Supervised Probabilistic Latent Semantic Analysis (sPLSA) for Estimating Technology Readiness Level

Donny Aliyanto[1], Riyanarto Sarno[2]
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
[1]donny13@mhs.if.its.ac.id, [2]riyanarto@if.its.ac.id

Bagus Setya Rintyarna[3]
Department of Informatics
Muhammadiyah University of Jember
Jember, Indonesia
[2]bagus.setya@unmuhjember.ac.id

***Abstract*—The increase of global competition today encourages universities in the world including Indonesia to be able to compete with world-class universities. QS World University Rankings is an annual publication of university rankings based on six key indicators by Quacquarelli Symonds (QS). One of the indicators is academic reputation which employs survey research to evaluate the score. This paper proposes a new insight for ranking the Universities in Indonesia by making use of Supervised Probabilistic Latent Semantic Analysis (sPLSA) for extracting the level of readiness of academic journals of university staff to estimate the score of academic reputation. A corpus of keywords is developed based on Bloom Taxonomy to determine the prior level of readiness of academic journal. Supervised PLSA is employed to determine the most probable level. An academic reputation score is then computed based on the level of readiness. Lastly, we rank Indonesian Universities based on the score. To validate the result, we collect 450 abstract of academic journals from several Indonesian Universities. The results of the experiment indicate that the proposed method is promising with a distance value of 10 and similarity of 0.8 compared to the ground truth.**

Keywords—Academic Reputation; Probabilistiv Latent Semantic Analysis; pLSA; Expectation Maximization; University Ranking.

I. INTRODUCTION

University ranking is one way to measure the quality of college [1]. Quality becomes important that must be achieved by universities [2]. Several assessment indicators are used to measure the quality of the university. One of them is the academic reputation of the college, which has an essential role because it has a high degree of relevance to other assessment indicators [3].

University academic reputation is one of the indicators describing the success of universities in conducting research and scientific development [4]. It is also used to measure the technology readiness level of the university. The level of readiness becomes the basis of government to map the follow-up action of continuation of research produced by researchers in college. A good academic reputation can be observed from how much research is generated, how important the impacts are generated for the community and how it can handle the continuation of the resulting research [4].

Information extraction of technology readiness level of universities using survey method is aimed to obtain objective results from the research context [6]. The use of this survey method is considered accurate, but not effective because it is labor intensive, which requires much time and high cost [7].

In that regards, this paper proposes a technique to optimize conventional survey method. The use of Supervised Probabilistic Latent Semantic Analysis (sPLSA) method is focused on the process of modeling topics from the abstract of research documents [8]. The resulting topics may reflect the focus of the research direction of the researchers representing the readiness level of the assessed colleges.

The main purpose of sPLSA is to extract the most probable topic of academic journal of Indonesian Universities previously determined by employing a corpus of keywords developed by using Bloom Taxonomy. The extracted topic is considered to represent the technology readiness level of the university[5]. The level of readiness is then used to estimate the score of academic reputation of universities to generate their ranking.

II. PREVIOUS RESEARCH

The pre-existing research explains the use of academic reputation indicators in QS World University Rankings. These indicators are obtained from a manual survey of academic expert respondents worldwide. The number of registered respondents of QS World University Rankings is more than 70.000 peoples up to 2017. This is how it becomes a time consuming and high-cost survey processes [4].

The background of the respondents of QS is mostly art, literature, engineering, biology, health sciences, natural sciences, and also social sciences. The information collected from the respondents represents university development based on the research and the quality of human resource [4]. The results of the survey are used to analyze and determine the university rank. This survey process requires high cost and time. It is considered inefficient.

The indicator that has the highest score among others is academic reputation. The weight of academic reputation is 40% of the overall weight of indicators. A previous work [4] explained that the most dominant aspect determining the academic reputation of a university is the quality of the publication. In this research, we assign the quality of publication of university by evaluating the level of readiness of the publication. We model the level by employing sPLSA as well as a corpus of keywords developed from Bloom Taxonomy.

III. PROPOSED APPROACH

There are 4 main step of the proposed approach i.e.: 1) dataset collection, 2) corpus development, 3) text pre-

processing and 4) the determination of the score of academic reputation.

A. Dataset Collection

The dataset used for evaluating the proposed method in this work are paper abstracts from nine most reputable universities in Indonesia listed in QS World University Rankings. We pick 50 abstracts of paper from each of those universities. The abstract documents are selected from the most cited paper in google scholar.

B. Corpus of Keywords Development

The corpus is a set of word that is associated with the level of readiness to map the suitable context of the dataset to the level of readiness [9]. The role of the corpus is to determine the initial assumption label in the dataset as input for sPLSA method [10]. The set of keywords contained in the corpus is originated from the Bloom Taxonomy which consists of 6 cognitive categories [8]. The number of words of Bloom Taxonomy in every category is shown in Table 1.Since level of readiness has nine level of category, it is necessary to split the Bloom Taxonomy to suit the 9 required categories [9].

TABLE. I KEYWORDS NUMBER OF IN BLOOM TAXONOMY

No	Bloom Taxonmy Cropus Category	Number of Words
1.	Knowledge	35
2.	Comprehension	29
3.	Application	36
4.	Analysis	51
5.	Sinthesys	51
6.	Evaluation	46
Sum of Words		248

In this work, we manually sort the keyword and split it into nine categories representing the level of readiness. The result of this mapping technique can be seen in Table 2.

TABLE. II NUMBER OF WORDS IN EVERY LEVEL OF READINESS AFTER MAPPING PROCESS

No	Keyword Corpus Level	Number of Words
1.	TRL 1	31
2.	TRL 2	23
3.	TRL 3	32
4.	TRL 4	15
5	TRL 5	32
6.	TRL 6	24
7.	TRL 7	26
8.	TRL 8	30
9.	TRL 9	34
Sum of Words		248

To provide a better performance of the corpus we enrich the keywords in every level of readiness by making use of WordNet library. WordNet is an English lexical database that organizes its collection in term of synonym set (synset). Every synset represents different sense of word. An effective way to enrich the keyword is by extracting the synonym of the keywords from WordNet library [10] and add it to the corpus. Table 3 indicates the number of keywords in each level of readiness after enriched with the collection of synonym of WordNet library. We use the result of this step as the final corpus to determine the prior level of readiness of the abstract document from 9 most reputable universities in Indonesia.

TABLE. III NUMBER OF WORDS IN EVERY LEVEL AFTER ENRICHING PROCESS

No	Keyword Corpus Level	Number of Words				
		Default	*Level 1*	*Level 2*	*Level 3*	*Level All*
1.	TRL 1	31	50	57	62	132
2.	TRL 2	23	43	54	59	75
3.	TRL 3	32	41	44	47	70
4.	TRL 4	15	25	29	30	50
5.	TRL 5	32	55	68	73	120
6.	TRL 6	24	28	29	29	31
7.	TRL 7	26	48	58	63	122
8.	TRL 8	30	55	66	72	109
9.	TRL 9	34	60	74	77	100
Sum of Words		248	405	479	512	809

C. Text Pre-processing

In this step, we perform text pre-processing to remove non-alphabetic characters as well as unimportant words. The initial stage is tokenization. Tokenization is to split documents into elements commonly called tokens. Next stage is stopword removal. This process begins with removal of all form of punctuation and removal of words that have no meaning or not important [14]. Usually, stopword removes connecting words and prepositions. The last stage is stemming. Stemming is the process of removing additive in a word that aims to obtain the basic form of the word [12]. In various documents, it can be found in various words actually comes from the same root, but written in different forms.

D. Determination of Academic Reputation Score

The process of determining the initial label assumption is the step that must be done to determine the topic class that is within the scope of the topic in the sPLSA method. Determination of early label assumptions is generally determined manually by analyzing any topic that includes content in the abstract paper document dataset. This analysis is done by the expert. In this paper, word matching method is done to reduce the time and cost in analyzing the paper. The word matching method is done by employing the corpus previously developed.

The frequency of words from the pre-processing step corresponds to one of the level of the corpus from the enriching step is then calculated. The formula is presented in equation (1).

$$Tf = tf_{dt} \quad (1)$$

The level with the three highest frequency of words is assigned as the level of the associated abstract. So every abstract is assigned with three labels of readiness level. Consequently, it is necessary to assign one most probable level of readiness for the abstract document. Probabilistic Latent Semantic Analysis (PLSA) developed by Hoffman (1999) is Latent Semantic Analysis which uses probabilistic to determine the probabilistic value of each topic of the text document. Latent Semantic Analysis (LSA) emerged as the first technique that can produce representations of documents comprising sets of words, LSA is the most widely known method for embedding feature of Bag-of-Words (Landauer, Foltz, and Laham, 1998) as a representation of the document.

PLSA is a statistically based method, which counts on the co-occurrence of terms and documents with a latent class. Consequently, it has a more robust statistical foundation and is able to provide a proper generative data model. In addition, it can deal with domain-specific

synonymy and polysemy. PLSA has proven to be effective and has been used in many applications. However since PLSA is based on Expectation-Maximization (EM) algorithm, it still suffers from long execution time when dealing with large datasets. [15]

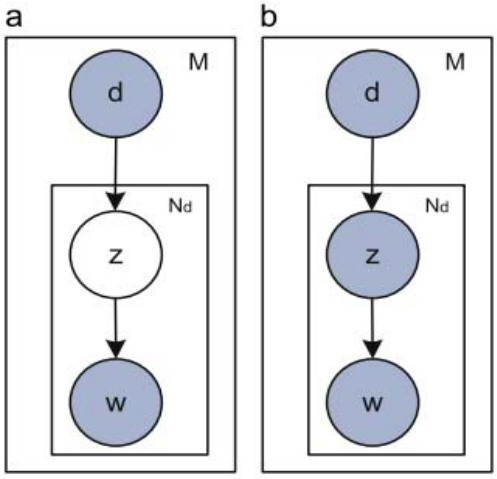

Fig. 1. Graphical models of the pLSA (a) and SpLSA (b). Nodes represent random variables. Shaded nodes are observed variables and unshaded ones are unseen variables. The plates stand for repetitions. In the framework of SpLSA, the latent aspect ‘‘z’’ is equal to the class labels of training data. So, it can be seen during training

The pLSA [12] is firstly proposed to model text collection in an unsupervised way. It assumes that the words are generated from a mixture of latent aspects which can be decomposed from a document. Here, we regard each aspect in the pLSA as one particular motion class. In another word, the number of aspects is equal to the number of classes. We notice the importance of the class label information in training data for the classification task. Considering this important information, we propose to learn the pLSA model in a supervised manner, which not only simplifies the learning process of the pLSA, but also improves its recognition accuracy.

Probabilistic Latent Semantic Analysis (PLSA) is used to calculate the probability of words and documents. PLSA can be used to identify words with multiple meanings and mapping those words on various topics. The relationship between the document, the topic, and the word [16] can be seen in Fig 2. Supervised PLSA is supervised topic modeling method where the topic is possible to be assigned initially.

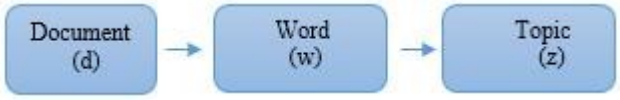

Fig. 2. Relationship among document, topic, and word

Supervised PLSA is used to classify words into topics that are observed (latent). So, each document is clustered based on topics. The algorithm is done by determining the number of topics (z) and initializing parameters of probabilities. P(z) is probability of topics, P(d|z) is probability document that contains topic and P(w|z) is the probability of words contained in the topic. For all k n j, calculate :

$$P(w_j, z_k) = \frac{n_{j,k}}{n_k} \quad (1)$$

As the initialization of the P(w|z) and random initialization of the P(z|d).

$$P(d_i, w_j) = \sum_{k=1}^{k} P(Z_k)P(d_i|Z_k)P(W_j|Z_k) \quad (2)$$

The calculation of word in the document is described in (2). The next step is to calculate the probability for each parameter using Expectation Maximization with two steps, namely E step and M step. E step is used to calculate the probability of the topics in the document and can be seen in (3).

$$P(z_k|d_i, w_j) = \frac{P(w_{j|z_k})P(z_{k|d_i})}{\sum_{k=1}^{k} P(w_{j|z_l})P(z_{l|d_i})} \quad (3)$$

The next step is used to update the value of the parameter and can be seen in the (4) and (5). The results of sPLSA calculation are the probability of the word in a topic and the probability of topics in a document.

$$P(w_{kj}|z_k) = \frac{\sum_{i=1}^{N} n(d_{i|w_j})P(z_{k|d_i w_j})}{\sum_{k=1}^{k}\sum_{k=1}^{k} n(d_{i|w_m})P(z_{k|d_i w_m})} \quad (4)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_{i|w_j})P(z_{k|d_i w_j})}{n(d_i)} \quad (5)$$

The result of the topic probability to the document from abstrac document is converted into academic reputation score. This score is obtained from the level weight as presented in Table 4.

TABLE. IV WEIGHT OF FINAL TOPIC

No	Final Topic	Level Weigth
1.	TRL 1	10
2.	TRL 2	20
3.	TRL 3	30
4.	TRL 4	40
5.	TRL 5	50
6.	TRL 6	60
7.	TRL 7	70
8.	TRL 8	80
9.	TRL 9	90

To calculate the score of academic reputation using :

$$Final\ Score = \frac{\Sigma\,(\vartheta^{(d_i)}\ x\ level\ weight(i))}{\Sigma\ all\ topic\ label} \quad (8)$$

The score reflects the quality of academic reputation of the university. This is to compare reputation results from every college and sort the final rank by comparing the score of academic reputation.

IV. EXPERIMENT AND RESULT

Experiment in this paper is conducted with 450 abstract documents of research paper collected from nine reputable universities in Indonesia. These papers are obtained from Google Scholar. We use nine level of technology readiness of the corpus as the guidelines. The first step is performing text pre-processing stage to abstract datasets. The word reduction in the dataset is shown in Table 5. Secondly, we determine the level of every abstract based on the corpus of

keyword previously developed. We make use of sPLSA to calculate the most probable level of readiness previously determined by employing the corpus of keywords.

TABLE. V NUMBER OF DATASET WORD AFTER REDUCTION PROCESS

University Name	Original Word	Result Preprocess
Institut Pertanian Bogor	10539	6293
Institut Teknologi Bandung	9287	5675
Institut Teknologi Sepuluh Nopember	9630	6015
Universitas Airlangga	9457	6009
Universitas Brawijaya	8651	5469
Universitas Diponegoro	8397	5327
Universitas Gadjah Mada	10214	6279
Universitas Indonesia	8823	5645
Universitas Muhammadiyah Surakarta	10330	6117

The second step is done by updating the corpus using the synonym set of words from WordNet library. The results of this step as in Table 3. The third step is matching the words in every dataset with the keyword corpus based on all level synonyms. Table 6 shows the number of word in the matching words of the dataset with the keyword corpus based on the level synonyms.

TABLE. VI TERM FREQUENCY IN KEYWORD CORPUS

University	Term Frequency in Keyword Corpus Level								
	1	2	3	4	5	6	7	8	9
IPB	14	12	14	9	17	5	17	13	13
ITB	18	14	18	24	22	8	24	17	12
UNAIR	20	14	13	11	18	6	20	11	16
UGM	20	16	20	13	25	8	22	15	17
UI	19	17	18	9	20	6	21	13	17
UMS	20	10	18	13	19	8	20	13	18
UB	18	18	15	10	17	8	17	15	16
UNDIP	12	10	12	7	18	10	17	13	14
ITS	15	13	16	11	23	8	15	15	17

Based on Table 6, the highest frequency of occurrence of words is in the top 3 classes. Next step after determining topic label to all documents then it will do the process of sPLSA mehod with probabilistic number with iteration. The result of topic to document probability show in Table 7.

TABLE. VII TOPIC TO DOCUMENT PROBABILITY USING sPLSA

Univ	Technology Readiness Level								
	1	2	3	4	5	6	7	8	9
IPB	0	0	0	0	0.1489360	0	0.1489360	0	0.170214
ITB	0	0	0	0	0.1489282	0	0.1489282	0	0.170230
UNAIR	0	0	0	0	0.1489219	0	0.1489219	0	0.170209
UGM	0	0	0	0	0.1489346	0	0.1489346	0	0.170214
UI	0	0	0	0	0.1489350	0	0.1489350	0	0.170218
UMS	0	0	0	0	0.1489395	0	0.1489395	0	0.170204
UB	0	0	0	0	0.1489357	0	0.1489357	0	0.170214
UNDIP	0	0	0	0	0.1489356	0	0.1489356	0	0.170213
ITS	0	0	0	0	0.1489369	0	0.1489369	0	0.170213

TABLE. VIII UNIVERSITY RANKING BASED ON 2 EXPERIMENT RESULTS WITH ACADEMIC REPUTATION INDICATOR

Rank	Gound Truth	Experiment I		Experiment II	
		Univ.	Score	Univ.	Score
1.	ITB	ITB	2.974926	ITB	3.688014
2.	UI	UII	2.978581	UI	3.687980
3.	UGM	UI	2.978678	UGM	3.687952
4.	UNAIR	UGM	2.978760	UMS	3.687952
5.	IPB	UNPAD	2.978736	UNDIP	3.687952
6.	UNDIP	ITS	2.978836	UB	3.687937
7.	ITS	UNAIR	2.978934	ITS	3.687931
8.	UMS	UPH	2.979156	UNAIR	3.687900
9.	UB	BINUS	2.978880	IPB	3.687713

The first experiment is based on university academic reputation. The result shows a similarity ranking of 80% and gap difference of 10, and percentage of tolerance of of 88,88% compared to the ground truth. While in the second experiment, the similarity ranking is 51%, the gap difference is 16 and the percentage of tolerance is 66.66%. Compared to the ground truth of QS World University Rankings 2016-2017, the first experiment indicates the better results than the second experiment. University's overall score is presented in Table 9.

TABLE IX. UNIVERSITY'S OVERALL SCORE

Rank	University	Score
1.	ITB	0.7434
2.	UI	0.7123
3.	UGM	0.68
4.	IPB	0.6154
5.	UNDIP	0.5877
6.	ITS	0.5841
7.	UB	0.5764
8.	UNPAD	0.5755
9.	UNAIR	0.5617

V. CONCLUSION

In this study, we propose a new approach to optimize assessement indicator in university academic reputation rankings. We developed a corpus based on Bloom Taxonomy to assign prior labels of the dataset. Supervised PLSA is employed to determine the most probable topic. A formula to calculate academic reputation score is proposed. The result of experiment seem to be promising compared to the ground truth of QS World University Rankings.

VI. ACKNOWLEDGMENT

Author would like to thank to Institut Teknologi Sepuluh Nopember for supporting this research.

VII. REFERENCES

[1] K. S Reddy, E. Xie and Q. Tang,"Higher education, high-impact research and world university rankings : A Case of India and Comparison with China," Pacific Science Review. B : Humanities Social Sciences, vol. 2, no. 1 pp 1-21, 2016

[2] Tatiana Sidorenko, Tatiana Gorbatova,"Efficiency of Russian Education Through The Scale of World University Rankings," *Procedia - Social and Behavioral Sciences,* vol. 166, pp. 464-467, 2015.

[3] Anne Wil K. Harzing, Ron van der Wal, "Google Scholar As a New Source for Citation Analysis," *Ethics in Science and Environtmental Politics,* vol. 8, no. 1, pp. 61-73, 2008.

[4] Peter Serdyukov , "Journal of Research," *Department of Teacher Education, School of Education,* vol. 320, no. 1, pp. 10-11, 2014.

[5] Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia,

"Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 42 Tahun 2016 Tentang Pengukuran dan Penetapan Tingkat Kesiapterapan Teknologi," Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia, Jakarta, 2016.

[6] Adina, Petruta Pavel,"Global University Rankings – A Comparative Analysis,"Procedia Economics and Finance vol. 26, pp. 54-63, 2015.

[7] Mu-Hsuan Huang,"Opening The Black Box of QS World University Taxonomy of Educational Objective," Stanford University, California, 1981

[8] T. Verma, "Tokenization and Filtering Process in RapidMiner," *Int. J. Appl. Inf. Syst. – ISSN 2249-0868 Found. Comput. Sci. FCS,* vol. 7 no. 2, p. 16–18, 2014 .

[9] Ferilli.S., F. Esposito and D. Grieco, "Automatic learning of linguistic resources for stopword removal and stemming from Text," *Procedia Comput. Sci,* vol. 38, no. 1, p. 116–123, 2014.

[10] Bagus Setya Rintyarna, Riyanarto Sarno, "Adapted weighted graph for Word Sense Disambiguation," in *International Conference on Information and Communication Technology (ICoICT)*, Bandung, Indonesia, 2016.

[11] H. K. Landauer, e. W. Foltz and a. Laham, "An Introduction to Latent Semantic Analysis," vol. 1, no. 1, pp. 259-284.

[12] E.. K. Kouassi, T. Amagasa and H. Kitagawa, "Efficient Probabilistic Latent Semantic Indexing using Graphics Processing Unit," *International Conference on Computational Science, ICCS 2011*, Japan, 2011

[13] F. Revindasari, R. Sarno and A. Solichah,"Traceability Between Business Process and Software Component using Probabilistic Latent Semantic Analysis," *International Conference on Informatics and Computing (ICIC),* vol. 1, no. 9, p. 245 – 250, 2016.

[14] D. Blei, L. Carin and D. Dunson, "Probabilistic topic models," *IEEE* Signal *Process. Mag,* vol. 27, no. 1, pp. 55-56, 2010.

[15] Daniel Jurafsky, James H. Martin, "Part-of-Speech Tagging," in *Speech and Language Processing*, 2016.

[16] K. S. Reddy, E. Xie and Q. Tang, "Higher education, high-impact research, and world university rankings: A Case of India and comparison with China," *Pacific Science Review B:Humanties and Social Sciences,* vol. 2, no. 1, pp. 1-21, 2016.

This page intentionally left blank