distribusi multinomial dari kumpulan kata-kata. Tiap topik terdiri dari keywords dan tiap keywords direpresentasikan dengan distribusi diriclet terhadap sebuah topik. Contohnya, sebuah dokumen dengan komposisi Topik 1: 0.8, Topik 2: 0.0, Topik 3: 0.1, Topik 4: 0.1 dan Topik 1 memiliki komposisi Presiden: 0.5, pemilu: 0.25, kpu: 0.25. Dalam LDA, terdapat dua parameter yang diprediksi, yaitu φ (*parts-versus-topics*) dan θ (*composites-versus-topics*). φ and θ dihitung dari α (*document-topic density*) dan β (*topic-word density*).

Secara matenatis, LDA dinyatakan dengan persamaan 3.

$$p(w) = \int_{\theta} \left(\sum_{z_n=1}^{k} p(w_n \mid z_n;\beta)p(z_n|\theta)\right)p(\theta;\alpha)\, d\theta, \quad (3)$$

di mana p(θ; α) adalah dirichlet, $p(z_n|\theta)$ adalah distribusi multinomial yang menggunakan parameter θ and $p(w_n|z_n)$; β adalah distribusi multinomial dari kumpulan kata-kata [8].

2.5 Elbow Method

Setelah diperoleh data dua dimensin, proses pembentukan klaster dapat dilakukan. Sebelum melakukanya proses klastering, perlu dicari nilai k yang optimal. Untuk memperoleh nilai k, digunakan *elbow method*, yang ditunjukan pada persamaan 4 [11].

$$\boldsymbol{SSE} = \sum_{k=1}^{k} \sum_{x \in x_i} || X_i - C_k || \quad (4)$$

Dari persamaan 5, *Sum Squared Error* (SSE) digunakan untuk menghitung *elbow method* dengan menjumlahkan rata-rata *eucledian distance* dari tiap data terhadap centroid. Pada gambar 2, ketika terjadi tikungan tajam pada perubahan nilai, itulah titik nilai optimal dari K. Dari K = 2 dan nilai SSE ditambah pada tiap itersi, di mana $K_n = K + 1$, margin terbesar SSE di mana SSE_{Kn-1} adalah titik di mana nilai optimal K diperoleh.

3. METODE USULAN

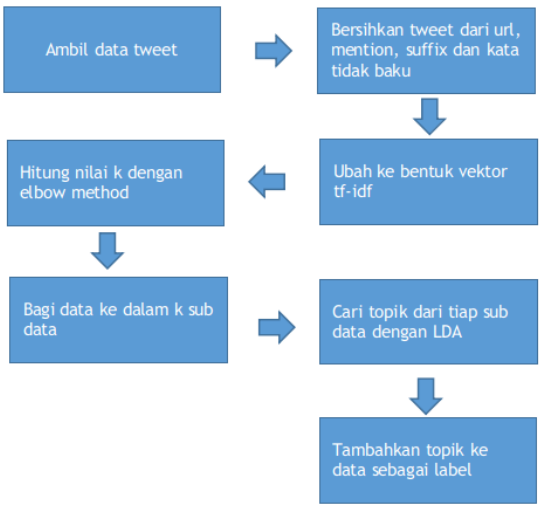
Setelah diperoleh teks yang telah diformalkan, data dirubah ke dalam vektor menggunakan TF-IDF. Ini dilakukan untuk melakukan *feature extraction*. *Feature* yang digunakan adalah *unigram* dan *bigram*. Karena Twitter membatasi karakter per post menjadi 280 karakter, menggunakan N-gram dengan jumlah tinggi akan menjadi tidak akurat.

Vektor hasil TF-IDF akan dikelompokan menggunakan *K-Means Clustering*. Karena tiap vektor berisi angka dari frekuensi kemunculan kata, vektor-vektor tersebut dapat langsung dioperasikan ke dalam algoritma K-Means. Jumlah klaster dikalkulasikan terlebih dahulu

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

menggunakan *elbow method*. Hasil dari *elbow method* akan digunakan untuk mengelompokan vektor TF-IDF.

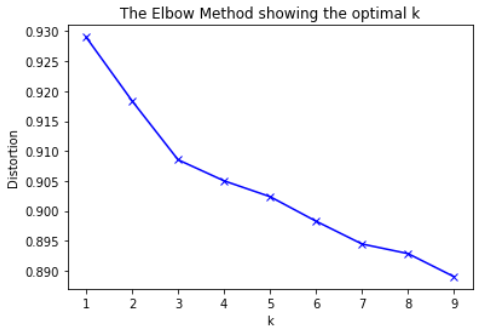
Dari data tiap klaster dilakukan dekomposisi topik dan *keywords* menggunakan LDA. Setelah topik dan data diperoleh dari LDA, topik dari LDA akan digunakan sebagai label pada tiap data. Gambar 1 menunjukan metode yang diusulkan.



Gambar 1. Metode usulan

4. HASIL DAN DISKUSI

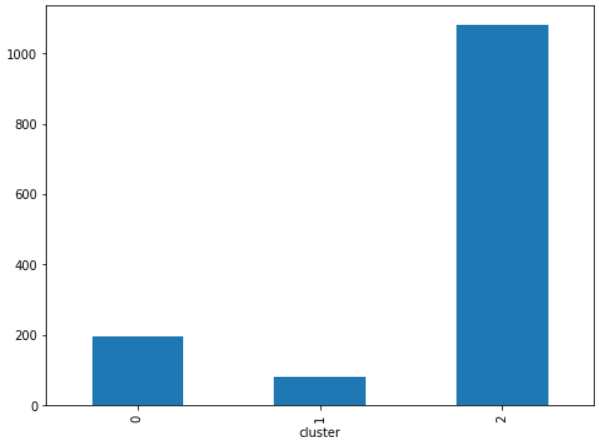
Menggunakan vektor dari TF-IDF, teks dikonvesi ke dalam corpus vektor. Karena jumlah klaster kecil, *elbow method* mengkalkulasikan kemunkinan jumlah klaster dari 1 hingga 2. Gambar 2 menunjukan nilai K yang dikalkukasikan dengan pada *elbow method* [11].



Gambar 2. Kurva *elbow method*

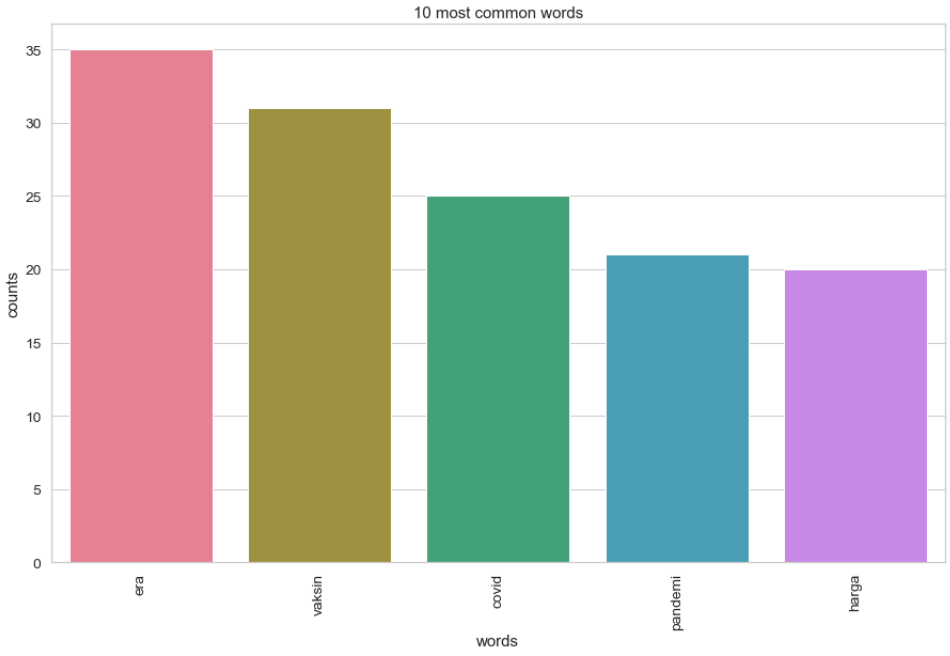
Dari gambar 2, terlihat bahwa nilai K yang optimal adalah 3. Setelah model *k-means* melalui fase training, data tes dimasukan ke dalam model sehingga diperoleh klaster untuk tiap data. Gambar 3 menunjukan proporsi klasterisasi data menggunakan *k-means*.

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020



Gambar 3. Sebaran data tiap klaster

Terlihat bahwa distribusi data pada tiap klaster tidak merata. Tidak meratanya distribusi ini Akan tetapi karena data yang diperoleh berjumlah 1300 dan jumlah data pada klaster terkecil adalah 100, proses dekomposisi tetap dilanjutkan. Dari proses dekomposisi ini, diperoleh topik dan kata yang ditunjukan oleh gambar 4 yang menunjukan komposisi pada salah satu klaster.



Gambar 4. Sebaran topik pada klaster

Tidak meratanya sebaran data pada ketiga klaster tersebut disebabkan oleh representasi numerik dari data teks pada klaster 3 memiliki frekuensi kemunculan yang lebih tinggi dibandingkan pada 2 klaster lainnya. Hal ini sejalan dengan sifat bawaan TF-IDF yang memperhitungkan frekuensi kemunculan kata sebagai representasi numeriknya. Meski data tidak seimbang, namun LDA berhasil melakukan dekomposisi dari teks pada tiap klaster.

Akurasi model LDA yang dibangun diukur menggunakan *log likelihood* dan perplexity. Semakin tinggi nilai *log likelihood*, semakin baik pula model yang dibangun dan semakin rendah nilai *perplexitiy*, semakin baik pula model yang dibangun. Dari data yang dikumpulkan dan

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

dimasukan ke dalam LDA, diperoleh *log likelihood* sebesar -24842.65 dan *perplexity* sebesar 1859.5279290335732. Angka itu diperoleh dengan menggunakan *learning rate* sebesar 0,7 dan jumlah topik sebanyak 5. Kedua nilai ini menunjukkan bahwa model LDA yang dibangun belum cukup optimal.

```
Log Likelihood:  -24835.458925378698
Perplexity:  1855.4761821263853
```

Gambar 5. Nilai *log likelihood* awal

Dengan menggunakan GridSearch, diperoleh parameter yang cocok untuk learning rate sebesar 0,7 dan jumlah topik optimal sebesar 10. Dengan mengubah kedua parameter tersebut, diperoleh nilai *log likelihood* sebesar -11538.783417564895 namun meningkatka *perplexity* menjadi 1863.0992690939918.

```
Best Model's Params:  {'learning_decay': 0.7, 'n_components': 10}
Best Log Likelihood Score:  -11539.261899430907
Model Perplexity:  1859.3985198317696
```

Gambar 6. Nilai *log likelihood* setelah optimasi

Meski dengan parameter yang telah diatur, nilai *log likelihood* masih sangat rendah dan bernilai negatif. Hal ini dikarenakan oleh data teks yang diambil dari twitter memiliki banyak singkatan dan kalimat tidak baku dalam bahasa Indonesia. Kata-kata singkatan tersebut menjadi tidak tersaring dan mempunyai representasi numerik yang berdekatan sehingga membuat salah satu klaster mempunyai nilai yang sangat besar.

4. KESIMPULAN

Dari hasil percobaan, dapat disimpulkan bahwa kmeans dan LDA dapat digunakan untuk pelabelan data secara otomatis namun dengan akurasi yang masih sangat kurang. Faktor dari platform media sosial twitter yang membuat banyak usernya menulis teks dengan singkatan membuat salah satu klaster memiliki lebih banyak angoota daripada klaster yang lain. Teks singkatan ini mempengaruhi nilai log likelihood dan perplexity. Untuk saran pengerjaan ke depan, sebaiknya menggunakan data dari reddit dibandingkan dari twitter.

DAFTAR PUSTAKA

[1] P. Parvathi and T. S. Jyothis, "Identifying Relevant Text from Text Document Using Deep Learning," *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, Kottayam, India, 2018, pp. 1-4, doi: 10.1109/ICCSDET.2018.8821192.

[2] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The computational limits of deep learning,” 2020.

[3] M. Al-Ghossein, P.-Murena, T. Abdessalem, A. Barré, and A. Cornuéjols. 2018. Adaptive collaborative topic modeling for online recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (*RecSys '18*). Association for Computing Machinery, New York, NY, USA, 338–346.

[4] W. Chen and X. Zhang, "Research on text categorization model based on LDA — KNN," *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2017, pp. 2719-2726, doi: 10.1109/IAEAC.2017.8054520.

ISSN (p) : 2407-1323
ISSN (e) : 2442-4404
DOI : https://doi.org/10.25124/jett.v7i2.3442
| Vol. 7 | No. 2 | Halaman : 885 - 892
JETT
Jurnal Elektro Telekomunikasi Terapan Desember 2020

[5] S. Twinandilla, S. Adhy, B. Surarso, and R. Kusumaningrum, "Multi-document summarization using k-means and latent dirichlet allocation (lda) – significance sentences," *Procedia Computer Science*, vol. 135, pp. 663 – 670, 2018. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

[6] A. Fiallos and K. Jimenes, "Using reddit data for multi-label text classification of twitter users interests," in *2019 Sixth International Conference on eDemocracy eGovernment (ICEDEG)*, pp. 324–327, 2019.

[7] D. Hemavathi, M. Kavitha, and N. Begum Ahmed, "Information extraction from social media: Clustering and labelling microblogs," in *2017 International Conference on IoT and Application (ICIOT)*, pp. 1–10, 2017.

[8] J. Wu. 2012. *Advances in K-means Clustering: A Data Mining Thinking*. Springer Publishing Company, Incorporated.

[9] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, 11 2018.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, p. 993–1022, Mar. 2003.

[11] D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," *2018 International Seminar on Application for Technology of Information and Communication*, Semarang, 2018, pp. 533-538, doi: 10.1109/ISEMANTIC.2018.8549751.