IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.14, No.2, April 2020, pp. 179~188
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.54507

Social-Child-Case Document Clustering based on Topic Modeling using Latent Dirichlet Allocation

Nur Annisa Tresnasari*[1], Teguh Bharata Adji[2], Adhistya Erna Permanasari[3]
[1,2,3]Department of Electrical Engineering & Information Technology, FT UGM, Yogyakarta, Indonesia
e-mail: ***[1]nurannisa89@mail.ugm.ac.id,** [2]adji@ugm.ac.id, [3]adhistya@ugm.ac.id

Abstrak

Anak adalah masa depan bangsa. Segala perlakuan dan pembelajaran yang diperoleh anak mempengaruhi masa depannya. Saat ini, ada berbagai macam masalah sosial yang berkaitan dengan anak. Untuk memastikan solusi yang tepat untuk masalah anak yang terjadi, pekerja sosial dapat merujuk pada dokumen kasus anak untuk menemukan kasus serupa di masa lalu dan mengadaptasi solusi yang diambil dari kasus tersebut. Namun, membaca banyak dokumen untuk menemukan kasus serupa adalah tugas yang melelahkan dan membutuhkan banyak waktu. Oleh karena itu, eksperimen ini bertujuan untuk mengelompokkan dokumen-dokumen tersebut sesuai dengan topik kasus. Pendekatan pemodelan topik dengan teknik Latent Dirichlet Allocation (LDA) digunakan untuk mengekstraksi topik dari dokumen serta mengelompokkan dokumen-dokumen tersebut berdasarkan kesamaan deskripsi kasusnya. Penentuan model terbaik dilakukan menggunakan perhitungan Coherence Score dan Perplexity. Hasilnya, diperoleh model dengan 5 topik sesuai tipe kasus yang ditargetkan. Hasil penelitian ini mendukung proses penggunaan kembali pengetahuan tentang penanganan kasus sosial anak, sehingga memudahkan untuk menemukan dokumen dengan deskripsi kasus yang sama.

Kata kunci*— clustering, dokumen teks, topic modeling, Latent Dirichlet Allocation, kasus sosial anak*

Abstract

Children are the future of the nation. All treatment and learning they get would affect their future. Nowadays, there are various kinds of social problems related to children. To ensure the right solution to their problem, social workers usually refer to the social-child-case (SCC) documents to find similar cases in the past and adapting the solution of the cases. Nevertheless, to read a bunch of documents to find similar cases is a tedious task and needs much time. Hence, this work aims to categorize those documents into several groups according to the case type. We use topic modeling with Latent Dirichlet Allocation (LDA) approach to extract topics from the documents and classify them based on their similarities. The Coherence Score and Perplexity graph are used in determining the best model. The result obtains a model with 5 topics that match the targeted case types. The result supports the process of reusing knowledge about SCC handling that ease the finding of documents with similar cases.

Keywords*— clustering, text document, topic modeling, Latent Dirichlet Allocation, social-child cases*

Received Februari 27th,2020; Revised April 28th, 2020; Accepted April 29th, 2020

1. INTRODUCTION

One of the most common social problems in Indonesia correlates with children. Various social problems of children are grouped by Kementerian Sosial Republik Indonesia (Kemensos RI), into several categories. These include abandoned babies, abandoned children, children in conflict with the law, children with disabilities, street children, children abused, and children with the needs of special protection [1]. Because children are the future of the nation, every problem related to children must be handled carefully. Based on *Undang-undang Nomor 14 Tahun 2019* about *Pekerja Sosial* and *Peraturan Pemerintah Nomor 44 Tahun 2017* about *Pelaksanaan Pengasuhan Anak,* the handling of children's social cases are carried out by the government and private institutions with the role of social workers.

In handling cases, social workers usually have to make the documentation [2] in a particular format called the Social-Child-Case (SCC) documents. Within the documents, social workers write at least the description of the case and what actions have been or will be taken in handling it. With various information on handling SCC in the past, these documents can be used as valuable references for social workers to determine the SCC solution they are currently responsible for. The assumption is that similar cases have the same problem-solving approach. Therefore, by reading the available SCC documents, social workers try to find similar cases to determine the solution. The problem is that there are too many documents and are not classified. So, it is quite difficult for social workers to find documents with a similar problem.

Thus, this research develops a computational approach to automatically categorizing the SCC documents based on the similarity of the case descriptions. Hence, we can consider our works as text/document clustering problems, since we have no certain label/category on our dataset. Recent technique on text clustering was dominated by two major approach, artificial neural networks based approach and ontology based approach. Wan et.al [3] using ANN to constructing word vector from large amount of documents and employing k-means algorithm to perform clustering task over the constructed word vectors. Another experiments based on ANN is conducted by Saini et.al [4], since they formulating Self Organizing Map (SOM) for generating various genetic operation to achieve the best clusters during the iteration of the algorithms. Another approach of document clustering is by employing pre-constructed ontology network which have condicted by Rupashinga & Park [5], Kang et.al [6], and Sandhiya & Sudarambal [7]. By using ANN and ontology, document clustering reach a promising accuracy. Nevertheles, there is a need of massive efforts for collecting large data in order to perform ANN effectifelly and also tedious work for constructing an ontology network. Hence, since we only have small amount of data and no pre-constructed ontology, we are trying to develop simpler keyword-based approach. It is assumed that a similar set of keywords reflecting a similar case description of the SCC documents. Based on this idea, we try finding documents with similar case descriptions by extracting keywords from each document into the SCC corpus. One of the most popular computational approaches for extracting keywords from documents is topic modeling. Topic modeling works statistically by exploring documents and representing them as collections of frequently co-occurrence terms [8].

The use of topic modeling approach in providing a useful view from a text collection has been done widely in various domains and cases. For example are, in journalism [9], information science [10] [11], and academic field [12]. Meanwhile, some researchers [13]–[15] have tried to use topic modeling for text clustering. According to [8] and [16], Latent Dirichlet Allocation (LDA) is the most popular topic modeling technique, which is also used in [9] - [15]. Therefore, we use the LDA topic modeling to create a topic model from a corpus of SCC documents. The model consists of topics with each keyword, in which each topic assumed to be a cluster. With the clustered documents, social workers are easier to find the available documents with a similar case to the current one.

2. METHODS

2.1. Dataset

The dataset that we used comes from a collection of SCC documents belong to the Dinas Sosial DIY, which are written in Indonesian and available in digital form (doc). They were chosen manually with the requirement that they must contain the same feature considered as the attributes of the case. There are two features found within all documents: problem description and recommendation. Both are available in descriptive text. As a result, 167 documents selected, which is each text's length, are from 50 to 900 words.

2.2. Experimental Design

The experiment design divided into two significant parts. The first part is data preprocessing, and the second part is the topic modeling of preprocessed data.

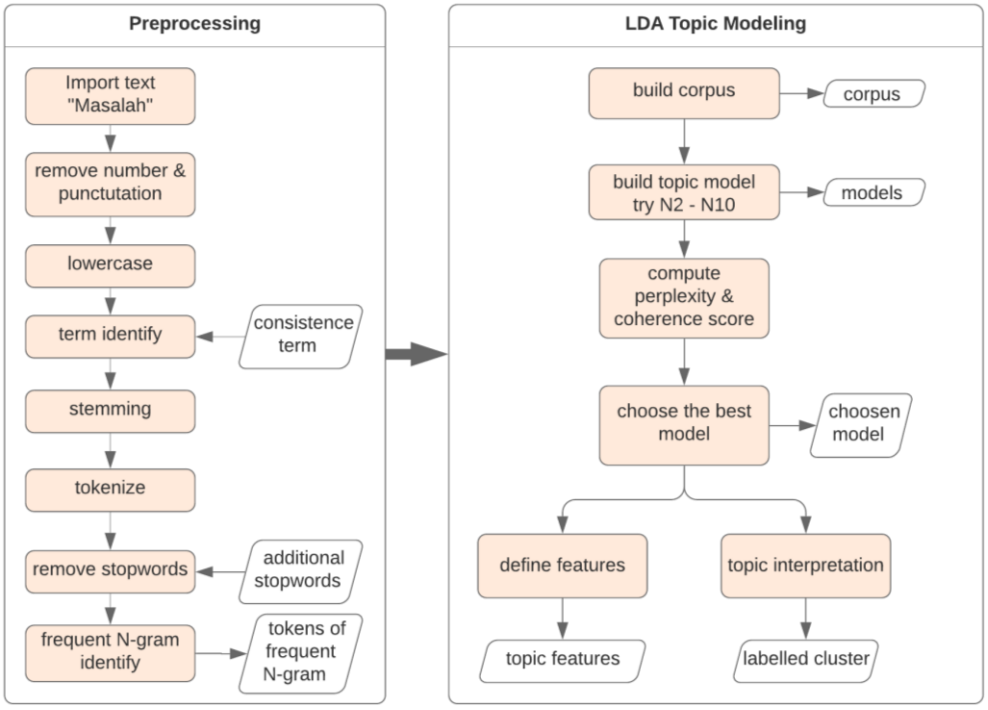

Figure 1. Experiment design

2.2.1. Data Preprocessing

According to [17] and [18], preprocessing is considered an essential step in text processing, as it provides a standard and consistent form of data that affecting the whole experiment result. As shown in Figure 1, we use a series of preprocessing steps consist of 1) number and punctuation removal; 2) case-folding (turn all the text into lowercase); 3) term identification; 4) stemming; 5) stopword removal and 6) frequent n-gram identify. Among the preprocessing, there are default steps (data cleaning and case-folding), the additional step (term identification), and the optional steps (stemming, stopword removal, and frequent n-gram identify).

The term identification usage is performed to handle too many words found in the text, which have the same meaning. So, we consider them as SCC terms with inconsistent writing, as shown in Table 1. Those inconsistencies are the results of differences in abbreviations, the use of words, letters, and space. Inconsistent writing can lead to errors in recognizing a term or word, which affects the whole experiment. Therefore, we add this step to make sure that on the next step, those SCC terms will be recognized consistently.

Table 1. Inconsistence term

NO	Inconsistence Term	FixTerm
1	satuan bhakti, satuan bakti	sakti
2	petugas, pekerja sosial, satuan peksos, sakti peksos	peksos
3	satuan polisi pamong praja, polisi pamong praja	satpol pp
4	dinas sosial	dissos
5	rumah perlindungan sosial	rps
6	yayasan sayap ibu, sayap ibu, yayasan ysi	ysi
7	assessment, assesment, asessment, asesment, assessmen, assesmen, asessmen	asesmen
8	camp asesmen	camp
9	balai rehabilitasi sosial dan pengasuhan anak, balai rspa	brspa
10	balai perlindungan dan rehabilitasi sosial remaja, balai prsr	bprsr
11	balai perlindungan dan rehabilitasi sosial wanita, balai prsw	bprsw
12	balai rehabilitasi terpadu penyandang disabilitas, balai rtpd	brtpd
13	balai rehabilitasis sosial bina karya dan laras, bina karya, bina laras, balai rsbkl	brsbkl
14	case conference	cc
15	orang tua	orangtua
16	rumah sakit	rs
17	sekolah dasar	sd
18	taman kanak-kanak	tk
19	tracing	penelusuran
20	activity daily living	adl
21	penyerahan kembali, diserahkan kembali, menyerahkan kembali	reunifikasi
22	support	dukungan
23	family	keluarga

The next steps that are quite significant contain of stemming and stopword removal. Actually, in [19] and [20], Schofield found that in some instances, the use of stemming and stopword removal does not affect the topic model. Since it can even reduce its stability [19], the use of those steps requires some consideration. In this study, we try experimenting using both steps with the consideration that the text addresses a fairly specific domain. For example, without stemming, some words with the same context (e.g. "pengasuhan", "diasuh", "mengasuh" and "asuhan") are recognized as different tokens. And without some additional stopword, some words which in this context are general terms (e.g. "klien", "kondisi" and "mengalami") frequently appear in many documents, even though they do not mean anything.

The frequent n-grams identification is made based on the assumption that a series of words frequently appear in the documents, have specific meanings that can become the text features, as done in [21]. As shown in Table 2, the results written in italics are named-entity. We get them by experimentally re-run this step while changing the threshold and min_count values. The higher the threshold values, the lesser the n-gram produced. Min_count is a minimum number of n-word's occurrences in sequence. In this experiment, we use bigram and trigram, where both use threshold = 75 and min_count = 3.

Table 2. N-gram

NO	WORDS	NO	WORD	NO	WORD	NO	WORD
1	dissos_diy	11	kena_razia_satpol_pp	21	media_sosial	31	senjata_tajam
2	satpol_pp	12	keras_fisik	22	habis_uang	32	retardasi_mental
3	lingkung_sosial	13	uang_saku	23	lampu_merah	33	obat_larang
4	*psbk_bekasi*	14	ojek_online	24	kelompok_punk		
5	rehabilitasi_sosial	15	interaksi_sosial	25	pondok_pesantren		
6	kelas_sd	16	penuh_butuh_hidup	26	biaya_salin		
7	kelas_smp	17	tonton_konser_musik	27	perangkat_desa		
8	ganggu_jiwa	18	tumpang_kendara	28	*gotong_royong*		
9	habis_bekal	19	konsumsi_obat	29	*pondok_sadar*		
10	kena_jangkau_satpol_pp	20	mantan_suami	30	penyalahgunaan_napza		

2.2.2. LDA Topic Modeling

The final process output from the preprocessing becomes a corpus consisting of n-gram tokens. From this corpus, the topic model is built using the LDA algorithm. As a generative probabilistic model of the corpus, LDA assumes that each document represented as a probabilistic distribution over latent topics, and each topic is characterized by a distribution over words [22]. The words that have the highest probability on each topic are usually used to determine what the topic is. Figure 3 shows the levels of an LDA topic model representation. *M*

represents the number of documents, while *N* represents the number of words in the document. The first level is the corpus level parameter (α and *ß*), which considered as samples in the corpus production process. The second level is the document level parameter (θ), which is a one-time sample of each document. Finally, the word level parameters (*z* and *w*) are sampled once for each word in each document.

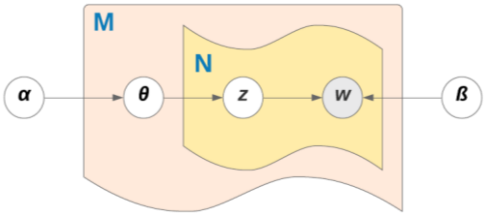

Figure 2. Graphical model representation of LDA

We first modeled the topic from the corpus and replicated this process several times, as was done in [23]. In generating a model, the LDA algorithm requires an input parameter (n) to determine the number of generated topics. Because there was no absolute knowledge about the topic number of SCC documents, we determined the value of n based on the expert's (social worker) assumption on the range of the topic's number. Based on the experts' assumption, we experimented using n = 2 to n = 10.

Determination of the best model (n topic) was carried out with two measurements, which are the Perplexity value [24] and the Coherence score [25]. The value of perplexity showed the confusion metrics or ways to capture the level of 'uncertainty' of a model's prediction result. In contrast, the coherence score indicated the level of semantic similarity between words on a topic. The formulation to calculate perplexity and coherence score are shown in equation (I) [24] and (II) [26].

$$perplexity(D_{test}) = \exp\{-\frac{\sum_{d=1}^{M} \log p\, W_d}{\sum_{d=1}^{M} N_d}\} \quad \text{(I)}$$

where :

W_d = words in document d
N_d = length of document d

$$coherence(V) = \sum_{(vi,vj)\in V} score(vi, vj, \in) \quad \text{(II)}$$

where :

V = a set of words describing the topic
$\in$ = a smoothing factor which guarantees that *score* returns real numbers

In the coherence score calculation, since this experiment uses no external corpus, the score is calculated by the UMass metric with equation (III) [26].

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)} \quad \text{(III)}$$

where :

$D(v_i, v_j)$ = counts the number of documents containing words vi and vj
$D(v_j)$ = counts the number of documents containing words vj

3. RESULTS AND DISCUSSION

Coherence score and perplexity are used to evaluate the proposed model. The resulting coherence score was reaching the top at the number of topics (n) = 6, while the perplexity was reaching the lowest value at the number of topics (n) = 9. Therefore, to determine the best topic amount, in this experiment, we used a trade-off (intersection) between both. The LDA topic

modeling experiment results with n = 2 to n = 10 was shown in Figure 3. The figure showed that the perplexity and coherence score graphs experience an intersection on the number of topics approaching 5. Thus, the number of topics that will be used in the next step is n = 5.

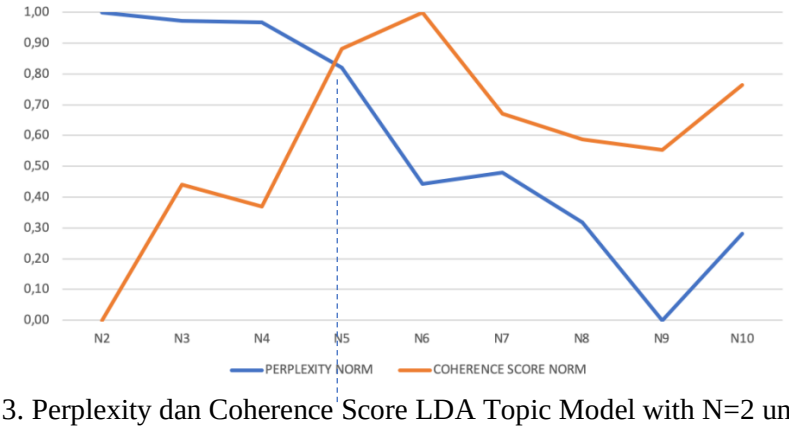

Figure 3. Perplexity dan Coherence Score LDA Topic Model with N=2 until N=10

From n-topic = 5, five document clusters were formed. Visualization of the 5 clusters appeared in Figure 4, which showed the distance between clusters in two-dimensional space. In the Gensim library, the distance between one cluster and another cluster was visualized by a multidimensional scaling technique. Figure 4 showed that 3 of the 5 clusters were entirely separated without overlapping, while the 2 clusters were slightly intersecting. The size of each cluster illustrateed the number of documents incorporated in it. The larger the cluster size, the more documents it contains.

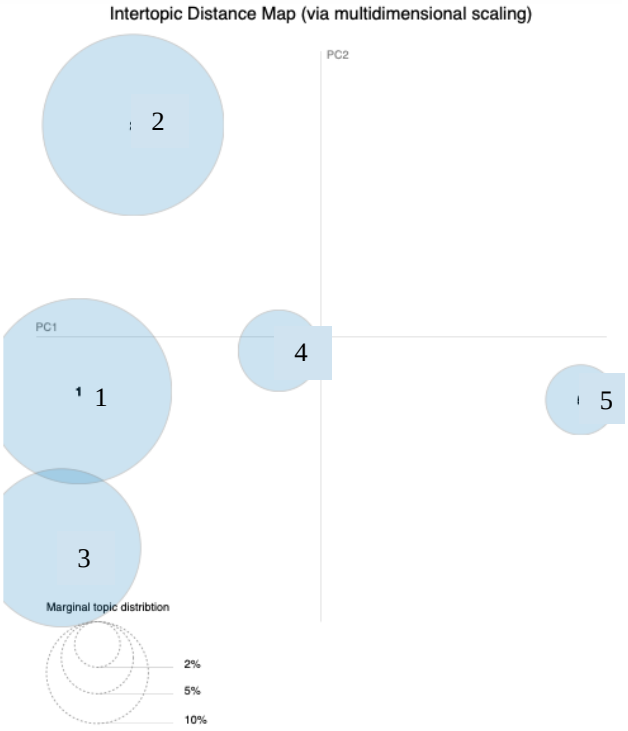

Figure 4. Visualization of the selected model

The words representing a cluster are the keywords of certain documents. This group of keywords is then called the topic. Based on observations of the 5 topics' selected keywords, and by referring to the documents contained in each cluster, we generally describe the dominant topic for each cluster. So the cluster labels (cluster 1 – 5) interpretation can be made briefly, as illustrated in Table 3.

Table 3. Representation of 5 topics

Cluster	Dominant Keywords	Unique Keywords	Brief Representation	Cluster Size (%)
1	tinggal, ayah, rumah, rps, keluarga, peksos, jalan, suami, pergi, yogyakarta, orangtua, asuh, temu, nikah, kerja	suami, kandung, informasi, paman, dissos, cari, ijin, tiri, komunikasi	Broken home (BrH)	33.0%
2	bayi, orangtua, keluarga, asuh, lahir, ayah, tinggal, hamil, nikah, kerja, nenek, rumah, sekolah, serah, titip	bayi, lahir, titip, istri, biologis, sah, hubung	Kelahiran tidak diinginkan (KtD)	31.7%
3	sekolah, teman, ayah, tinggal, kerja, rumah, amen, rps, keluarga, pergi, dissos_diy, putus, saudara, jalan, satpol_pp	amen, saudara, satpol_pp, camp, uang, jangkau, kakek, rp	Beraktivitas di jalanan (AJ)	24.1%
4	keluarga, yogyakarta, bprsr, sekolah, rtm, jalan, laki, tindak, ekonomi, rumah, orangtua, mi, semarang, pondok, bapas	bprsr, tindak, pondok, bapas, alamat, lingkung_sosial, hadir, desa, curi, polisi, rapat, terima, warga	Berhadapan dengan hukum (ABH)	6.4%
5	wahyu, tinggal, bogor, tahan, hanif, dissos_diy, aldizar, annur, sum, rusak, bawa, rujuk, bantul, marbun, bks	tahan, rusak, bawa, bks, dusun, bak, psbk, kantor, korupsi, bab, pulang	Keluarga dalam rehabilitasi (Rh)	4.7%

In Table 3, the Dominant Keywords of each cluster contains 15 words with the highest coherence score. However, in the Dominant Keywords, some words are also keywords in other clusters. While Unique Keywords contains words that only appear in one cluster. Those keywords are used as lexical identifiers. Although they do not directly provide semantic meaning, these lexical features provide clues about the dominant topics of the documents incorporated in a cluster.

The first cluster discusses the condition of families experiencing broken homes (BrH), causing separation and abandonment. It is indicated by the words "tinggal", "rumah" and "pergi" which can be interpreted by the child or one of the family members who invited the child to leave the house or family, with or without permission ("ijin"). The words "jalan" and "kerja" describe the consequences of leaving home, i.e., losing their homes and having to earn a living to survive. Also, the findings of the words "cari", "informasi", "temu", "komunikasi", reinforce that in some cases, there are efforts to be able to re-gather with family.

The second cluster contains a series of documents having similarities in the case of unwanted births (KtD). It is proven by the findings of "bayi", "lahir", "hamil", "sah", "biologis" and "hubung". The word "biologis" indicates that the baby's biological father and mother are not bound in a legal marriage. The mother's pregnancy and the baby's presence are serious problems for her whole family. Some of them occur in economically weak families, so families find it more difficult to accept the presence of babies. Therefore, in some cases, the babies are entrusted ("titip") or handed over ("serah") to children's social welfare institutions — especially the ones with services for abandoned babies and toddlers.

In **the third cluster**, the highlight problem is the children who is caught doing a specific activity on the streets (AJ). Unlike the first cluster, this cluster's discussion revolves around the children who drop out ("putus") from school ("sekolah"), hanging around and do activities to earn money ("uang") on the streets ("jalan") such as beggar ("amen"). Most of them are then caught by the officer ("satpol_pp") and end up in a shelter called "camp". Other findings in this cluster are the words "pergi", "tinggal", "rumah" which is also a keyword in the first cluster. It happens because some documents describe that broken home was the cause of children 's activities on the streets. The relationship between AJ and BrH problems also appears in Figure 4, where both clusters slightly intersect.

For **the fourth cluster**, the most obvious keywords are “bprsr”, “bapas”, “curi” and “polisi” which generally considered as characteristics of children in conflict with the law problem. Starting from a child who violated the law (for example: “curi”), then acted on (“tindak”) by law enforcement (“polisi”). According to the judicial decisions, the children were getting rehabilitation in “bprsr” and or “bapas”. In several documents, the discussions are even reached out to the children’s condition after completing rehabilitation and returning to the community. For example, the process of preparing the environment (“lingkung_sosial”), so that the ‘post-rehabilitation’ children do not get a rejection, and well-accepted (“terima”) by the society (“warga”).

Next, the clearest keywords from **the fifth cluster** are “bks” and “psbk” which are institutions with the rehabilitation services for homeless, beggars, and psychotics (people with mental disorders). This type of case is related to the condition of parents with psychiatric disorders and or tendencies to lead homelessness or begging habits. These conditions conduce their children got improper care, suffering from growth disorders, or even deviant behavior. Some documents contain deviant behavior, starting from individual behavior related to daily activities, e.g., urinate ("bak") and defecate ("bab") habits, up to social behavior related to interactions with others, e.g., communication skills and conflicting tendencies. The word “rujuk” represents a child or family who has received rehabilitation from an institution but then referred to another institution due to certain conditions. There are also documents mentioning destructive behaviors (“rusak”) so, security actions (“tahan”) are required.

Meanwhile, words such as “keluarga”, “orangtua”, “ayah”, “asuh” appear in several clusters, indicating that, in some documents, there are pieces of information about the background of the children’s family. According to [27], family engagement is an influential factor for the success of children’s social welfare practices. So, information relating to family conditions are really helpful. Besides providing clues about the causes of children’s problems, those pieces of information also contributes to provide the best intervention to solve the children's problems.

For comparison, manual classification of the same 167 SCC documents was carried out in the 5 labeled clustering results by the expert. As a result, the proportion of the number of documents included in the 5 labels was illustrated in Figure 5. It seems quite clear that there is a difference of more than 10% proportion on the KtD label. While on other labels, there was a difference between 1-6%. Further observation on the documents included in the KtD cluster found some of the documents which were not suppose to be in the cluster (nonKtD). From the keywords contained, there was one document falls into the KtD cluster because it mentioned the origin of the child, so it has keywords such as "biologis" and "hamil". But because the child’s origin was not the focus of the problem, experts did not include that document to the KtD group. As for other non-KTD documents, it might be detected as KtD because they has keywords such as "asuh", "orangtua", "keluarga", "bayi" or "ayah" appearing together. When in fact there are also keywords such as "curi", "hukum", "aktivitas", "jalan", "pergi", "tinggal", "rumah", which clearly characterizes other clusters. This was a bit confusing considering that other clusters having under 6% difference in proportion, with relatively similar findings.

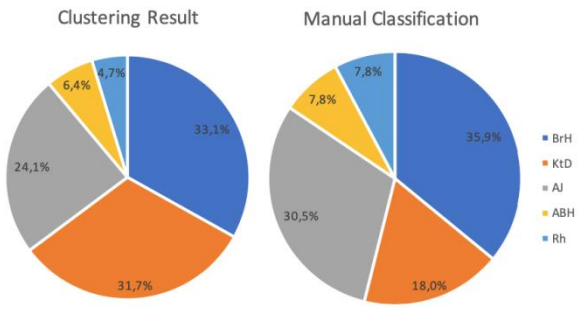

Figure 5. Comparison of LDA Clustering Vs. Manual Classification by Expert with the 5 resulting label

The comparison result showed that this model still has weaknesses in predicting the type of SCC. The possible factors affecting the prediction results are the un-clean preprocessing result or the possibility of other features that can be used as an SCC identifier besides the frequent co-occurring keywords simultaneously. For example, the determination of certain keywords for certain types of SCC by experts. Therefore, to get better SCC similarity, the preprocessing and further exploration of other candidate features from the resulting keywords can be improved in the future.

4. CONCLUSIONS

From the experiments, LDA topic modeling gives promising results in clustering SCC documents according to the topic's similarity. The clusters are obtained using the graph of coherence score and perplexity. The best resulting clusters can be found when coherence score and perplexity plots intersect. The intersection occurs as the number of topics approach 5. The five clusters can be interpreted and labeled according to the targeted case types. It supports the process of reusing knowledge about SCC handling, making it easier to find documents with the same case description. However, when compared with the manual classification result, there was still a big difference in one cluster. This difference could be influenced by the results of un-clean preprocessing, or the possibility of other features that can be used as SCC identifiers besides the frequent of co-occurring keywords simultaneously. Therefore, to obtain better SCC similarity, finding other candidate features from the resulting keywords of the preprocessing and further exploration can be improved in the future.

REFERENCES

[1] Indonesian Ministry of Social, *Pedoman Pendataan dan Pengelolaan Data Penyandang Masalah Kesejahteraan Sosial dan Potensi dan Sumber Kesejahteraan Sosial*. Indonesia, 2012, pp. 1–7.

[2] R. S. H. Ellya Susilowati, Krisna Dewi, Meiti Subardhini, Dwi Yuliani, Tuti Kartika, Rini Hartini Rindra, "Kompetensi Pekerja Sosial dalam Pelaksanaan Tugas Respon Kasus Anak Berhadapan dengan Hukum di Cianjur," *PEKSOS J. Ilm. Pekerj. Sos.*, vol. 16, no. 1, pp. 71–87, 2017.

[3] L. J. Wan H., Ning B., Tao X., "Research on Chinese Short Text Clustering Ensemble via Convolutional Neural Networks," in *Artificial Intelligence in China*, 2020, pp. 622–628.

[4] N. Saini, S. Saha, and P. Bhattacharyya, "Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution," *Cognit. Comput.*, vol. 11, no. 2, pp. 271–293, 2019.

[5] R. A. H. M. Rupasingha and I. Paik, "Alleviating sparsity by specificity-aware ontology-based clustering for improving web service recommendation," *IEEJ Trans. Electr. Electron. Eng.*, vol. 14, no. 10, pp. 1507–1517, Oct. 2019.

[6] S. Kang *et al.*, "Ontology-Based Ambiguity Resolution of Manufacturing Text for Formal Rule Extraction," *J. Comput. Inf. Sci. Eng.*, vol. 19, no. 2, Feb. 2019.

[7] R. Sandhiya and M. Sundarambal, "Clustering of biomedical documents using ontology-based TF-IGM enriched semantic smoothing model for telemedicine applications," *Cluster Comput.*, vol. 22, no. 2, pp. 3213–3230, 2019.

[8] X. Sun, X. Liu, B. Li, Y. Duan, H. Yang, and J. Hu, "Exploring topic models in software engineering data analysis: A survey," in *IEEE/ACIS 17th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2016*, 2016, pp. 357–362.

[9] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journal.*, vol. 4, no. 1, pp. 89–106, 2016.
[10] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *J. Inf. Sci.*, vol. 43, no. 1, pp. 88–102, 2017.
[11] M. Shovkun, K. R. Fleischmann, and B. Xie, "Computational social science using topic modeling: Analyzing patients' values using a large hospital survey," *Proc. Assoc. Inf. Sci. Technol.*, vol. 55, no. 1, pp. 892–893, 2018.
[12] Y. H. Kee, C. Li, L. C. Kong, C. J. Tang, and K. L. Chuang, "Scoping Review of Mindfulness Research: a Topic Modelling Approach," *Mindfulness (N. Y).*, vol. 10, no. 8, pp. 1474–1488, 2019.
[13] A. Onan, H. Bulut, and S. Korukoglu, "An improved ant algorithm with LDA-based representation for text document clustering," *J. Inf. Sci.*, vol. 43, no. 2, pp. 275–292, 2017.
[14] C. Li *et al.*, "LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering Changzhou," in *WWW '18 Companion April 23-27, 2018, Lyon, France.*, 2018, vol. 2, pp. 1699–1706.
[15] H. Ma and T. Zhang, "Research on policy text clustering algorithm based on LDA-Gibbs model," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 23, no. 2, pp. 268–273, 2019.
[16] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.
[17] W. Etaiwi and G. Naymat, "The Impact of applying Different Preprocessing Steps on Review Spam Detection," *Procedia Comput. Sci.*, vol. 113, pp. 273–279, 2017.
[18] M. Petrović, Dorde and Stanković, "The Influence of Text Preprocessing Methods and Tools on Calculating Text Similarity," *Ser. Math. Inform.*, vol. 34, no. 5, pp. 973–994, 2019.
[19] A. Schofield and D. Mimno, "Comparing Apples to Apple: The Effects of Stemmers on Topic Models," *Trans. Assoc. Comput. Linguist.*, vol. 4, pp. 287–300, 2016.
[20] A. Schofield, M. Magnusson, and D. Mimno, "Pulling out the stops: Rethinking stopword removal for topic models," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, pp. 432–436, 2017.
[21] V. H. A. Soares, R. J. G. B. Campello, S. Nourashrafeddin, E. Milios, and M. C. Naldi, "Combining semantic and term frequency similarities for text clustering," *Knowl. Inf. Syst.*, vol. 61, no. 3, pp. 1485–1516, 2019.
[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003.
[23] M. V. Mantyla, M. Claes, and U. Farooq, "Measuring LDA topic stability from clusters of replicated runs," *Int. Symp. Empir. Softw. Eng. Meas.*, 2018.
[24] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "Research on Topic Detection and Tracking for Online News Texts," *IEEE Access*, vol. 7, pp. 58407–58418, 2019.
[25] S. K. Habibabadi and P. D. Haghighi, "Topic Modelling for Identification of Vaccine Reactions in Twitter," *ACM Int. Conf. Proceeding Ser.*, 2019.
[26] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," *EMNLP-CoNLL 2012 - 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. Proc. Conf.*, no. July, pp. 952–961, 2012.
[27] K. Toros, D. M. DiNitto, and A. Tiko, "Family engagement in the child welfare system: A scoping review," *Child. Youth Serv. Rev.*, vol. 88, no. July 2016, pp. 598–607, 2018.