JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

PEMODELAN TOPIK SKRIPSI MENGGUNAKAN METODE LDA

Pijar Candra Mahatagandha[a1], Ngurah Agus Sanjaya ER[a2]

[a]Program Studi Informatika, Universitas Udayana
Kuta Selatan, Badung, Bali, Indonesia
[1]pijarcandra22@gmail.com
[2]agus_sanjaya@unud.ac.id

Abstract

Writing is one of the most important activities to express ideas in visual form. However, many students have difficulty in writing research. The difficulty experienced by students lies in choosing a topic that will be developed in the background. The topics needed in research writing are topics that are currently being discussed by the community so that initial observations are needed to get the appropriate topic. However, not infrequently the observation stage will take a long time, so a solution is offered by building an LDA model to assist students in choosing a thesis topic. This study uses an unsupervised learning algorithm, namely LDA or (Latent Dirichlet Allocation) for topic modeling. The best model produced is the LDA model with a combination of 40 topics, alpha 0.4, beta symmetric, and corpus tf-idf with a coherence score of 0.43 and a perplexity of 199.09.

Keywords: *LDA, Twiter, Topic Modeling, Research Topic, Text Mining*

1. Pendahuluan

Menulis merupakan salah satu kegiatan yang sangat penting untuk menuangkan ide dalam bentuk visual. Menulis tidak hanya aktivitas untuk mentransfer ide, namun juga berkaitan dengan pengetahuan yang dimiliki oleh penulis termasuk substansi, pengembangan *thesis sentence*, dan relevansi penampilan topik [1]. Adanya hubungan antara tulisan dengan pengetahuan penulis menyebabkan kemampuan ini menjadi salah satu syarat kelulusan di beberapa Universitas untuk menilai kemampuan lulusan dari skripsi atau tugas akhir yang dibuat. Namun menulis menjadi hambatan bagi mahasiswa dalam mengerjakan skripsi dikarenakan adanya beberapa faktor seperti pemilihan topik, pengembangan ide, kekurangan kosakata, dan lain sebagainya [2].

Pada penelitian yang berjudul "Analisis Kesulitan Mahasiswa dalam Mengembangkan Ide pada *Basic Writing*" oleh Bela Aprilia menunjukan bila 80% responden menyatakan bila kesulitan utama dalam menulis skripsi adalah memilih topik yang berkualitas [1]. Hasil yang sama juga ditunjukan pada penelitian "Identifikasi Kesulitan Mahasiswa dalam Penyusunan Skripsi Prodi PGSD Universitas Mataram" oleh Muhammad Irawan Zain, menunjukan bila 46% kesulitan menulis skripsi terletak pada pembuatan latar belakang, yang merupakan pengembangan awal dari topik yang dipilih oleh penulis [2]. Syarat sebuah topik penelitian yang baik adalah dekat dengan masalah yang saat ini tengah dihadapi oleh masyarakat, namun untuk mendapatkanya penulis perlu melakukan observasi awal terhadap kebutuhan masyarakat serta mengetahui topik hangat yang saat ini tengah diperbincangkan oleh masyarakat. Tahap observasi awal ini tentu akan memakan waktu yang lama sehingga diperlukan sebuah solusi untuk mempercepat tahap ini.

LDA atau *Latent Dirichlet Allocation* merupakan algoritma yang dapat digunakan untuk melakukan pemodelan topik dari beberapa dokumen yang dimasukan. Pemodelan topik menggunakan LDA dapat menjadi solusi untuk mengatasi kesulitan mahasiswa untuk mendapatkan topik penelitian. Pada

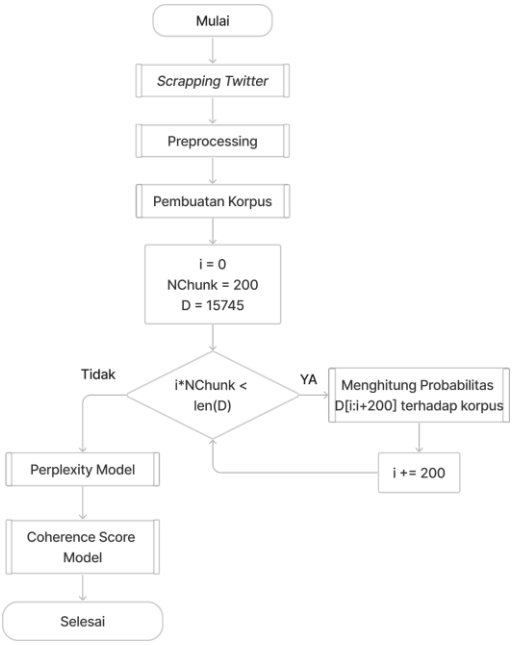
penelitian “Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (LDA)” menunjukan bila hasil yang meyakinkan dengan faktor distribusi probabilitas dokumen secara mayoritas ada pada rentang 0.5 sampai 0.99.

Berdasarkan latar belakang yang telah disampaikan maka peneliti menulis jurnal yang berjudul “Pemodelan Topik Skripsi menggunakan metode LDA (*Latent Dirichlet Allocation*)”. Penelitian ini menggunakan data *tweet* dari *Twitter* sebagai dataset. Dataset ini dipilih karena *Twitter* merupakan salah satu platform berbasis teks yang paling banyak digunakan oleh banyak user di seluruh dunia. Selain dengan luasnya jangkauan aplikasi *Twitter*, aplikasi ini juga menyediakan fitur *trending topic* sehingga memudahkan peneliti untuk mendapatkan kata kunci topik yang paling banyak dibahas oleh pengguna *Twitter.*

Tujuan dari penelitian ini adalah untuk mengetahui nilai *coherence score* dan *perplexity* yang dihasilkan dari model LDA. *Coherence score* merupakan penilaian untuk mengetahui seberapa mudah model untuk dapat dipahami manusia. Semakin tinggi nilai *coherence score* maka semakin baik. Sementara *perplexity* merupakan kemampuan model untuk menggeneralisasi dokumen setelah memperkirakan model.

2. Metodologi Peneliitian

Penelitian ini akan menggunakan 15816 *tweet* yang dicari menggunakan kata kunci politik mulai dari tanggal 27 Pebruari 2022 sampai 28 Pebruari 2022 dan diperoleh menggunakan API *Twitter*. Pemilihan kata kunci politik dikarenakan jurnal ini ditulis ti tahun 2022 dimana terjadi banyak isu politik seperti perang Rusia dan Ukraina dan G20. Tahapan dari pelaksanaan penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Diagram Alir Pemodelan LDA

Penelitian dimulai dengan (1) *scraping* data menggunakan API Twitter. Data yang diambil merupakan data *tweet* pengguna *Twitter* Indonesia sebanyak 15816 *tweet* dengan kata kunci politik mulai dari tanggal 27 Pebruari 2022 sampai 28 Pebruari 2022 (2) Selanjutnya data *tweet* akan di *preprocessing*. Tahap *preprocessing* mencangkup menghapus *stopword*, username twitter, emoji, dan link menggunakan *regular expression* dan proses *stemming* menggunakan *library* Sastrawi. Setelah data berhasil di preprocessing selanjutnya dilakukan (3) pembuatan model LDA menggunakan *library* PyLDAvis, gensim, dan *skit-learn*. (4) Selanjutnya setelah model berhasil dibuat selanjutnya dilakukan evaluasi model LDA menggunakan *perplexity* dan *coherence score. Perplexity* akan menggunakan sebagian *corpus* sebagai korpus uji pada model yang dihasilkan. Tujuannya adalah untuk mendapatkan nilai probabilitas tiap dokumen terhadap *corpus* yang dirumuskan sebagai berikut.

$$p(w \mid D) = \int p(w \mid \Phi, \alpha) p(\Phi, \alpha \mid D) d\alpha d\Phi \quad (1)$$

Keterangan :

$p(w \mid D)$ = probabilistik tiap dokumen dalam korpus uji

$p(\Phi, \alpha \mid D)$ = probabilistik metode rata rata harmonik untuk tiap dokumen dalam korpus uji

$p(w \mid \Phi, \alpha)$ = estimasi probabilistik metode rata-rata harmonik untuk tiap dokumen

$$p(w \mid \Phi, \alpha) \simeq \frac{1}{\frac{1}{s}\sum_s \frac{1}{p(z, \Phi)}} \quad (2)$$

Setelah nilai probabilistik berhasil didapatkan selanjutnya diambil nilai *perplexity* dari dokumen uji menggunakan rumus.

$$perplexity(D_{uji}) = exp\left\{-\frac{\sum_{d=1}^{M} loglog\, p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (3)$$

Keterangan:

M = Jumlah Dokumen

$p(w_d)$ = Probabilistik tiap kata tiap dokumen d

N_d = Jumlah kata untuk tiap dokumen

D = Korpus

Semakin rendah nilai *perplexity* maka semakin kualitas model semakin baik. Selanjutnya setelah nilai *perplexity* didapat maka dilanjutkan dengan menghitung *coherence score* dari model menggunakan rumus berikut.

$$c(t, W_t) = \sum_{w1,w2 \in Wt} log\, log\, \frac{d(w_1, w_2) + \varepsilon}{d(w_1)} \quad (4)$$

Coherence score akan menilai seberapa baik topik yang dihasilkan oleh model dengan mencari kedekatan antara topik dengan set kata dalam W. Semakin tinggi *coherence score* maka model yang dibuat semakin baik.

3. Hasil dan Pembahasan

Penelitian ini menggunakan dataset 15816 *tweet* yang didapatkan menggunakan metode *scraping* menggunakan API *Twitter.* Setelah data didapat, data akan masuk kedalam proses *preprocessing* yang dilanjutkan dengan pemodelan menggunakan algoritma LDA dan evaluasi untuk mendapatkan model pemodelan topik terbaik untuk dataset. Hasil dan proses dari setiap tahapan dapat dilihat sebagai berikut.

3.1. *Preprocessing* teks

Tahap *preprocessing* merupakan tahap pertama dalam pembuatan model LDA yang mana proses didalamnya mencangkup penghapusan emotikon, penghapusan tautan web, penghapusan username *Twitter*, penghapusan karakter berulang, penghapusan *stopword*, dan *stemming text*. Pada tahap ini data yang sebelumnya berjumlah 15816 *tweet* berkurang menjadi 15745 *tweet*. Hal ini dikarenakan setelah data selesai di *preprocessing* ditemukan ada data yang hanya berisi *stopword* atau setelah di *stemming* tidak ada hasil yang sesuai sehingga data dianggap kosong.

3.2. Pemodelan LDA (*Latent Dirichlet Allocation*)

Pemodelan LDA bertujuan untuk klasterisasi topik yang dari data tweet yang telah di preprocessing. Pada tahap pemodelan LDA ada dua model yang dibuat yakni model LDA dengan *corpus* dasar dan model dengan *corpus tf-idf*. Kedua model ini selanjutnya akan di uji menggunakan metode *hyperparameter tuning* untuk mendapatkan jumlah topik terbaik serta parameter *alpha* dan *beta* terbaik. Parameter *alpha* mewakili kepadatan topik untuk tiap dokumen. Semakin tinggi nilai *alpha* maka tiap dokumen akan terdiri dari lebih banyak topik begitu pula sebaliknya. Sementara parameter *beta* merupakan kepadatan topik untuk tiap kata. Semakin tinggi nilai *beta*, maka topik akan terdiri dari sejumlah besar kata dalam *corpus* dan sebaliknya. Nilai alpha dan beta akan mempengaruhi besar nilai

perplexity dan *coherence score* yang diperoleh dari model. Berikut merupakan besar nilai *perplexity* dan *coherence score* dari model LDA sebelum diimplementasikan metode *hyperparameter tuning*.

Tabel 1. Model LDA Tanpa *Hyperparameter Tuning*

	Model Dasar	Model *TF-IDF*
Perplexity	49,26030441317917	178.77300475609354
Coherence Score	0,37804202528401254	0.3717364224558765

Pada tabel 1, model LDA yang menggunakan *corpus* dasar disebut dengan model dasar dan model LDA yang menggunakan *corpus tf-idf* disebut dengan model *tf-idf*. Terlihat pada tabel bila model LDA yang dibentuk menggunakan Model dasar memiliki nilai *perplexity* yang rendah yang berarti bila model dasar memiliki kemampuan yang lebih baik dalam menggeneralisasi dokumen. Hal ini disebabkan oleh *corpus* yang digunakan. Pada model dasar menggunakan corpus yang mana dibentuk menggunakan logika *bag of words* yang berarti bobot tiap kata dihitung berdasarkan jumlah kemunculannya dalam tiap dokumen. Hal ini tentu baik untuk mendapatkan model dengan *perplexity* yang rendah, namun kelemahan dari model ini adalah tidak ada bobot untuk tiap kata sehingga topik yang dimunculkan kemungkinan tidak merupakan kata yang penting. Sementara penggunaan model *tf-idf* akan memunculkan topik dengan kata yang penting yang dihitung berdasarkan analisis *tf-idf* pada tiap dokumen, namun kelemahannya terletak pada nilai *perplexity*-nya yang tinggi.

Setelah model dasar didapat maka selanjutnya dilakukan analisis parameter paling optimal untuk model menggunakan *hyperparameter tuning*. Parameter model yang menjadi target pada proses ini adalah nilai alpha dan beta pada model. Berikut merupakan lima nilai tertinggi dari proses *hyperparameter tuning* pada model dengan dua *corpus* berbeda.

	num_topics	alpha	eta	coherence_score	perplexity
58	40	0.7	0.7	0.441639	117.585988
35	30	0.1	symmetric	0.398587	51.601062
60	40	symmetric	0.1	0.397622	54.536883
41	30	0.7	0.4	0.396099	91.190851
22	20	0.4	0.7	0.395936	79.354865

Gambar 2. Hasil *Hyperparameter Tuning Corpus* Dasar

	num_topics	alpha	eta	coherence_score	perplexity
53	40	0.4	0.4	0.521576	322.014842
54	40	0.4	0.7	0.443179	368.792438
55	40	0.4	symmetric	0.434121	199.092208
37	30	0.4	0.4	0.431318	270.580092
42	30	0.7	0.7	0.430297	305.804212

Gambar 3. Hasil *Hyperparameter Tuning Corpus* TF-IDF

Berdasarkan gambar 2 dan 3 terlihat kelima hasil *hyperparameter tuning* untuk setiap corpus memiliki perbedaan nilai *perplexity* dari hasil sebelumnya. Pada model *corpus* dasar dapat dilihat bila untuk mendapatkan model terbaik, maka kombinasi yang harus digunakan untuk membentuk model adalah 40 topik, *alpha* sebesar 0,7 dan *beta* sebesar 0,3. Namun dikarenakan nilai *perplexity* yang masih tinggi maka kombinasi yang optimal digunakan adalah kombinasi dengan urutan kedua atau ketiga yakni 30 topik, *alpha* sebesar 0,1, dan *beta symmetric* dan 40 topik, alpha *symmetric,* dan *beta* sebesar 0,1. Hal ini dikarenakan *coherence score* dan perplexity yang dihasilkan tidak memiliki selisih yang jauh dengan model dasar, dan memiliki nilai *coherence* yang lebih baik. Sementara pada model *corpus tf-idf* menghasilkan *coherence score* yang lebih baik namun nilai *perplexity* yang semakin besar. Sehingga

dengan membandingkan dengan model *tf-idf* sebelumnya, maka dipilih kombinasi parameter ketiga yakni 40 topik, *alpha* 0,4 dan *beta symmetric*. Pemilihan kombinasi ini dikarenakan *coherence score* yang didapat lebih tinggi dari *model tf-idf* sebelumnya dan memiliki selisih *perplexity* yang paling kecil dari model sebelumnya.

Berdasarkan kelima model LDA yang telah dihasilkan selanjutnya dipilih satu model LDA yang paling sesuai untuk menyelesaikan kasus pemilihan topik skripsi. Dalam pemilihan topik skripsi perlu diperhatikan topik yang dimunculkan dan interpretasi manusia terhadap topik tersebut sehingga kombinasi yang paling sesuai adalah 40 topik, *alpha* 0,4, *beta symmetric*, dan *corpus tf-idf* dengan *coherence score* sebesar 0,43 dan *perplexity* sebesar 199,09. Hal ini dikarenakan model ini dapat memunculkan *coherence score* yang mendekati 0,5 serta menggunakan *corpus tf-idf* yang memperhatikan bobot tiap kata sehingga sangat sesuai dengan keperluan dalam menggali topik penelitian. Topik yang dihasilkan dari model ini dapat dilihat pada tabel 2.

Tabel 2. Hasil Topik Model LDA

No Topik	Corpus
1	0.090*"tdk" + 0.084*"jangan" + 0.044*"langgar" + 0.043*"krn" + 0.035*"biasa" + 0.030*"kibar" + 0.029*"atur" + 0.027*"bawa2" + 0.025*"serang" + 0.024*"kamu"
2	0.085*"rakyat" + 0.074*"milu" + 0.073*"hak" + 0.073*"alas" + 0.073*"tunda" + 0.071*"elite" + 0.062*"rampas" + 0.030*"lawan" + 0.029*"tak" + 0.027*"amat"
3	0.079*"ga" + 0.064*"terus" + 0.043*"masuk" + 0.035*"tanda" + 0.033*"cara" + 0.029*"stabil" + 0.027*"gulir" + 0.024*"fifa" + 0.024*"tanggap" + 0.023*"meski"
4	0.102*"rusia" + 0.090*"perang" + 0.071*"gak" + 0.042*"kalah" + 0.038*"jadi" + 0.037*"deklarasi" + 0.036*"terima" + 0.035*"bangkang" + 0.029*"merdeka" + 0.029*"chechnya"
5	0.047*"lihat" + 0.038*"ganti" + 0.038*"wayang" + 0.037*"fayhsal" + 0.037*"amira" + 0.034*"konflik" + 0.030*"perlu" + 0.030*"luas" + 0.026*"putus" + 0.026*"timbul"
6	0.125*"penting" + 0.045*"sifat" + 0.041*"calon" + 0.033*"baca" + 0.028*"tinggal" + 0.027*"parpol" + 0.025*"utama" + 0.024*"bp2mi" + 0.023*"rawan" + 0.023*"juang"
7	0.077*"apa" + 0.056*"ukraina" + 0.039*"hari" + 0.037*"ilmu" + 0.036*"bahasa" + 0.033*"amerika" + 0.032*"jadi" + 0.027*"pakar" + 0.026*"dinamika" + 0.024*"mearsheimer"
8	0.069*"kata" + 0.064*"makin" + 0.063*"tinggi" + 0.062*"of" + 0.062*"bagai" + 0.061*"sop" + 0.061*"kes" + 0.061*"cancel" + 0.061*"uni" + 0.061*"majlis"
9	0.064*"elit" + 0.049*"kena" + 0.048*"cuma" + 0.035*"org" + 0.031*"udah" + 0.031*"kat" + 0.028*"kurang" + 0.024*"kau" + 0.020*"posisi" + 0.019*"bkn"
10	0.103*"indonesia" + 0.101*"tolak" + 0.072*"tunda" + 0.070*"ulama" + 0.070*"milu" + 0.068*"sekretaris" + 0.065*"majelis" + 0.065*"jendral" + 0.065*"tambun" + 0.057*"usul"
11	0.067*"amirsyah" + 0.059*"ikut" + 0.036*"aman" + 0.033*"barat" + 0.030*"tum" + 0.029*"pertama" + 0.027*"anggota" + 0.025*"menang" + 0.025*"dewan" + 0.021*"tni"
12	0.157*"cari" + 0.080*"ajak" + 0.080*"sembang" + 0.080*"topik" + 0.080*"tukar" + 0.080*"all" + 0.080*"in" + 0.079*"level" + 0.079*"ideologi" + 0.079*"pasang"
13	0.109*"politik" + 0.088*"kalau" + 0.063*"hukum" + 0.043*"bubarin" + 0.037*"ubah" + 0.030*"ini" + 0.030*"lacur" + 0.028*"naik" + 0.023*"hancur" + 0.023*"bagus"
14	0.065*"kuasa" + 0.064*"dah" + 0.060*"nak" + 0.049*"sekarang" + 0.039*"kan" +

	0.036*"la" + 0.036*"pasal" + 0.035*"luar" + 0.027*"so" + 0.026*"ni"
15	0.046*"dgn" + 0.038*"begini" + 0.033*"ajar" + 0.027*"bela" + 0.025*"selalu" + 0.024*"catat" + 0.024*"ada" + 0.023*"jilat" + 0.023*"khianat" + 0.022*"tahu"
16	0.040*"pihak" + 0.036*"mampu" + 0.036*"erick" + 0.027*"nelayan" + 0.027*"tani" + 0.025*"thohir" + 0.025*"tema" + 0.024*"menteri" + 0.023*"nilai" + 0.023*"milik"
17	0.146*"parah" + 0.146*"just" + 0.144*"masyarakat" + 0.144*"kotor" + 0.135*"tak" + 0.133*"lebih" + 0.012*"politik" + 0.005*"pake" + 0.004*"amandemen" + 0.004*"degree"
18	0.044*"ekonomi" + 0.040*"bagi" + 0.032*"russia" + 0.031*"buah" + 0.030*"pegang" + 0.028*"olahraga" + 0.026*"mulai" + 0.026*"oligarki" + 0.020*"kpd" + 0.020*"ri"
19	0.096*"agama" + 0.057*"daya" + 0.056*"tipu" + 0.045*"guna" + 0.030*"si" + 0.029*"bawa" + 0.024*"bilang" + 0.022*"paling" + 0.020*"hindar" + 0.018*"mungkin"
20	0.129*"aku" + 0.042*"pkb" + 0.035*"antar" + 0.032*"kab" + 0.032*"biar" + 0.030*"hingga" + 0.024*"lima" + 0.023*"siap" + 0.021*"fraksi" + 0.020*"kekal"
21	0.050*"benar" + 0.045*"anak" + 0.045*"kait" + 0.040*"john" + 0.038*"prediksi" + 0.028*"ambil" + 0.020*"politis" + 0.018*"duk" + 0.017*"politik" + 0.017*"gws"
22	0.124*"dukung" + 0.050*"kelompok" + 0.035*"tetap" + 0.034*"identitas" + 0.033*"yusril" + 0.033*"ilegal" + 0.031*"ihza" + 0.024*"bangun" + 0.022*"presiden" + 0.021*"hati"
23	0.048*"pernah" + 0.043*"sangat" + 0.039*"memang" + 0.036*"masalah" + 0.035*"konstitusi" + 0.031*"sepakbola" + 0.028*"sih" + 0.025*"taat" + 0.025*"dlm" + 0.024*"matang"
24	0.495*"politik" + 0.023*"uang" + 0.020*"signifikan" + 0.018*"pragmatis" + 0.018*"tingkat" + 0.017*"transaksional" + 0.016*"gagal" + 0.013*"pro" + 0.011*"nur" + 0.010*"mobil"
25	0.038*"cak" + 0.038*"imin" + 0.036*"punya" + 0.035*"zulhas" + 0.033*"demokrasi" + 0.032*"praktis" + 0.030*"konyol" + 0.028*"ismail" + 0.028*"beda" + 0.028*"campur"
26	0.111*"sama" + 0.106*"mau" + 0.052*"dulu" + 0.046*"nya" + 0.030*"jauh" + 0.025*"ngga" + 0.025*"ngomongnya" + 0.025*"awam" + 0.021*"for" + 0.019*"sebab"
27	0.047*"ni" + 0.039*"kuat" + 0.038*"soal" + 0.029*"awal" + 0.028*"emang" + 0.028*"ingat" + 0.027*"macam" + 0.027*"bicara" + 0.024*"besi" + 0.022*"tunjuk"
28	0.113*"partai" + 0.075*"tunda" + 0.063*"perintah" + 0.054*"hasil" + 0.053*"usaha" + 0.049*"manuver" + 0.048*"koalisi" + 0.047*"operasi" + 0.040*"milu" + 0.039*"bakal"
29	0.063*"nu" + 0.050*"satu" + 0.039*"muda" + 0.038*"salah" + 0.032*"dekat" + 0.030*"menteri" + 0.028*"gus" + 0.028*"jokowi" + 0.024*"dapat" + 0.023*"intelektual"
30	0.066*"laku" + 0.052*"islam" + 0.050*"wacana" + 0.050*"akhir" + 0.043*"anggap" + 0.036*"muhammadiyah" + 0.034*"umat" + 0.030*"jadi" + 0.028*"kepala" + 0.027*"balik"
31	0.083*"aja" + 0.065*"dunia" + 0.061*"banyak" + 0.048*"kalo" + 0.042*"mana" + 0.035*"tp" + 0.029*"bikin" + 0.026*"ranah" + 0.018*"turun" + 0.016*"objek"
32	0.062*"baik" + 0.057*"bumn" + 0.029*"buat" + 0.027*"pikir" + 0.025*"sehat" + 0.024*"diri" + 0.022*"manusia" + 0.022*"uktaina" + 0.022*"efektif" + 0.019*"ubaid"

33	0.082*"semua" + 0.063*"sosial" + 0.057*"survei" + 0.049*"isu" + 0.036*"rilis" + 0.035*"lewat" + 0.029*"tajuk" + 0.024*"gw" + 0.024*"jd" + 0.023*"hasil"
34	0.139*"bukan" + 0.106*"orang" + 0.081*"tuju" + 0.045*"bangsa" + 0.026*"jalan" + 0.025*"konstelasi" + 0.024*"depan" + 0.023*"center" + 0.016*"belakang" + 0.016*"peran"
35	0.069*"pak" + 0.063*"jokowi" + 0.061*"paham" + 0.054*"urus" + 0.039*"dasar" + 0.034*"harus" + 0.033*"penuh" + 0.026*"surat" + 0.026*"tiga" + 0.021*"isi"
36	0.049*"gera" + 0.043*"jabat" + 0.042*"presiden" + 0.042*"masa" + 0.041*"panjang" + 0.039*"tahun" + 0.039*"n" + 0.038*"lalu" + 0.038*"sejak" + 0.038*"pimpin"
37	0.086*"ahli" + 0.059*"utk" + 0.039*"suka" + 0.039*"tokoh" + 0.034*"nama" + 0.031*"umno" + 0.027*"sekaligus" + 0.022*"arifin" + 0.021*"kini" + 0.021*"panigoro"
38	0.071*"pbnu" + 0.057*"ketua" + 0.055*"baru" + 0.051*"pilih" + 0.044*"nyata" + 0.044*"jadi" + 0.039*"besar" + 0.034*"yahya" + 0.033*"harap" + 0.033*"beliau"
39	0.045*"rasa" + 0.040*"parti" + 0.034*"kerja" + 0.029*"pecah" + 0.027*"lama" + 0.025*"negeri" + 0.025*"jelas" + 0.021*"malaysia" + 0.018*"belah" + 0.017*"putin"
40	0.087*"negara" + 0.078*"main" + 0.059*"tu" + 0.039*"suara" + 0.033*"lah" + 0.030*"berobahnya" + 0.029*"siapa" + 0.026*"duduk" + 0.021*"bodoh" + 0.020*"ajoi"

Seperti yang dapat dilihat pada tabel 2, model LDA yang dipilih dapt menghasilkan 40 topik, namun masih diperlukan kemampuan manusia untuk melakukan penalaran terhadap corpus yang dihasilkan sehingga dapat dengan mudah dipahami. Berdasarkan penelitian "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)" karya Kusnanta pada tahun 2017, mendapatkan hasil perplexity terkecil sebesar 213.41, sehingga dapat disimpulkan bila model yang dibuat terbukti lebih baik dari nilai *perplexity* dari penelitian sebelumnya [3]. Sementara pada penelitian "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma *Naive Bayes* dan *Latent Dirichlet Allocation*" karya Merawati tahun 2021 menghasilkan *coherence score* sebesar 0,528, sehingga dapat disimpulkan bila model yang dibuat belum dapat lebih baik dari penelitian sebelumnya berdasarkan aspek *coherence score* [4]. Hal ini dapat disebabkan karena perbedaan corpus yang digunakan dan jumlah data dalam pemodelan topik. Pada penelitian ini jumlah total tweet yang digunakan sebesar 15745 *tweet* sementara pada penelitian Merawati menggunakan 9.496 *tweet*.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan maka dapat disimpulkan bila tanpa *hyperparameter tuning* akan menghasilkan nilai *perplexity* yang lebih kecil dari pada model dengan *hyperparameter tuning.* Namun model dengan *hyperparameter tuning* akan memiliki nilai *coherence score* yang lebih baik dari pada model tanpa *hyperparameter tuning.* Hal ini dikarenakan metode *hyperparameter tuning* hanya mempengaruhi penyebaran topik untuk tiap dokumen dengan mencari setiap kemungkinan *alpha* dan *beta* pada model. Sementara nilai *perplexity* lebih dipengaruhi oleh *corpus* yang digunakan pada model. Semakin rumit *corpus* maka nilai *perplexity* akan semakin tinggi, hal ini terlihat dengan penggunaan dua *corpus* berbeda yakni *corpus* yang dibentuk menggunakan bag of word dan *corpus* yang dibentuk menggunakan *tf-idf*. *Corpus* yang menggunakan *tf-idf* cenderung akan menghasilkan nilai *perplexity* yang lebih tinggi namun *corpus* yang digunakan pula akan menghasilkan topik yang lebih relevan dikarenakan melakukan pembobotan untuk tiap kata. Sehingga untuk membuat model LDA kedepannya perlu memperhatikan tujuan pembuatan model. Pada penelitian ini disimpulkan bila model terbaik adalah model dengan kombinasi 40 topik, *alpha* 0,4, *beta symmetric*, dan *corpus tf-idf* dengan *coherence score* sebesar 0,43 dan *perplexity* sebesar 199,09. Hal ini dikarenakan model ini dapat memunculkan *coherence score* yang mendekati 0,5 serta menggunakan *corpus tf-idf* yang memperhatikan bobot tiap kata sehingga sangat sesuai dengan keperluan dalam menggali topik penelitian.

Referesi

[1] Bela Aprilia, Dhimas Romadhoni AP, Lestari Widyaningsih, dan Chusna Apriyanti, "Analisis Kesulitan Mahasiswa dalam Mengembangkan Ide pada Basic Writing" *Jurnal Penelitian Pendidikan,* vol. 12, no. 1, p. 1669-1719, 2020.

[2] Muhammad Irawan Zain, Radiusman, Muhammad Syazali, Hasnawati, dan Lalu Wira Zain Amrullah, "Identifikasi Kesulitan Mahasiswa dalam Penyusunan Skripsi Prodi PGSD Universitas Mataram" *Jurnal Penelitian Ilmu Pendidikan,* vol. 4, no. 1, p. 73-85, 2021.

[3] Kusnanta Bramantya Putra, I Made dan Renny Pradina Kusumawardani, “Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)” Jurnal Teknik ITS, vol. 6, no. 2, p. 2337-3520, 2017.

[4] Merawati, Ni Luh Putu , Ahmad Zuli Amrullah, dan Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation" *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi),* vol. 5, no. 1, p. 123-131, 2021.