JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer | ISSN: 2615-577X (Online)
https://ejournal.upi.edu/index.php/JATIKOM

Deteksi Topik *Fashion* pada Twitter dengan *Latent Dirichlet Allocation*

Fashion Topic Detection on Twitter with Latent Dirichlet Allocation

Yupa Umigi Al-khairi [#1], Yudi Wibisono [#2], Budi Laksono Putro [#3]

Departemen Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia
Bandung, Indonesia
[1]yupa.umigi@student.upi.edu
[2,3]{yudi,budilp}@upi.edu

Abstrak— **Bagi orang-orang yang bergerak di bidang *fashion* mengetahui tren *fashion* adalah hal yang penting. Salah satu cara untuk mengetahui tren adalah dengan mendeteksi topik mengenai *fashion* yang dibicarakan di media sosial. Penelitian ini mengimplementasikan algoritma Latent Dirichlet Allocation untuk mendeteksi topik *fashion* di Twitter. *Tweet* yang didapat, diklasifikasi dengan metode Naive Bayes lalu dibersihkan dengan cara menghapus URL, simbol, angka dan merubah setiap kata menjadi huruf kecil. *Tweet* lalu dibentuk menjadi kumpulan kata dan dikelompokan dengan algoritma Latent Dirichlet Allocation. Berdasarkan hasil eksperimen, konfigurasi paramater 20 topik dengan 1000 iterasi memperoleh skor UMass terbaik dengan nilai -56.342, dan konfigurasi parameter 50 topik dengan 1000 iterasi memperoleh skor PMI terbaik dengan nilai 6.272.**

***Kata Kunci*:** *topic detection*, *fashion trend*, Twitter, Latent Dirichlet Allocation.

Abstract— **For people within the fashion industry, being up-to-date with fashion trends is an important matter. One method of finding these trends is to detect fashion-related topics within social media. This research implements Latent Dirichlet Allocation algorithm to detect fashion topics in Twitter. Procured tweets are then classified with Naive Bayes method by erasing URL, symbols, numbers and changing every word into lowercase. Tweets are then made into compilation of words and then grouped with Latent Dirichlet Allocation algorithm. According to the experiment results, configuration with the parameters of 20 topics and 1000 iterations obtains the best UMass score of -56,432, and configuration with the parameters of 50 topics and 1000 iteration obtains the best PMI score of 6,272.**

***Keywords*:** Topic Detection, Fashion Trend, Twitter, Latent Dirichlet Allocation.

I. PENDAHULUAN

Dunia pakaian merupakan bidang yang berkembang dengan cepat, setiap musim bahkan setiap bulan selalu muncul produk baru dari berbagai produsen [1]. Konsumen perlu mengetahui perkembangan *fashion* agar penampilanya tidak ketingalan zaman. Para produsen dan distributor perlu memahami tren sebagai sarana untuk mencari inspirasi dalam pembuatan produk selanjutnya dan dasar strategi untuk mengetahui selera pasar saat ini, sehingga dapat memaksimalkan keuntungan.

Pendeteksian topik dilakukan dengan menelompokan kata pada *tweet* yang membahas *fashion* menggunakan algoritma Latent Dirichlet Allocation. Data *Tweet* yang didapat akan dibentuk menjadi kumpulan kata dan dikelompokan berdasarkan probabilitas kata tersebut termasuk ke dalam topik mana.

II. PENELITIAN TERKAIT

Salah satu cara untuk mengetahui tren *fashion* adalah dengan mengakses Google Trends pada www.google.com/trends atau mengkases aplikasi berbayar WSGN pada www.wgsn.com. Terdapat penelitian yang menganalisis aktivitas konsumen di situs jual beli [2], dengan menganalisis kata-kata pada *fashion* blog dan situs berita *fashion* menggunakan metode Neuro-Linguistic [3], atau dengan memanfaatkan data dari sosial media, contohnya menghitung jumlah *tweet* yang berkaitan dengan *fashion* [4].

Selain cara tersebut, *fashion trend* dapat diidentifikasi dengan melakukan *Topic Detection* pada data dari media sosial yang membahas mengenai *fashion.* Penelitian mengenai *Topic Detection* pada media sosial sudah pernah dilakukan sebelumnya. Penelitian [5] meneliti cara kerja aplikasi TwitterMonitor. Aplikasi tersebut medeteksi topik dengan mencacah isi dari *Twit* yang diambil menjadi sekumpulan kata kunci, kata kunci yang jumlahnya meningkat drastis dalam waktu yang singkat akan dipisahkan dan akan dimaknai dengan menggunakan algoritma ekstraksi konten. Pada penelitian [6], jumlah lonjakan kata kunci dideteksi dengan cara menghitung

frekuensi kemunculan suatu kata kunci dengan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan juga algoritma n-gram. Lalu pada penelitian [7], topik dideteksi dengan menghitung probabilitas kemunculan bersama suatu kata kunci dengan metode Jensen-Shannon Divergence dan mengelompokan kata kunci tersebut dengan algoritma K-Means.

III. LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation atau LDA merupakan algoritma untuk mendeteksi topik melalui pemodelan probabilistik dalam sekumpulan data [8].

LDA mengasumsikan setiap dokumen memiliki topik dan dibentuk dari kata-kata yang berkaitan dengan topik tersebut [9], sehingga suatu dokumen dapat direpresentasikan sebagai campuran dari topik topik tersembunyi dengan proporsi yang berbeda beda [10].

Pada penelitian ini, teknik Gibss Sampling digunakan dalam pengaplikasian LDA [11]. Tahapannya adalah sebagi berikut :

1. Tentukan jumlah topik dan jumlah iterasi.
2. Untuk setiap kata yang ada dalam suatu *tweet*, masukan kata tersebut ke dalam suatu topik secara acak.
3. Pilih satu *tweet*
4. Pilih satu kata dalam *tweet*
5. Hitung nilai probabilitas kata tersebut terhadap setiap topik yang ada dengan menggunakan persamaan (1).
6. Masukan kata ke dalam topik yang memiliki nilai tertinggi.
7. Lakukan tahap 5-6 untuk setiap kata dalam *tweet* hingga seluruh *tweet* telah terproses.
8. Ulangi tahap 4-8 sebanyak iterasi yang ditentukan

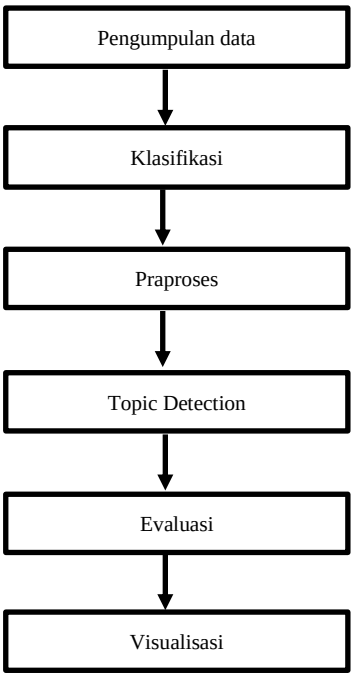
Dengan menggunakan persamaan (1), probabilitas suatu dokumen masuk ke dalam topik apa dihitung dengan melihat jumlah topik pada suatu dokumen. Probabilitas suatu kata termasuk topik apa, dihitung dengan melihat jumlah kata pada suatu topik.

$$P(j|\, w_i, d_i) = \frac{C^{\,WT}_{w_i j}}{\sum_{w=1}^{W} C^{\,WT}_{w j}} \cdot \frac{C^{\,DT}_{d_i j}}{\sum_{t=1}^{T} C^{\,DT}_{d_i t}} \qquad (1)$$

j = topik yang sedang dikalkulasi
w_i = kata yang sedang dikalkulasi
d_i = dokumen yang sedang dikalkulasi
$C^{\,WT}$ = matriks jumlah kata dalam suatu dokumen
$C^{\,DT}$ = matriks jumlah topik dalam suatu dokumen
$C^{\,WT}_{w_i j}$ = Jumlah kata w_i dalam topik j
$C^{\,WT}_{w j}$ = Jumlah kata w dalam topik j
$C^{\,DT}_{d_i j}$ = Jumlah kata dalam d_i termasuk topik j
$C^{\,DT}_{d_i t}$ = Jumlah kata dalam d_i termasuk topik t

IV. Skenario Eksperimen

Gambar 1 menjelaskan tahapan pada penelitain ini. Tahapan pertama adalah pengumpulan data, data merupakan data *tweet* yang didapat dari Twitter.



Gambar. 1 Tahapan Eksperimen

Tweet yang dikumpulkan akan diklasifikasi untuk memisahkan antara *tweet* yang relevan untuk diolah dan yang tidak relevan untuk diolah. *Tweet* yang relevan akan dibersihkan pada tahap praproses. *Topic Detection* akan digunakan untuk memproses *tweet* yang telah dibersihkan dan hasilnya akan dievaluasi.

Selanjutnya hasil dari tahap *Topic Detection* akan ditampilkan dalam bentuk grafik sehingga mempermudah pengguna untuk memahami suatu topik [12].

V. HASIL

Eksprimen dilakukan dengan menggunakan 73.074 *tweet* mengenai *fashion* yang diambil dari tanggal 27 Agustus 2017 sampai 9 September 2017. *Tweet* diklasifikasi dengan algoritma Naive Bayes menggunakan 5000 tweet yang telah dilabeli secara manual dan dibagi menjadi 80% data latih dan 20%. Nilai akurasi klasifikasi adalah sebesar 87%.

Tweet yang relevan akan dibersihkan pada tahap praproses dengan menghapus URL, simbol, angka dan merubah setiap kata menjadi huruf kecil. Contoh *tweet* hasil praproses dapat dilihat pada tabel I.

Korespondensi: *yupa.umigi@student.upi.edu*

TABEL I
PRAPROSES TWEET

Tweet asal		*Tweet* setelah praproses
@GCDSwears 2017 fall/winter collection is now available	gcdswears fall winter collection is now available	

Setelah tahap praproses, *tweet* akan diolah dengan menggunakan metode LDA. Tabel II menunjukan contoh topik yang dihasilkan. Topik yang baik dapat dilihat dari kata-kata yang mudah dimaknai dan memiliki korelasi yang baik satu sama lain.. Sedangkan topik yang buruk memiliki kata-kata yang sulit dimaknai dan tidak membahas suatu tema yang spesifik.

TABEL I
CONTOH TOPIK YANG DIHASILKAN

Term	Deskripsi
nike = 172; air = 128; max = 87; look = 76; virgil = 64; abloh = 64; white = 59; collab = 56;	Topik ini membahas kolaborasi brand Nike dan brand Off-White milik Virgil Abloh

Dengan menggunakan evaluasi nilai *coherence* Umass dan PMI, keterkaitan antar kata akan dinilai. Jika kata-kata pada suatu topik sering mengalami kemunculan bersama maka topik bersebut memiliki nilai *coherence* yang besar. Untuk skor Umass, nilai yang lebih baik adalah nilai yang mendekati 0. Sedangkan untuk skor PMI, nilai yang lebih baik adalah nilai yang lebih besar.

Tabel III menampilkan skenario eksperimen yang dilakukan. Eksperimen dijalankan berdasarkan dua parameter, yaitu jumlah topik dan jumlah iterasi. Untuk pemilihan jumlah topik, diasumsikan topik yang dibicarakan tidak akan kurang dari 5 topik dan tidak akan lebih dari 50 topik maka digunakan 5,10,20 dan 50 topik. Sedangkan untuk jumlah iterasi, digunakan 100,500, dan 1000 iterasi.

TABEL II
SKENARIO EKSPERIMEN

No	Topik	Iterasi
1	5 Topik	100 Iterasi
2		500 Iterasi
3		1000 Iterasi
4	10 Topik	100 Iterasi
5		500 Iterasi
6		1000 Iterasi
7	20 Topik	100 Iterasi
8		500 Iterasi
9		1000 Iterasi
10	50 Topik	100 Iterasi
11		500 Iterasi
12		1000 Iterasi

Gambar 2 menampilkan nilai evaluasi skenario 5 topik untuk setiap iterasi. Pada 100 iterasi, skor evaluasi bernilai -61.30 dan – 6.99, dan skor meningkat pada jumlah iterasi yang lebih besar. Pada 500 iterasi, skor meningkat menjadi -60.87 dan -6.82 dan skor terbesar di skenario ini terdapat pada 1000 iterasi dengan -60.05 dan -4.72.

TABEL III
CONTOH HASIL PADA SKENARIO 5 TOPIK

Term	Deskripsi
nike = 545; fall = 356; air = 349; lt = 282; winter = 260; look = 234; max = 215; adidas = 197;	Topik sulit dimaknai karena kata tidak memiliki kolerasi yang kuat

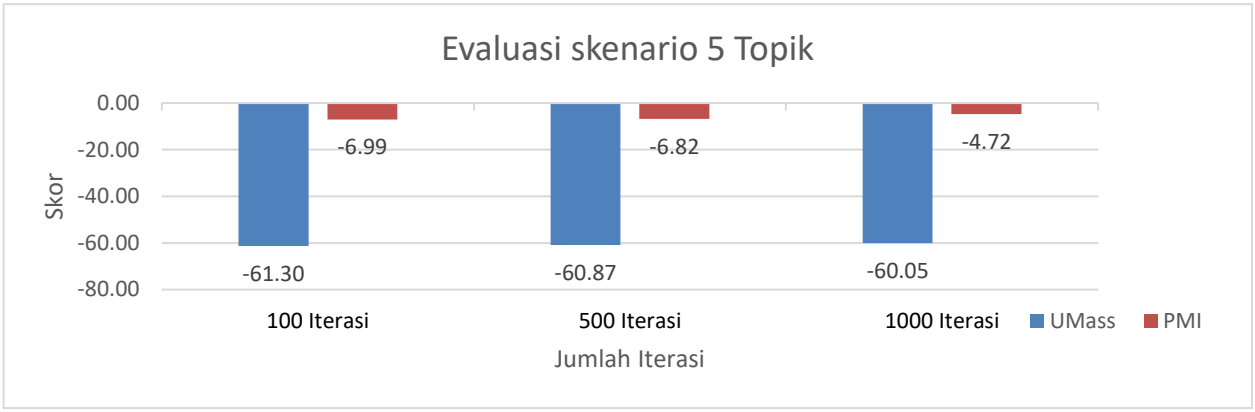
Tabel IV menampilkan contoh topik yang dihasilkan pada skenario 5 topik denga 1000 iterasi. Kata-kata yang membentuk topik pada skenario ini tidak menghasilkan informasi yang spesifik sehingga topik pun sulit dimaknai dikarenakan korelasi antar kata yang membentuk tidak baik.

Gambar 3 menunjukan bahwa nilai terbaik pada skenario 10 topik merupakan hasil dengan iterasi terbanyak, yaitu 1000 iterasi dengan nilai -59.20 untuk skor Umass dan nilai 1.20 untuk skor PMI, sedangkan skor terburuk merupakan hasil dengan iterasi paling sedikit yaitu 100 iterasi dengan nilai -59.94 dan -2.81.

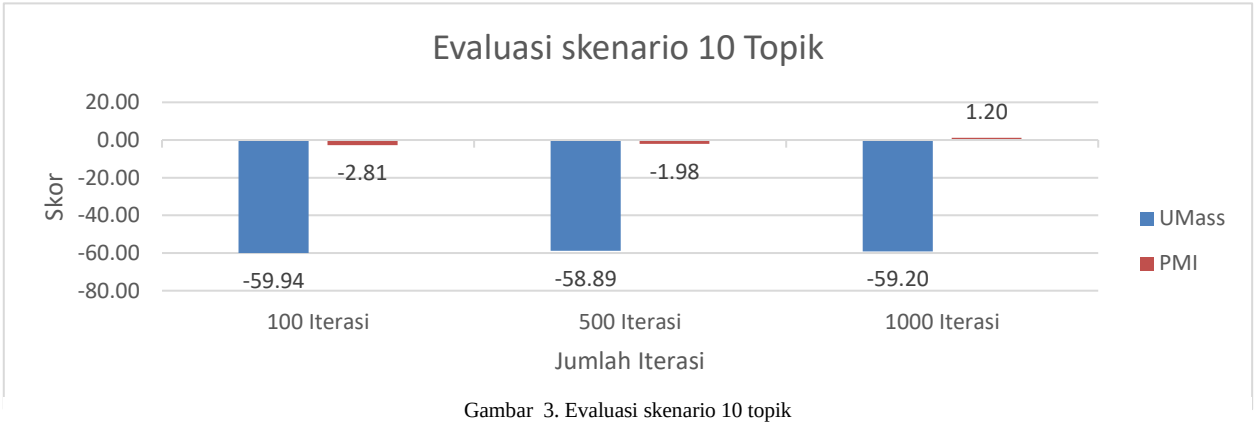
Tabel V menampilkan contoh topik yang dihasilkan pada skenario 10 topik dengan 1000 iterasi. Terdapat 1 topik yang dapat dimaknai yaitu membahas kolaborasi antara *brand* Nike dan Supreme, sedangkan topik lain yang dihasilkan masih sulit untuk dimaknai karena korelasi antar kata yang masih tidak baik.

TABEL IV
CONTOH HASIL PADA SKENARIO 10 TOPIK

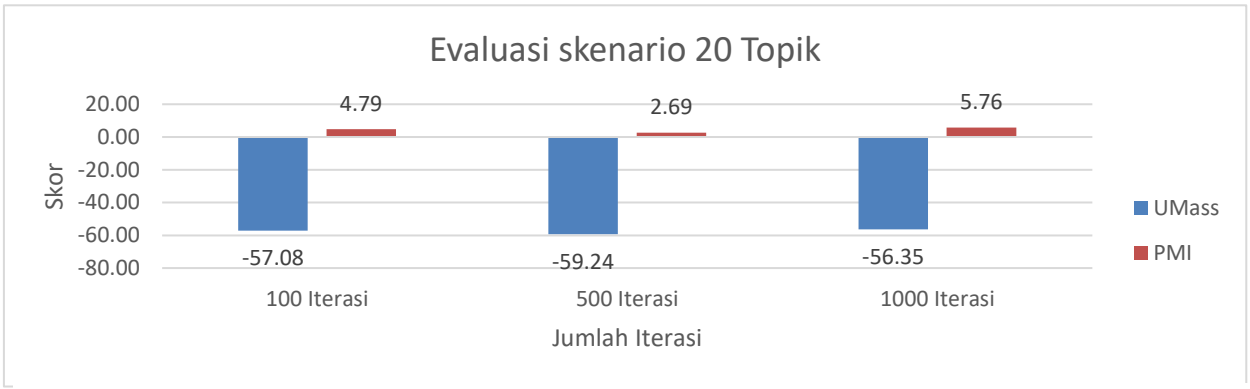
Term	Deskripsi
nike = 190; air = 134; first = 134; supreme = 125; look = 123; like = 123; lt = 115; video = 104; us = 102;	Topik ini membahas tampilan pertama kolaborasi antara *brand* Nike dan Supreme



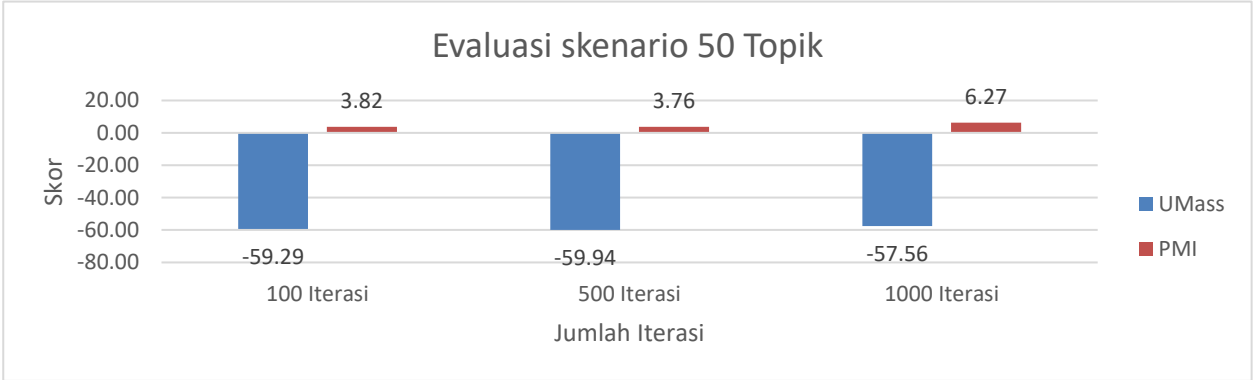
Gambar 2. Evaluasi skenario 5 topik



Gambar 3. Evaluasi skenario 10 topik



Gambar 4. Evaluasi skenario 20 topik



Gambar 5. Evaluasi skenario 50 topik

Gambar 4 menujukan hasil evaluasi skenario 20 topik untuk setiap iterasi. Jumlah iterasi terkecil pada skenario ini tidak memiliki skor paling buruk, malah skor paling buruk didapat pada 500 iterasi dengan nilai -59.24 dan 2.69. Skor terbaik tetap diperoleh oleh jumlah iterasi terbesar yaitu 1000 iterasi dengan nilai -56.35 dan 5.76. Dibandingkan dengan ke-2 skenario sebelumnya, skenario dengan 20 topik memiliki skor PMI yang lebih baik karena skor PMI untuk setiap iterasi memiliki nilai positif.

TABEL V
CONTOH HASIL PADA SKENARIO 20 TOPIK

Term	Deskripsi
nike = 172; air = 128; max = 87; look = 76; virgil = 64; abloh = 64; white = 59; collab = 56;	Topik ini membahas kolaborasi *brand* Nike dan Off-White

Tabel VI menampilkan contoh topik yang dihasilkan pada skenario 20 topik dengan 1000 iterasi. Beberapa topik yang dihasilkan pada skenario ini sudah memiliki korelasi antar kata yang baik sehingga topik menghasilkan informasi yang spesifik dan mudah dimaknai. Contohnya terdapat topik yang membahas kolaborasi antara *brand* Nike dan Off-White.

Selanjutnya gambar 5 menunjukan hasil evaluasi skenario 50 topik untuk setiap iterasi. Jumlah iterasi terbesar yaitu 1000 iterasi memiliki skor terbaik dengan nilai -57.56 untuk skor Umass dan 6.27 untuk skor PMI. Skenario 50 topik dengan 1000 iterasi juga memiliki skor PMI terbesar dibandingkan skenario topik lain.

TABEL VI
CONTOH HASIL PADA SKENARIO 50 TOPIK

Term	Deskripsi
adidas=47;amp=43;boost=37;collaboration=34; yeezy=31;adidasfootball=31;introducing=30; lt=30;first = 28;	Topik ini membahas Adidas Yeezy Boost

Tabel VII menampilkan contoh topik yang dihasilkan pada skenario 50 topik dengan 1000 iterasi. Sama seperti skenario 20 topik, skenario 50 topik dapat menghasilkan beberapa topik yang mudah dimaknai dan memberi informasi yang spesifik.

VI. PEMBAHASAN HASIL

Dari hasil eksperimen, dapat diketahui bahwa semakin banyak jumlah iterasi yang dilakukan maka hasil akan semakin baik. Iterasi terbanyak yaitu 1000 iterasi memiliki nilai *coherence* yang paling besar.

Untuk menentukan jumah topik yang optimal, kedua metode evaluasi yang dilakukan memiliki hasil yang berbeda. Pada metode Umass dapat dilihat pada gambar 3, jumlah topik yang paling optimal adalah 20 topik dengan nilai -56.342, sedangkan untuk metode PMI, dapat dilihat pada gambar 4 jumlah topik yang paling optimal adalah 50 topik dengan nilai 6.272.

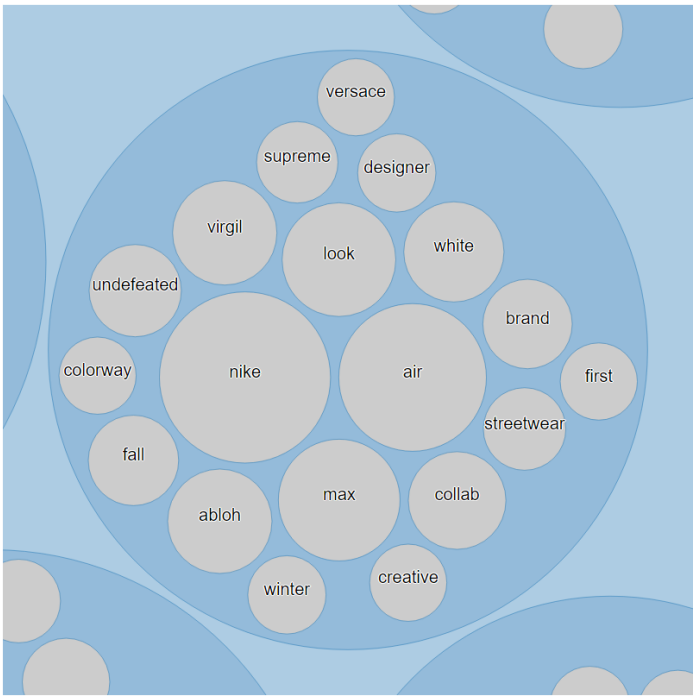
Jika melihat komposisi topik yang dihasilkan dari dua skenario terbaik, meskipun masih tidak dapat mengidentifikasi seluruh topik secara sempurna, kedua konfigurasi menghasilkan topik yang cukup baik. Contoh topik yang baik dapat dilihat pada tabel VIII, kata-kata dalam suatu topik akan menghasilkan tema yang spesifik ketika dilakukan pencarian pada Google. Untuk topik satu, topik yang dihasilkan membahas berita tentang desainer bernama Zuhair Murad. Untuk topik 2, topik yang dihasilkan membahas pakaian tenis dari Adidas yang didesain oleh Pharrell. Sedangkan topik 3, topik ini membahas mengenai rilisan baru dari Adidas Football yang bekerja sama dengan Carlos Soler.

TABEL VII
CONTOH TOPIK YANG BAIK

No	Term	Deskripsi
1	red = 36;dior = 36;fall = 32;favorite = 31;vmas = 30;brand = 26;right = 24;love = 23;winter = 22;carpet = 21	Topik ini membahas brand Dior di ajang Red Carpet VMAS
2	adidas = 45;style = 29;shop = 28;like = 28;sneakers = 27;designer = 26;right = 26;tennis = 23;pharrell = 22	Topik ini membahas pakaian tenis dari Adidas yang didesain oleh Pharrell
3	collaboration = 97; adidasfootball = 90; supreme = 88; amp = 82; created = 76; look = 76; carlossoler = 75; introducing = 75; ace = 75; heretocreate = 74;	Topik ini membahas rilisan Adidas Football

Topik yang dihasilkan selanjutnya akan ditampilkan dalam bentuk *Circle Packing*, agar topik dapat lebih mudah dimaknai. Gambar 6 menampilkan isi suatu topik, yaitu kumpulan kata-kata dalam bentuk lingkaran. Ukuran lingkaran menunjukan frekeunsi suatu kata dalam topik tersebut. Semakin besar lingkaran maka semakin sering kata tersebut muncul dalam suatu topik.

Korespondensi: *yupa.umigi@student.upi.edu*



Gambar 6. Visualisasi Topik

VII. KESIMPULAN

Metode LDA dapat menghasilkan topik mengenai *fashion* yang sedang dibicarakan di media sosial Twitter. Meskipun tidak semua topik yang dihasilkan memberikan hasil yang diinginkan, terdapat beberapa topik dengan hasil yang baik sehingga *item* atau *brand* dapat diidentifikasi dengan mudah.

Untuk penelitian berikutnya, pengambilan data dapat dilakukan dengan rentang waktu yang lebih panjang sehingga data yang diperoleh lebih banyak dan variatif. Eksperimen dapat dilakukan dengan skenario yang lebih banyak agar analisa yang dilakukan memiliki hasil yang lebih baik.

Penelitian ini tidak hanya terbatas untuk diterapkan di bidang *fashion*, metode ini dapat diterapkan di bidang lainnya contohnya untuk mencari topik di bidang politik yang sedang ramai dibicarakan di media sosial. Penelitian ini juga dapat dikembangkan dengan menambahkan fitur untuk melakukan *Topic Tracking* atau fitur untuk mengikuti perkembangan topik tersebut.

REFERENSI

[1] T. Hines, *Fashion Marketing: Contemporary Issues*. 2007.

[2] R. Sanchis-Ojeda, D. Sibley, and P. Massimi, “Detection of fashion trends and seasonal cycles through the analysis of implicit and explicit client feedback,” 2016.

[3] S. Beheshti-kashi, M. Lütjen, L. Stoever, and K. Thoben, “TrendFashion - A Framework for the Identification of Fashion Trends,” 2015.

[4] S. Beheshti-kashi, “Twitter and Fashion Forecasting : An Exploration of Tweets regarding Trend Identification for Fashion Forecasting,” 2015.

[5] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,” *SIGMOD ’10 Proc. 2010 ACM SIGMOD Int. Conf. Manag. data*, pp. 1155–1158, 2010.

[6] L. M. Aiello *et al.*, “Sensing trending topics in Twitter,” no. c, 2013.

[7] C. Wartena and R. Brussee, “Topic detection by clustering keywords,” *Belgian/Netherlands Artif. Intell. Conf.*, pp. 379–380, 2008.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” vol. 3, pp. 993–1022, 2003.

[9] R. Krestel, P. Fankhauser, and W. Nejdl, “Latent Dirichlet Allocation for Tag Recommendation,” 2009.

[10] J. Mazarura and A. De Waal, “A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text,” pp. 1–6, 2016.

[11] M. Steyvers and T. Griffiths, “Probabilistic Topic Models,” 2012.

[12] J. Allan, A. V Leouski, and R. C. Swan, “Interactive CLuster Visualizaton for Information Retrieval.” 1997.