Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 1, No. 1 (2019), pp. 17-31
ISSN 2686-5335
doi: 10.15408/inprime.v1i1.12726

Sequential Topic Modelling: A Case Study on Indonesian LGBT Conversation on Twitter

Arsy Arlina and Muhaza Liebenlito
Department of Mathematics, Faculty of Sciences and Technology
Universitas Islam Negeri Syarif Hidayatullah Jakarta
Email: arsy.arlina14@mhs.uinjkt.ac.id, muhazaliebenlito@uinjkt.ac.id

Abstract

As a country with the largest Muslim population in the world, the Lesbian, Gay, Bisexual, and Transgender (LGBT) issue in Indonesia has always been a hot topic to investigate. Social media such as Twitter is normally the main media where people normally discuss this LGBT topic. In this paper, we collect 18,552 tweets dated from 2015 up to 2018 to analyze the dynamics of the LGBT conversation among Indonesian peoples. In this research, we will explore the main topic of the LGBT conversation using Linear Discriminant Analysis (LDA). LDA is one of the most popular methods of soft clustering. This technique is effective to identify latent topic information (hidden) in a collection of big data using bag of words approaches that treat every document as a vector of total words and is represented as a probability distribution on several topics. The result shows that there are seven main categories that people normally talked about regarding LGBT i.e. politics, religion, government, ethics, nationality, culture, and technology. Looking at the topic probability distributions on each semester we found that it is generally homogenous. An exception occurs during the government election period where politic tends to have a significantly higher probability. In other words, we have found that there is a tendency that the LGBT issues are used in Indonesian politics.
Keywords: LGBT; politics; topic modeling; twitter.

Abstrak

Sebagai negara dengan penduduk muslim terbesar di dunia, isu mengenai *Lesbian*, *Gay*, *Bisexual*, dan *Transgender* (LGBT) di Indonesia adalah isu sensitif yang senantiasa menarik untuk diteliti. Media sosial seperti twitter adalah salah satu media yang biasa digunakan masyarakat untuk mendiskusikan tentang topik LGBT ini. Penelitian ini menggunakan 18.552 tweet tahun 2015 – 2018 dikumpulkan untuk melihat perbedaan pola perbincangan dari waktu ke waktu. Dalam penelitian ini, eksplorasi topik utama perbincangan LGBT dianalisis menggunakan metode Linear Discriminant Analysis (LDA). LDA adalah metode yang paling populer dalam *soft clustering*. Teknik ini efektif untuk mengidentifikasi informasi topik laten (tersembunyi) dalam koleksi dokumen besar menggunakan pendekatan *bag of words* yang memperlakukan setiap dokumen sebagai vektor jumlah kata dan direpresentasikan sebagai distribusi probabilitas atas beberapa topik, sementara setiap topik direpresentasikan sebagai distribusi probabilitas atas sejumlah kata. Hasil menunjukkan bahwa terdapat tujuh topik dominan yang sering muncul pada perbincangan tentang LGBT, yaitu politik, agama, pemerintahan, keasusilaan, kewarganegaraan, budaya dan teknologi. Pada kategori ini kemudian distribusi probabilitas topik dihitung dan dianalisa pada setiap semesternya. Hasilnya menunjukkan bahwa ada kecenderungan distribusi topik seragam, kecuali pada masa-masa pergantian pemerintahan dimana kategori politik cenderung meningkat secara signifikan. Dengan kata lain, ada kecenderungan bahwa isu LGBT dikaitkan dengan kehidupan perpolitikan di Indonesia.
Kata kunci: LGBT; politik; topic modelling; twitter.

1. INTRODUCTION

Lesbian, Gay, Bisexual, and *Transgender* (LGBT) are one of sexual disorders. To get the exact number of individuals with these disorientations is rather difficult to obtain from civil registry because its existence is not acknowledged by many countries particularly in Indonesia. According to Provincial Health Office records and SIHA (HIV/AIDS Information System) regarding the number of people infected with HIV/AIDS due to several factors such as bisexual, homosexual, and men who have sex with men (MSM) which is a form of LGBT as shown in Figure 1, shows that the number of LGB has increased every year.

LGBT defenders do various way to get a strong defense of their existence, as published in the study of Pan et al [2] who said that news media were used as a means for LGBT defenders so that LGBT can get their legal recognition. In this era globalization, news is no longer difficult to obtain. This is because of the existence of communication media with wider coverage and easily to be accessed like social media. One of the most popular social media in Indonesia is twitter. It was shown that twitter's users in Indonesia in July 2009 had reached 41 million users and continued to experience rapid increase [3]. The increase in the number of LGBT people and twitter's users raises the question of how the development of the number of tweets about LGBT on twitter. As can be seen in Figure 2, the development in terms of the number of tweets about LGBT has increased. With the number of tweets about LGBT that keep increasing, this will make us difficult to understand the conversation topics and require a long time, because that topics that emerge will contain each other. In order not to be done manually, objectively, and can be done more quickly, people can use machines to understand a collection of tweets by uniting the words, topics, and contexts at the same time.

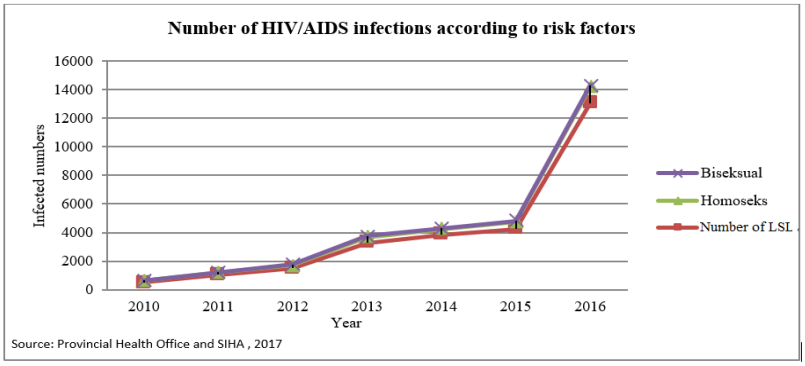

Figure 1. Number of HIV/AIDS infections according to risk factors

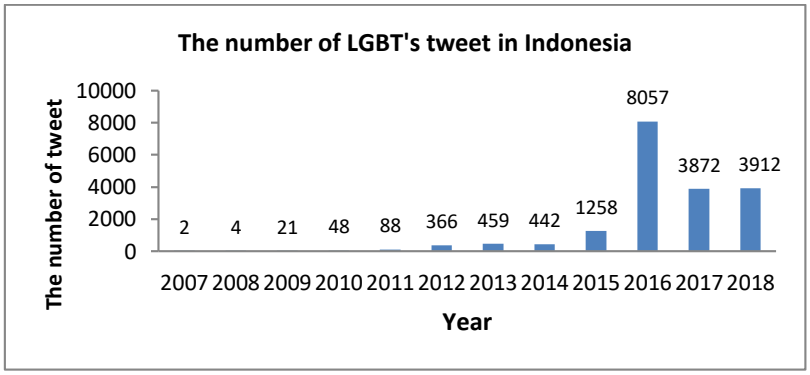

Figure 2. The number of tweets related to LGBT.

Topic modelling is an algorithm to find the main theme which includes a collection of big and unstructured documents, and also can arrange the collection of that documents according to the themes found [7]. This algorithm can be adapted to many types of data such as genetic data, images, and social networks [7]. One form of topic modelling algorithm is called Latent Semantic Analysis (LSA) which approaches to automatic indexing and information retrieval which tries to overcome the problem of differences in using words by mapping the documents and terms of representation [8] [9] [10], then a new approaches is performed to strengthen the statistical foundation and to perform factor analysis called Probabilistic Latent Semantic Analysis (PLSA) [11] [12] [13]. From these two algorithms, it was found a problem that if there is a weighting matrix for each word with negative value, then a method is developed called Non-Negative Matrix Factorization (NMF) [14] [15]. According to previous researchers, LDA is effective in topic modelling, easy to understand, and there are even many application developments from LDA [4] [5] [6]. This research will conduct an understanding of Twitter social media on topics related to LGBT in Indonesia using LDA.

2. METHOD

2.1. Data

The data used in this study is obtained from secondary data of tweets in Bahasa sent each day from January 1st, 2015 to June 30th, 2018 with keyword "LGBT" because this keyword is very popular in the discussion related to lesbian, gay, bisexual, and transgender. In the storage process, the pre-processing data process is included to clean the tweets from abbreviations, symbols, and irregular capital letters. The final data consists of 10,562 lines of tweets containing several information as shown in Table 1.

Table 1. Data obtained from scraping and preprocessing in CSV.

Time	Username	Replies	Retweets	Likes	Language	Tweet	Cleaned_ Tweet
27-Jan-15	@pancuniyoldas	1	3	12		#Ã§are #LGBT Bakanligi	care lgbt bakanligi
25-Jan-15	@yourPlainJean	1	4	2	Indonesian	@twt_LGBT punyalah susah trans nak dapat rights pastu citer bodoh macam ni bagi pulak lulus. pic.*twitter*.com/SOi0C0NzKR	lgbt punyalah susah trans pastu citer bodoh lulus soi0c0nzkr
25-Jan-15	@BACK2STONEWALL	0	2	3	Indonesian	Barbaric Anti-Gay Saudi King Abdullah DEAD http://www.back2stonewall.com/2015/01/barbaric-anti-gay-saudi-king-abdullah-dead.htmlÂ â€¦ #p2 #WorldNews #lgbt #gay	barbaric anti gay saudi king abdullah mati worldnews lgbt gay
24-Jan-15	@prosadio	3	8	26	Indonesian	Wala kaming pake sa sinasabi niyo na pagiging LGBT IS A SIN, Kasi ganito talaga kami simula nung pinanganak kami. Sinasayang mo lang oras mo	wala kaming sinasabi niyo pagiging lgbt sin kasi ganito talaga simula nung pinanganak sinasayang lang oras

The collected data from 2015-2018 is divided into semester (every 6 months) to be observed. The numbers of the recorded tweets are varying over the seven semesters as displayed in Figure 3, with the highest peak during the first semester of 2016.

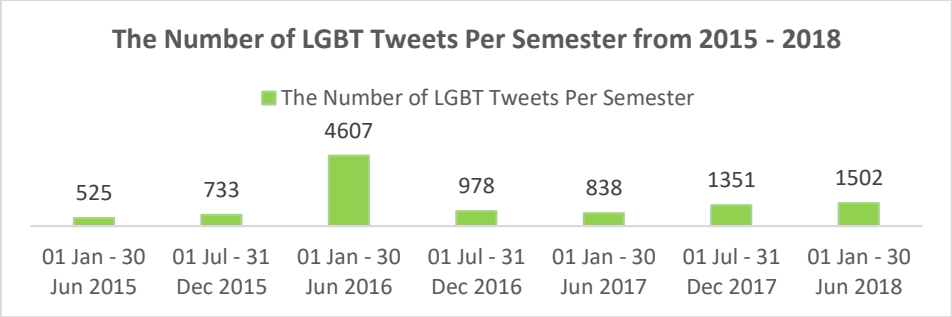

Figure 3. The number of LGBT tweets per semester from 2015-2018.

There are several interesting words in the data, they are one word with two meanings such as "*dr*" that can refer to "*dari*" or "*dokter*"; "*u*" that can refer to "*untuk*" as well as "*anda*", and also "*tk*" that can refer to "*taman kanak-kanak*" as well as "*tidak*". Since the number of words with multiple meanings is not much, then to overcome this problem is by changing them manually adjusted to the context of the sentences in the tweet. Furthermore, this data also has problem in terms of language used; some tweets contain foreign language or mixed languages such as Malay, Spanish, and Filipino languages. According to the recorded data, several information related to LGBT, as presented in Table 2, shows that the number of responses from users in one tweet about LGBT reached the maximum number of 938 responses with average number of 5.82 out of 61,422 responses. If we observe a tweet that is redistributed by other users related to LGBT by looking at the numbers of retweets, indicating that those tweets have multiple users who agree, with the highest number of retweets is 17,077 and average of 50.14 out of 529,528 retweets. The number of twitter users who like one tweet about LGBT is 10,327 with average of 46.65 out of 492,720 number of likes. This suggests that public's responses and judgments about LGBT in social media such as twitter is still low.

Table 2. Descriptive statistics of LGBT tweets

	N	Minimum	Maximum	Sum	Mean
Replies	10562	0	938	61422	5.82
Re*tweet*	10562	0	17077	529528	50.14
Likes	10562	0	10327	492720	46.65
Valid N (listwise)	10562				

2.2. Research Stages

The stages of this study include data preparation, model development, until evaluation as well as interpretation of research results so that useful insights can be concluded from this study. The research stages can be seen in Figure 4 and 5.

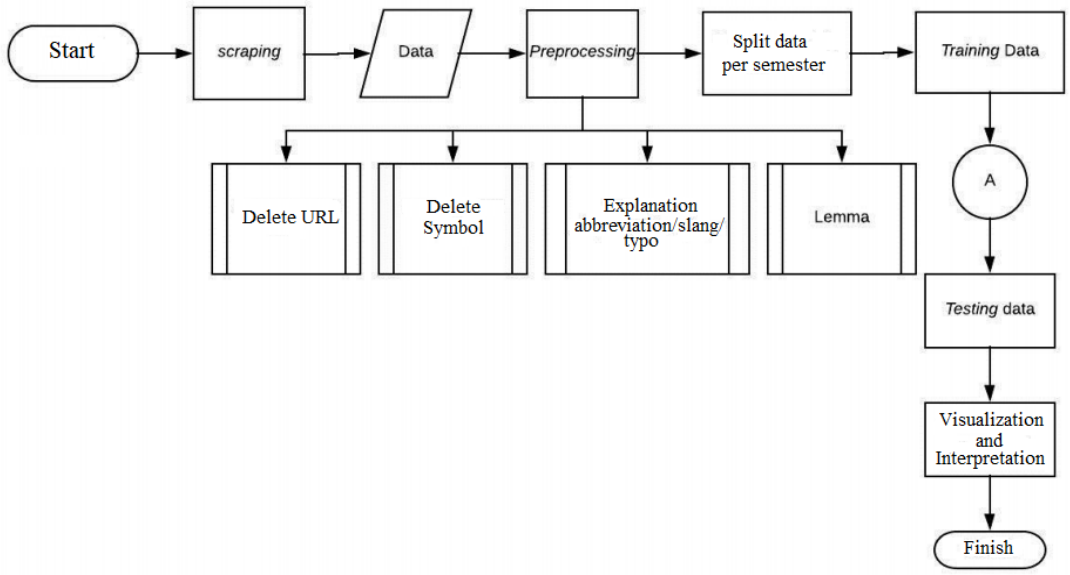

Figure 4. Flowchart of learning classification.

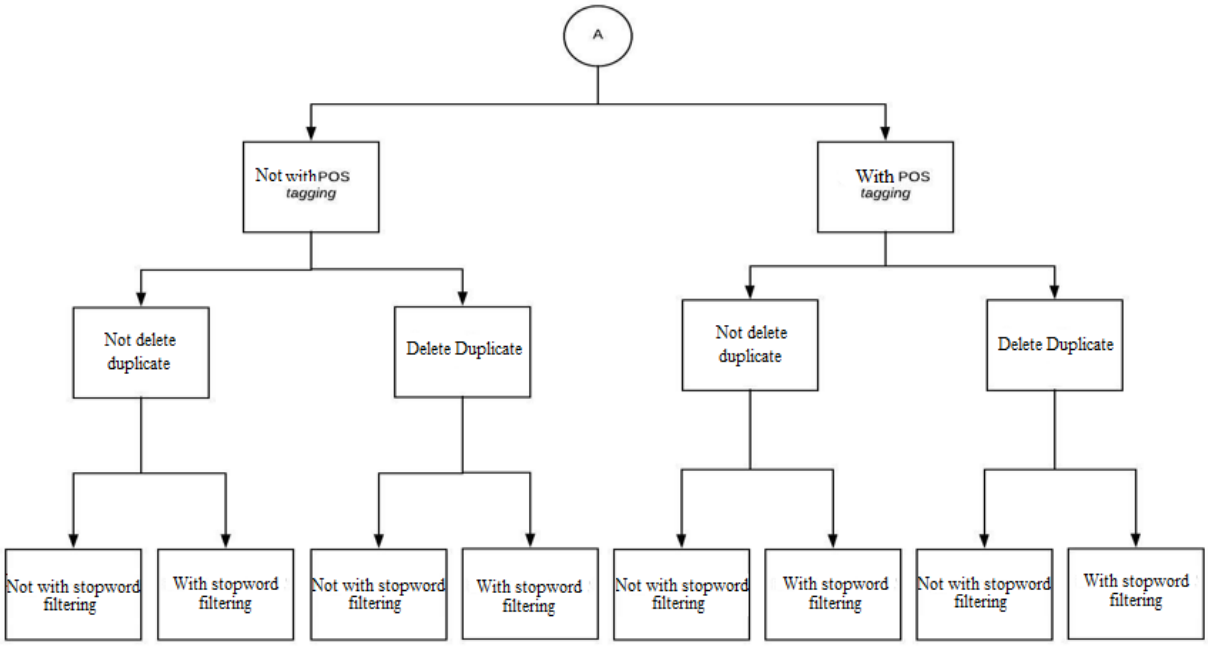

Figure 5. Experiments to obtain the model.

3. RESULTS AND DISCUSSIONS

3.1. Topic Categories of Wordcloud and Wordlink

The first time to perform topic modeling is by determining several main variables, one of them is the number of topics. The number of topics in this study is selected based on the observation results on manual labelling of topics on tweets. The results of manually labeling topic in texts obtained 14 topics that are often associated with LGBT talks on social media twitter. However, among these 14 categories, we select only 8 categories that often appear such as Politic, Religion, Government, Ethnicity, Citizenship, Culture, Technology and others (contains unselected topics). Afterwards, each

tweet with category label is visualized using wordcloud and wordlink using voyant tools application to see words that can be used as characteristics from topic categories so that it is easier to identify the issues that being discussed. The following explains the results of visualization using wordcloud and wordlink on the eight topics.

As displayed in Figure 6, issue on religion that is circulating around LGBT discussions is about people's perceptions to support LGBT that is a wrong thing to do based on religion point wrong of view, like tweet *"!!@onlyaidee: Istri nabi Luth ga diselamatin Allah krn mendukung gay, dibinasakan Allah. Siapa Loe pede banget dukung LGBT? Yakin selamat?"* and many other people's perceptions with majority do not justify the behavior of supporting LGBT people. Not much different from religious issues, cultural issues that are circulating on LGBT talks on Figure 7 shows a form of public inconvenience about the things that is perceived as characteristics of LGBT supports and are often used as a habit or culture as in the tweet: "*Wallahi, kaum LGBT telah merusak indahnya warna dan dalamnya makna PELANGI. - (via herricahyadi) bener... http://tmblr.co/ZsTayt1oTMo-7A*". Government issues that arise in LGBT talks as displayed in Figure 8 shows that the support form government members towards LGBT is a thing that cannot be justified by the community and can cause various rejections, as stated in one of the tweets: "*dulu LGBT dicibir, skrg diterima di seluruh AS. kaum pedofili sdg menempuh jalan yg sama yg ditempuh kaum LGBT*".

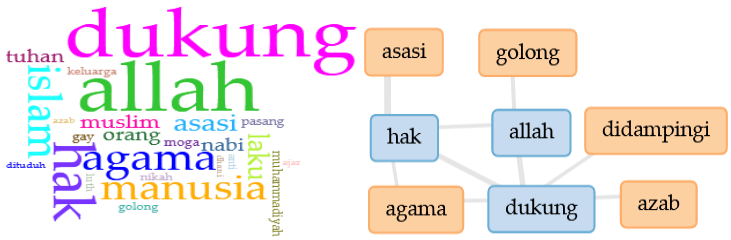

Figure 6. Wordcloud and wordlink on religious topic category.

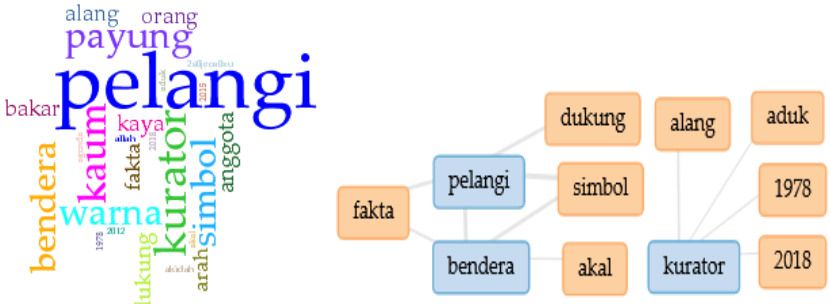

Figure 7. Wordcloud and wordlink cultural topic category.

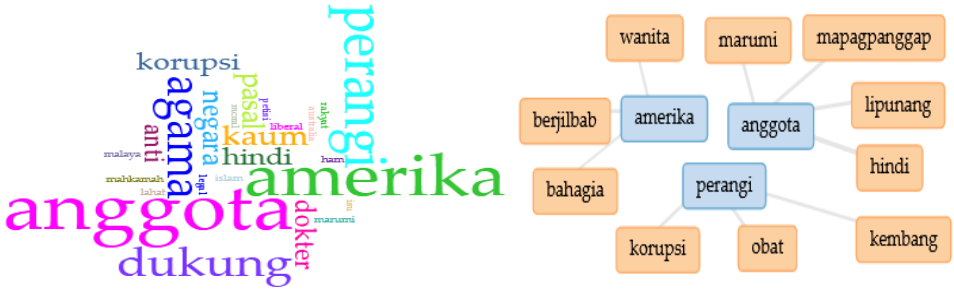

Figure 8. Wordcloud and wordlink government category topic.

LGBT talks are often associated with the recognition of the position of LGBT actors. A result, the issue of citizenship is often arising in LGBT talks as shown in Figure 9. It reveals that the freedom of LGBT actors is something that cannot be accepted in society, as stated in one of tweets: "*wkwkwk pake dikultwit in pula pro kontranya* "@hilaz_28: *LGBT disahkan di USA, yang kebakaran jenggot orang Indonesia* :)"". Other things that often appear in society regarding LGBT talks is political issues as shown in Figure 10. This tells us that political issues that arise in LGBT issues, mostly in the form of public denial for LGBT supports who try to get support from political party leaders in government election as in one of tweets: "*saat pilpres si artis ngetwit dukung jokowi, diblowup media jokower. nah kalo skrg ditulis 'Artis Pendukung Jokowi Dukung LGBT' kok sewot?*". Another thing that also appear in LGBT talks as an effort to get support is with ussue of technology that can be seen in Figure 11. The results suggest that various media are used as a facilitator to spread any form of LGBT so that this kind of sexual disorders is perceived as common things in society and as a results their existence can be acknowledged, as stated in the tweet: "*Perusakan Perilaku Publik Oleh Media Dengan Penayangan Perilaku Banci/LGBT*" by @SrigalaPagi has been Chir..http://chirpstory.com/li/262370Â".

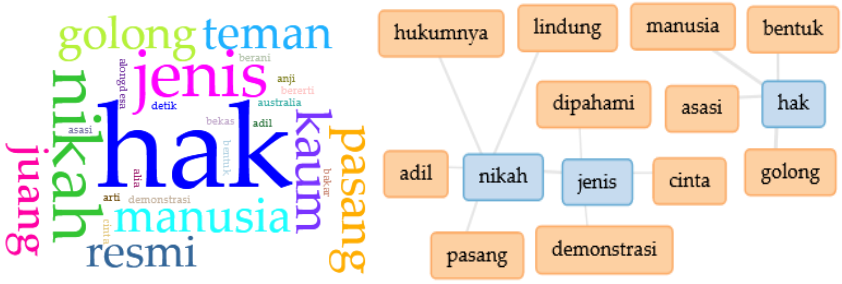

Figure 9. Wordcloud and wordlink citizenship topic category.

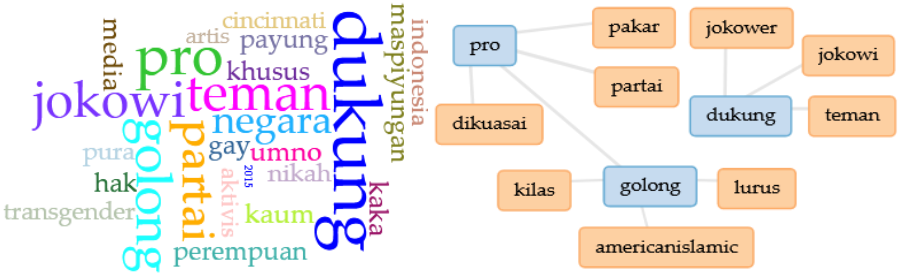

Figure 10. Wordcloud and wordlink politic topic category.

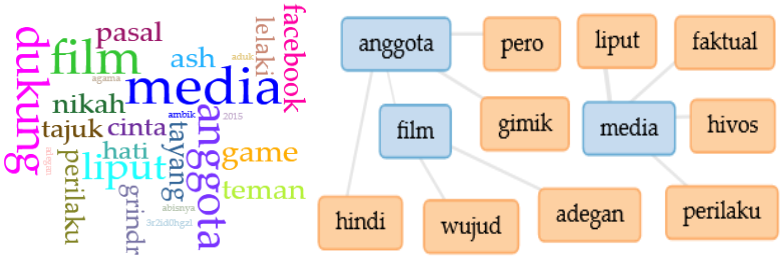

Figure 11. Wordcloud and wordlink technology topic category.

LGBT is one of sexual disorders and it is not surprising that many rejections arise that resulting in immoral acts that are accepted by LGBT actors from their environment either from their friends or even their children such as discrimination, and lack of freedom of human rights as shown in Figure 12. It is listed on one of the tweets showing that there is immoral act such as: "*LGBT itu dianggap kriminal buat mereka, lha iya kalo trs mereka melakukan kejahatan thd orang lain. Lha kalo hidup tenang2 aja?*". Most of these LGBT talks are often associated with many issues such as on tweet: *"Selepas buku kerja mesra-LGBT, muncul satu lagi buku kerja cerita pengalaman "diraba punggung". [Pix @hydir_ruslin] pic.twitter.com/OdOno4GLMi"*. This shows that the support of LGBT is done in various ways and is not liked by society because it can endanger and have its own mystery behind it as displayed in Figure 13.

3.2. Results of modelling with Latent Dirichlet Allocation

A text can contain more than one topic, therefore to extract information for a text can be carried out with soft clustering. One of the most popular methods of soft clustering is called LDA. This technique identify the latent topic information (hidden) in a collection of big data using bag of words approaches that treat every document as a vector of total words and is represented as a probability distribution on several topics. Meanwhile, every topic is represented as a probability distribution over several words [6]. The probability here means a value that is used to measure the level of occurrence of a random event [18].

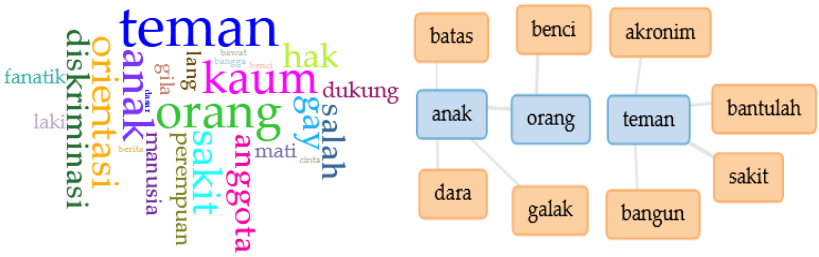

Figure 12. Wordcloud and wordlink immorality topic category.

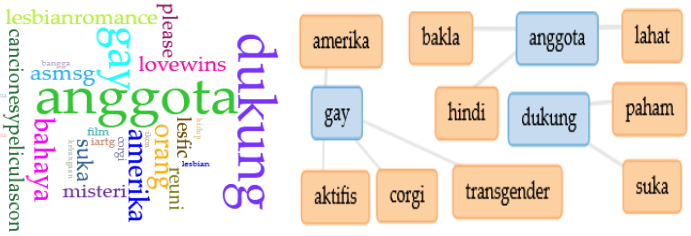

Figure 13. Wordcloud and wordlink other topic category.

The following describes the steps of LDA work steps as shown in Figure 8. The following is description of the work steps of LDA according to Figure 14:

1. Determination of main parameters (β, α, k)
   The values of α dand β are positive real numbers that is less than 1 or can be written as $0 \leq \alpha, \beta \leq 1$, and for the optimal value is approached numerically using Gibbs sampling algorithm in determining the number topics first as shown in Figure 15, where for each $i = 1,2, \ldots N$ (N is the number of words) will be drawn a value that is denoted with u variable that is uniform from 0 to 1. Then, for each topic $k = 1,2, \ldots, K$ (K is the number of topics), it will be calculated the

probability of each topic or $p(k)$. This is done until $u < \frac{p(k)}{p(K)}$, so that the number of topics equals to the number of topics in the i-th word and j (z_{ij}) document.

for $i \leftarrow 1$ **to** N
do
$u \leftarrow$ draw from Uniform$[0, 1]$
for $k \leftarrow 1$ **to** K
do
$P[k] \leftarrow P[k-1] + \frac{(N_{kj}^{\neg ij}+\alpha)(N_{x_{ij}k}^{\neg ij}+\beta)}{(N_{k}^{\neg ij}+W\beta)}$
for $k \leftarrow 1$ **to** K
do
if $u < P[k]/P[K]$
then $z_{ij} = k, stop$

Figure 15. Gibbs sampling algorithm for LDA.

As the value of α increases then we can interpret that every document contains several big topics and as the value of α decrease then a document is more likely to be represented by only several topics. Meanwhile, as β gets larger than a topic contains a mixed of majority of words and as β gets lower than a topic only contains mixed of several words.

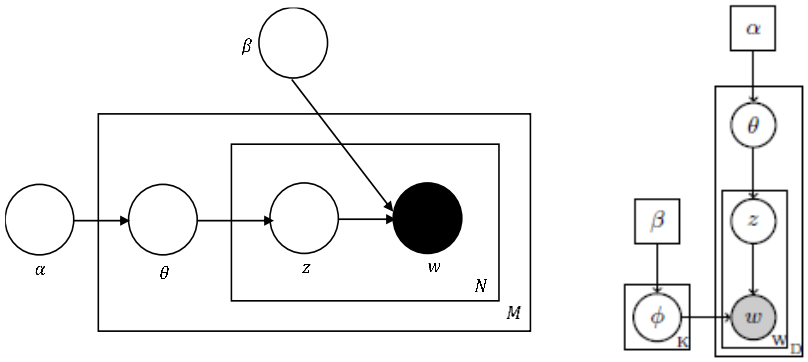

Figure 14. Representation of graphical model from LDA (a) Representation of LDA graphical model (2002) [16][17] (b) Representation of LDA graphical model (2008) [19].

2. Topic assignment for every word in every document/tweet by random
A unique word in tweet for overall document/tweet is obtained from word weighting, then setting the topic can be performed for every word in every document/tweet by random. This can express as a matrix as in the following:

$$wt_{kn} = \begin{bmatrix} wt_{11} & \cdots & wt_{1n} \\ \vdots & \ddots & \vdots \\ wt_{k1} & \cdots & wt_{kn} \end{bmatrix}$$

where $\boldsymbol{wt_{kn}}$ is the *k*-th topic for *n*-th unique word.

3. Determination of topic category for every word in each document/tweet.

4. Create topic matrix for each document where the numbers are in line with the number fo token set up in each topic. The number of tweet token on topic for each document is then the probability is calculated with the following formula:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}. \quad (1)$$

So that θ is obtained as a collection of mixed topics in topic probability matrix on document as shown in the following matrix:

$$\begin{bmatrix} \theta_{11} & \cdots & \theta_{1K} \\ \vdots & \ddots & \vdots \\ \theta_{M1} & \cdots & \theta_{MK} \end{bmatrix}$$

This means that θ_{MK} is the *K-th* collection of mixed on M*-th* document.

5. Randomly renew the topic in one token at the same time. From the collection of mixed topics (θ), each topic can be separated from that collection of topics. Therefore, we can obtain a new matrix containing the probability values of word on document for each topic as follows:

$$\begin{bmatrix} z_{11} & \cdots & z_{1N} \\ \vdots & \ddots & \vdots \\ z_{M1} & \cdots & z_{MN} \end{bmatrix}$$

where z_{MN} is the topic for the M-th document and the N-th word, with $z_{\{1,\dots,K\}}$ or it can be defined that every z as the number of topics in the mixed topics, and each probability can be computed using the following formula:

$$p(z|\theta) = \prod_{m=1}^{M}\prod_{k=1}^{K}\theta_{d,k}^{n_{m,k}}. \quad (2)$$

Variable θ and z are often denoted as latent parameter or posterior Dirichlet parameter.

6. The obtained probability of topics and word distribution on topics (β) yield probability of words that appear as a final result of model development (w). Therefore, the resulting model for one document will bring up words from a formed of topic group, and these words can help us in defining the category for each group. Thus, the total probability based on LDA graphical model can be computed using the following equation:

$$p(w,z,\theta,\phi|\alpha,\beta) = \prod_{j=1}^{M} p(\theta_j|\alpha)\prod_{i=1}^{k} p(\phi_i|\beta)\prod_{t=1}^{N} p(z_{jt}|\theta_j)\ p(w_{jt}|\phi,z_{jt}) \quad (3)$$

Distribution of ϕ is formulated as follows:

$$p(\phi|\beta) = \prod_{k=1}^{K}\frac{\Gamma(\beta_k)}{\prod_{v=1}^{V}\Gamma(\beta_{k,v})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_{k,v}-1} \quad (4)$$

where V is the number of words. Then, the probability for each word for every topic can be calculated using the following equation:

$$p(w|\phi, z) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v}}. \quad (5)$$

Next step in this research is to obtain a good grouping or clustering. Therefore, an experiment is conducted to model the topic using LDA with several different treatment for preprocessing. This difference of treatments consists of combination of three treatments, i.e. delete duplicate, extract word with POS tag, and stopword filtering. The difference that occurs after several trials from these three combinations is the change in the size of Vector Space Model (VSM) where VSM vector with row values indicate the number of document or the number of tweet in this study and columns indicate the appearance of the number of words. This vector is calculated by word weighted as discuses in previous section.

Overall the treatment by deleting duplicates results a decrease in the number of tweets but appearance of the number of words does not change because this treatment only deletes one of the tweets with the same contents. Therefore, the frequency of important tweets goes down, but at the same time the same tweet shows a word that does not contain a meaning such as "yang", "atau", or "dengan" so it becomes difficult to define the groups by those words.

Experiment using POS tag causes change in the number of tweets and the number of words that appear due to this treatment only retains tweets that have words with the chosen tag, i.e. noun. This experiment results in elimination of words with unselected tag, so this causes lack of words within a group and this complicates the process of defining the group. However, with the POS tag, the results of words that appear become cleaner when compared with the results of deletion of duplicates, but it cannot be said that the fitted model with POS tag is the best model because the definition of groups is still unclear.

Finally, treatment with stop word filtering results in a change in the number of tweets and words that appear. Words that appear are much cleaner from POS tag and deletion of duplicates. This indicates that the results by adding stopword filtering treatment is much better in terms of displaying words that appear, therefore, we have 4 possible treatments to obtain the best model including:

1. With POS tag, deletion of duplicates and stopword filtering,
2. With POS tag and stopword filtering,
3. With deletion of duplicates and stopword filtering,
4. With stopword filtering only.

Other differences after performing clustering / grouping by using LDA can be seen from the results of LDA visualization. The best visualization results are obtained from treatment with only adding process of stopword filtering. This is because of the definition of each group by words that appear on LDA visualization can be seen, so that it is easier to determine categories of groups as show in in Figure 16. Successively, the LDA model with stopword filtering defines the categories of groups formed from 1 to 8 each related to religion, government, technology, ethics, culture, politics, citizenship, and others.

Then this best model is used as data training to be used as a reference for the next steps in data testing and visualization and also overall interpretation of the data. Analysis of the results from twitter data related to LGBT processed by LDA can be analyzed with three ways:

1. Distribution of Topics based on Time

The best model selected from testing in the data sample (first semester of 2015) produces an abnormal topic distribution, therefore to see the topic distribution, the data is normalized to

obtain the topic distribution as depicted in Figure 17. From this figure, it can be seen that there are clear differences in probability words in LGBT talks that experience fluctuation that are quite diverse and do not have certain patterns over time. This indicates that the variety of words usage related to LGBT issues remains unchanged although in period of January-June 2015, LGBT talks on twitter media social were mostly associated with words related to culture. For "other" category was omitted in the next visualization because it was too diverse and did not contribute much.

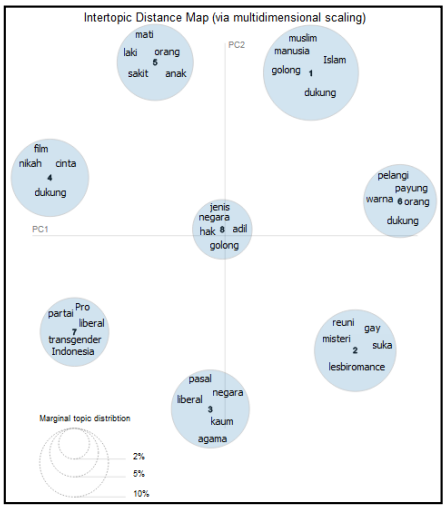

Figure 16. LDA visualization on sample data by adding stopword filtering process.

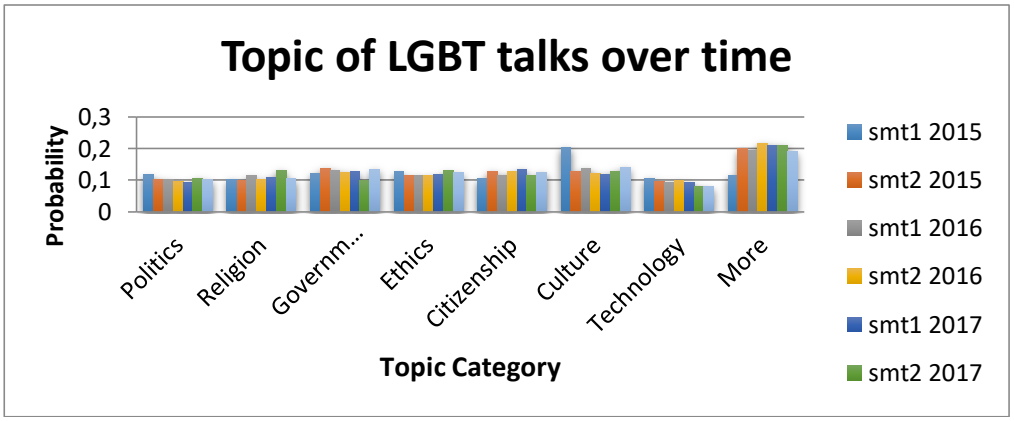

Figure 17. Topic distribution based on time

2. Analysis of Topic Categories over Time

Changes in LGBT talks from time to time can be seen from the probability of word distribution using plot radar as shown in Figure 20. From Figure 18, it can be concluded that some important information such as topic category related to politics, the probability of words diversity has relatively higher numbers in the first semester of 2015 and the second semester of 2017. If we observe in the period of the first semester of 2015, there was an event post presidential election and legalization of LGBT cases in United State that influence the movement of LGBT worldwide, and this is perceived as a new job for the selected president at that moment. Meanwhile, in the second half of 2017, an event occurred where one of the candidates for the regional head of West Java was considered to support LGBT and often gave responses to LGBT

so people think that it is used as red herring prior to the local election of "pilkada" in West Java on June 27th 2018. The connection with this incidence indicates there is a tendency of linking LGBT issues with political issues in Indonesia (Figure 18).

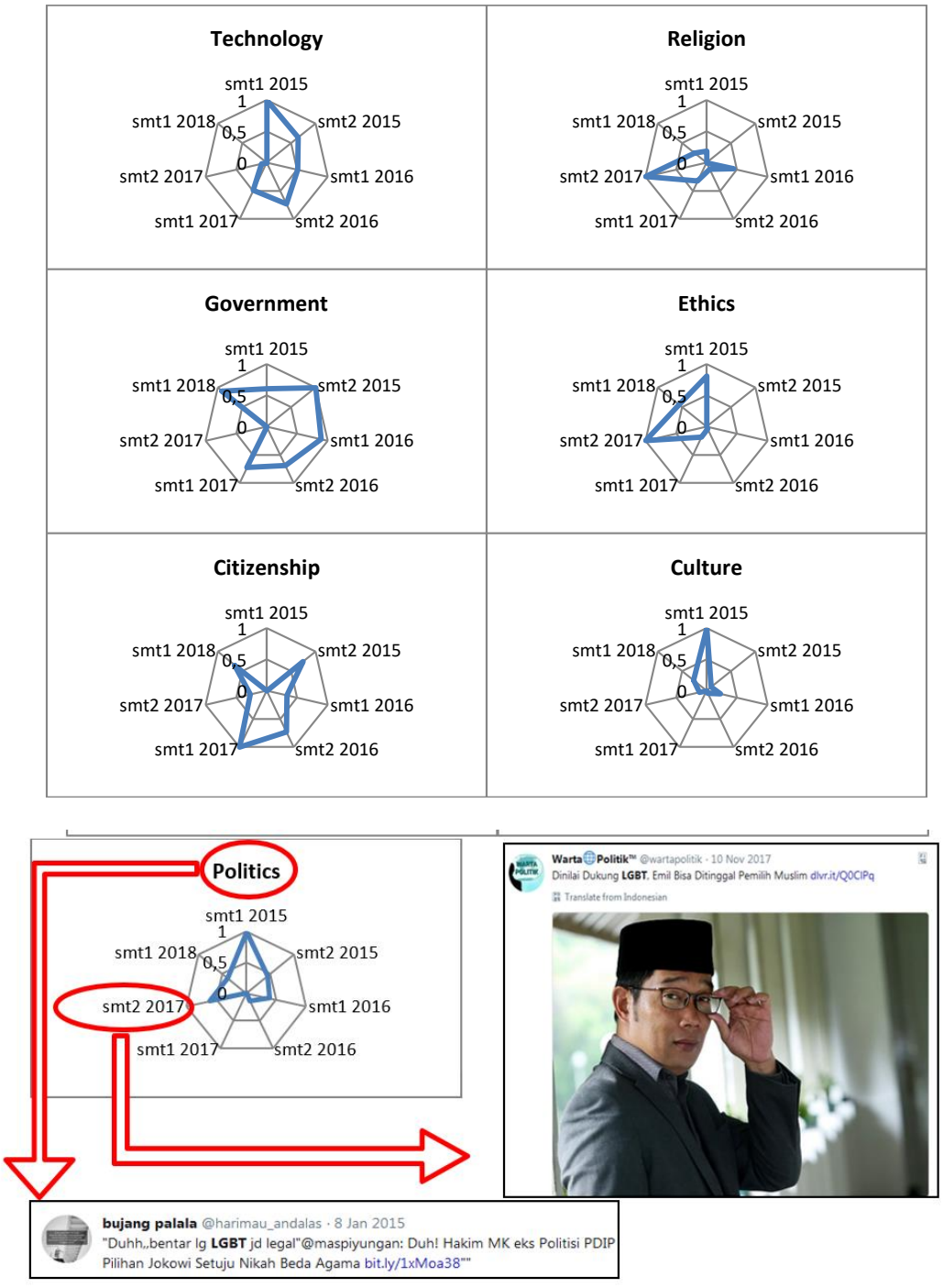

Figure 18. Analysis of topic category over time.

3. Summary of Overall Topic Analysis

To see the pattern or picture of word probability in all categories selected in every period, we can see by using petal plot as shown in Figure 19. The result of petal plot shows that the seven selected categories, if observed every year, have different pattern of probability of word diversity. It is shown that the most diversity of words appear in the first semester of 2015, i.e. about government issues. Meanwhile, in the second semester of 2015, first semester of 2016, and first semester of 2018, the most word diversity is found related to government issues. In the second semester of 2016 and first semester of 2017, the most said words are about citizenship, while second semester of 2017 concerns about religion issue. Therefore, it can be concluded that these LGBT topics do not have specific pattern over time, and also lack of description of a strong indication in terms of tendency for specific topic categories to be associated with LGBT issues.

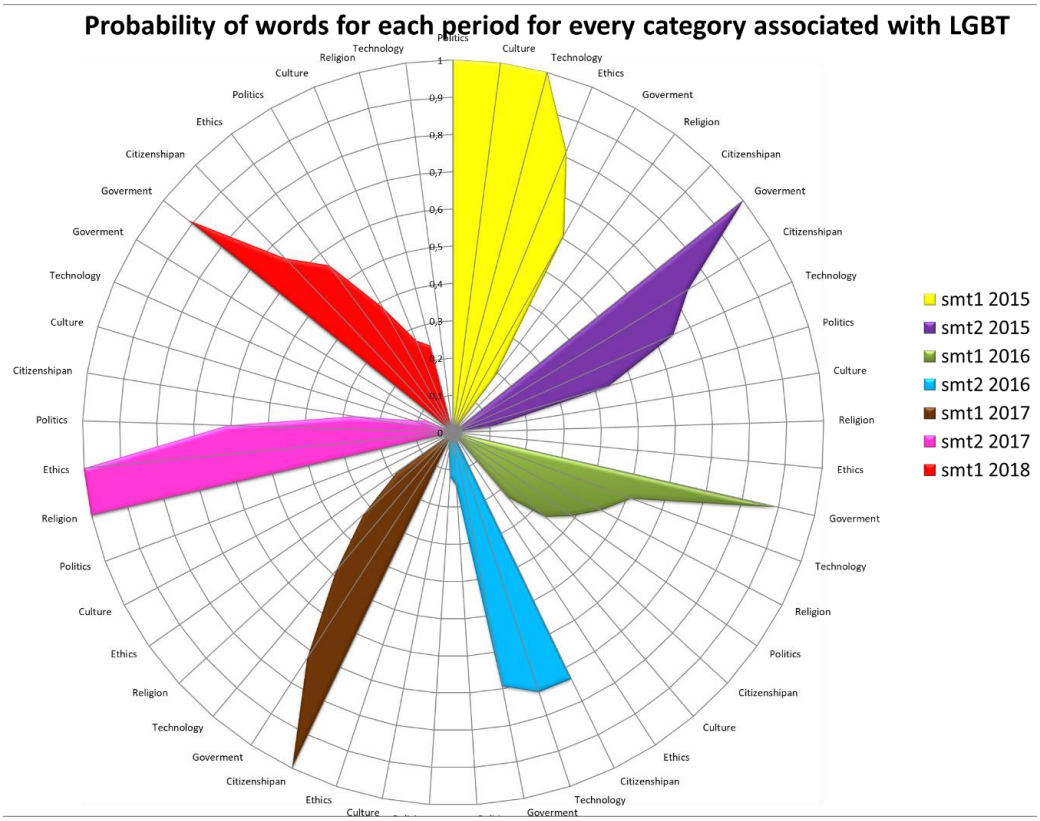

Figure 19. Probability of words for each period for every category associated with LGBT

4. CONCLUSION

From this study, it can be concluded several things that might be considered and explored further. Based on the data and model used, it can be seen that in general there is no tendency or pattern of topics in LGBT talks in twitter social media in Indonesia. The probability of word diversity in one of topic categories such as politics shows that in the second semester of 2017, the probability of word diversity has relatively higher numbers, meaning topic category in politics is often associated with LGBT talks in that period of time. This coincides with the implementation of Pilkada (local elections)

in West Java on June 27th, 2018, indicating that there is tendency to associate LGBT issues with political issues in Indonesia.

REFERENCES

[1] D. J. P. and P. Penyakit, "Laporan Perkembangan HIV-AIDS & Infeksi Menular Seksual (IMS) Triwulan IV Tahun 2017," Jakarta, 2017.

[2] P. L. Pan, J. Meng, and S. Zhou, "Morality or equality? Ideological framing in news coverage of gay marriage legitimization," *Soc. Sci. J.*, vol. 47, no. 3, pp. 630–645, 2010.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is *Twitter*, a Social Network or a News Media?," *Int. World Wide Web Conf. Comm.*, pp. 1–10, 2010.

[4] M. J. Paul and M. Dredze, "You are what you *Tweet*: Analyzing *Twitter* for public health," *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media*, pp. 265–272, 2011.

[5] W. X. Zhao *et al.*, "Comparing *Twitter* and Traditional Media Using Topic Models," *Ecir*, vol. 6611, pp. 338–349, 2011.

[6] L. Hong and B. D. Davison, "Empirical study of topic modeling in *Twitter*," *Proc. First Work. Soc. Media Anal. - SOMA '10*, pp. 80–88, 2010.

[7] D. M. Blei, "Probabilistic topic models (Lecture)," *Commun. ACM*, vol. 55, no. 4, p. 77, 2012.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[9] N. E. Evangelopoulos, "Latent semantic analysis," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 4, no. 6. pp. 683–692, 2013.

[10] N. Evangelopoulos, X. Zhang, and V. R. Prybutok, "Latent semantic analysis: Five methodological recommendations," *Eur. J. Inf. Syst.*, vol. 21, no. 1, pp. 70–86, 2012.

[11] T. Hofmann and Dan Oneata, "Probabilistic latent semantic analysis," *UAI'99 Proc. Fifteenth Conf. Uncertain. Artif. Intell.*, pp. 1–7, 1999.

[12] T. Hofmann, "Probabilistic Latent Semantic Indexing," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 211–218, 2017.

[13] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001.

[14] D. Lee and S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 556–562.

[15] L. Li and Y. J. Zhang, "Non-negative Matrix-Set Factorization," in *Proceedings of the 4th International Conference on Image and Graphics, ICIG 2007*, 2007, pp. 564–569.

[16] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[17] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 424–433, 2006.

[18] J. Supranto, "Statistik teori dan aplikasi jilid 1 / oleh J. Supranto," *Stat. Teor. dan Apl. jilid 1 / oleh J. Supranto*, vol. 2000, no. 2000, pp. 1–99, 2000.

[19] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed Gibbs sampling for latent Dirichlet allocation," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 2008, p. 569.