

Pemeringkatan Perguruan Tinggi Menggunakan Metode Probabilistic Latent Semantic Analysis (pLSA) untuk Mengukur Tingkat Kesiapterapan Teknologi di Indonesia

Donny Aliyanto, Riyanarto Sarno, dan Bagus Setya Rintyarna

Departemen Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)

e-mail: riyanarto@if.its.ac.id

Abstrak—Meningkatnya persaingan global di dunia pendidikan saat ini mendorong perguruan tinggi di dunia termasuk Indonesia bisa bersaing dengan perguruan tinggi kelas dunia (World Class University / WCU). QS World University Rankings adalah salah satu publikasi tahunan pemeringkatan universitas berdasarkan enam indikator oleh Quacquarelli Symonds (QS). Salah satu indikatornya adalah Reputasi Akademik yang menggunakan metode survei secara manual untuk mengevaluasi bobotnya. Makalah ini mengajukan konsep baru untuk Pemeringkatan Universitas di Indonesia dengan menggunakan metode *Probabilistic Latent Semantic Analysis* (PLSA) dengan menggunakan topik utama yaitu jurnal akademik yang dilakukan secara otomatis untuk memperkirakan nilai kriteria secara kualitatif berdasarkan tingkat kesiapterapan teknologinya. Penggunaan metode *Probabilistic Latent Semantic Analysis* (PLSA) dapat digunakan sebagai alternatif untuk menentukan reputasi akademik perguruan tinggi dengan survei manual, dengan ground truth 80,00%, perbedaan gap yaitu 10, dan toleransi perbedaan peringkat yaitu 88,88%.

Kata Kunci—Reputasi Akademik, Probabilistic Latent Semantic Analysis, PLSA, Expectation Maximization, Pemeringkatan Universitas

I. PENDAHULUAN

PERGURUAN tinggi adalah salah satu sarana yang efektif untuk mengembangkan ilmu pengetahuan dan teknologi. Peringkat perguruan tinggi adalah cerminan kualitas yang dimiliki oleh perguruan tinggi. Salah satu upaya untuk meningkatkan kualitas Perguruan Tinggi adalah dengan melakukan pemeringkatan Perguruan tinggi di seluruh dunia [1]. Kualitas menjadi salah satu hal yang harus dicapai oleh perguruan tinggi [2]. Mengukur kualitas perguruan tinggi dapat dilihat dari berbagai indikator penilaian, salah satunya adalah reputasi akademik perguruan tinggi yang memiliki peran penting karena memiliki tingkat hubungan yang tinggi dengan indikator penilaian lainnya [3].

Reputasi akademik perguruan tinggi merupakan salah satu indikator yang menggambarkan keberhasilan perguruan tinggi dalam melakukan pengembangan penelitian dan sains kepada para mahasiswa [4] dan juga digunakan untuk mengukur tingkat kesiapterapan teknologi dimana level tersebut menjadi dasar

pemerintah untuk memetakan tindak lanjut penelitian yang dihasilkan oleh peneliti di perguruan tinggi [5]. Reputasi akademis yang baik dapat diamati dari seberapa banyak penelitian dihasilkan, seberapa penting dampak yang ditimbulkan bagi masyarakat dan melanjutkan penelitian yang dihasilkan [4].

Salah satu cara untuk menilai reputasi akademik perguruan tinggi dengan menggunakan survei. Mendapatkan informasi menggunakan metode survei bertujuan untuk mendapatkan hasil yang objektif dari konteks penelitian [6]. Penggunaan metode survei ini dinilai akurat namun tidak efektif karena masih dilakukan dengan cara manual yang membutuhkan waktu dan biaya tinggi [7]. Oleh karena itu, makalah ini mengusulkan metode baru yang dapat meningkatkan metode survei yang manual dengan menggunakan algoritma text mining Metode *Probabilistic Latent Semantic Analysis* (PLSA) yang berfokus pada proses pengelompokan topik dari abstraksi dokumen penelitian yang diambil dari google scholar [8].

II. URAIAN PENELITIAN

Penelitian yang sudah ada sebelumnya menjelaskan tentang penggunaan indikator reputasi akademik pada QS World University Rankings. Indikator ini diperoleh dari survei yang dilakukan secara manual oleh mitra responden ahli akademik dari QS World University Rankings yang tersebar di seluruh dunia. Jumlah responden yang terdaftar sampai 2017 lebih dari 70.000 orang. Ini menjadi proses survei yang paling banyak [4].

Latar belakang responden ini adalah seni dan sastra, teknik dan teknologi, biologi dan ilmu kesehatan, ilmu alam, dan ilmu sosial. Metode survei bertujuan agar informasi yang didapatkan bisa sama dan subjektif. Informasi ini menjelaskan tentang perguruan tinggi di berbagai negara dengan responden yang paling baik dalam penelitian dan pengembangan sumber daya manusia [4]. Hasil survei perlu dianalisis lagi untuk menentukan peringkat universitas. Proses ini membutuhkan biaya yang tinggi dan waktu yang lama sehingga tidak efisien untuk dilakukan. Indikator reputasi akademik ini memiliki posisi tertinggi dari indikator lainnya. Indikator ini memiliki

bobot 40% dari penilaian sehingga memiliki pengaruh yang besar terhadap peringkat universitas saat dipublikasikan. Dalam penelitian lain [4], jelaskan bahwa indikator reputasi akademik sangat dominan di beberapa universitas peringkat publikasi. Menentukan indikator pemeringkatan menjadi kunci kesuksesan dalam menghasilkan peringkat universitas yang baik. Salah satu cara untuk mengoptimalkan reputasi akademik adalah dengan menggunakan optimasi pemodelan topik. Disini saya menggunakan metode PLSA sebagai pengganti survei kualitatif. Kami juga menganalisa topik dari abstraksi paper yang mencerminkan kualitas penelitian di setiap universitas.

III. PETUNJUK TAMBAHAN

Dalam penelitian ini, terdapat 3 tahap dalam kinerja sistem, termasuk penyusunan dataset, tahap memperkaya kata corpus, tahap preprocessing, metode generatif, dan tahap akhir penentuan bobot. Setiap tahap saling berkaitan satu sama lain, karena hasil dari satu tahap adalah masukan untuk tahap selanjutnya.

Penyusunan dataset dan kata pada corpus adalah tahap persiapan masukan yang digunakan oleh sistem untuk menghasilkan topik akhir dengan mengumpulkan dataset dalam bentuk dokumen penelitian dari universitas dan kata kunci korpus (TKT). Pada tahap preprocessing dan generative, proses yang dilakukan adalah text preprocessing. memperkaya kata corpus, dan tahap text mining yaitu PLSA. Tahap ini merupakan tahap inti dari proses penentuan probabilitas dari topik yang dihasilkan. Sedangkan pada tahap akhir, penentuan bobot adalah tahap konversi dari topik probabilitas yang dihasilkan menjadi nilai reputasi akademik perguruan tinggi yang akan menjadi gambaran besar peringkat perguruan tinggi.

A. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini adalah dokumen dengan ekstensi (.txt) yang berisi abstraksi makalah penelitian yang dimiliki oleh perguruan tinggi. Dataset ini akan menjadi masukan dari tahap preprocessing dan akan menjadi dasar untuk menentukan probabilitas topik akhir. Pemilihan dokumen abstrak dilakukan dengan beberapa persyaratan, seperti dokumen penelitian abstrak yang hanya berfokus pada teknik berbasis penelitian, dokumen abstrak yang diperoleh dari 16 universitas yang terdaftar dalam peringkat reputasi akademik QS World University Rankings. Keterbatasan perguruan tinggi yang termasuk dalam kriteria penilaian adalah perguruan tinggi institusi pendidikan akademik dalam berbagai sains dan teknologi [6]. Sumber data kertas abstrak yang digunakan berasal dari Google Scholar yang bisa ditemukan pada (<http://scholar.google.co.id/>). Setiap perguruan tinggi menggunakan 50 dokumen abstraksi sehingga total dataset adalah 800 dokumen.

B. Korpus

Korpus adalah sekumpulan tingkat kata kunci yang terkait dalam tingkat kesiapterapan teknologi. Sistem dapat

memetakan dari dataset korpus ke tingkat kesiapterapan teknologi.

Korpus yang digunakan berlandaskan pada Taxonomy Bloom Corpus yang terdiri dari 6 level [9]. Dalam penelitian ini, tingkat kesiapterapan teknologi digunakan sebagai dasar pemetaan kualitas penilaian reputasi akademik perguruan tinggi yang terdiri dari 9 level, sehingga perlu membuat pemetaan terhadap Taxonomy Bloom agar sesuai dengan 9 level yang dibutuhkan [10]. Kata kunci dalam setiap level tidak mungkin terjadi tumpang tindih, karena akan mempengaruhi probabilitas akhir metode PLSA. Jumlah kata dari Korpus Taksonomi Bloom seperti pada Tabel 1.

Tabel 1.
Jumlah Kata pada 6 Level Korpus Taxonomy Bloom

No	Level Korpus Taxonomy Bloom	Jumlah Kata
1.	Knowledge	35
2.	Comprehension	29
3.	Application	36
4.	Analysis	51
5.	Sinthesys	51
6.	Evaluation	46
Total Kata		248

C. Teks Preprocessing

Dataset berupa abstraksi dokumen perguruan tinggi akademik yang dilakukan proses teks preprocessing dimana dokumen ini akan dibersihkan menjadi dokumen yang tidak mengandung karakter non-abjad dan penghapusan kata-kata yang tidak penting atau berpengaruh dalam hubungan konten dokumen.

Tahap awal adalah *tokenization*. Tokenisasi adalah proses untuk membagi teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token-token. Tahap selanjutnya adalah *stopword removal*. Proses ini dimulai dengan penghapusan semua bentuk tanda baca dan penghapusan kata-kata yang tidak memiliki arti atau tidak penting [9]. Biasanya, kata penghubung dan kata depan. Tahap terakhir adalah *stemming*. *Stemming* merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) [11]. Dalam berbagai dokumen, bisa ditemukan berbagai kata yang sebenarnya berasal dari arti yang sama, namun ditulis dalam bentuk yang berbeda.

50 abstraksi dokumen akademik dari data yang dikumpulkan setiap universitasnya akan digabungkan menjadi satu file dokumen besar sebelum proses preprocessing sehingga hasil preprocessing yang diinput dalam tahap PLSA tidak memerlukan perhitungan yang besar dan memudahkan dalam proses analisis dan penentuan nilai dari probabilitas topik akhir. Probabilitas ini kemudian akan diubah menjadi nilai reputasi akademik.

D. Memperkaya Kata Korpus

Korpus yang digunakan memiliki jumlah kata yang terbatas dan beragam frekuensi. Hal ini dapat mempengaruhi proses penentuan asumsi label awal sebagai masukan dalam proses pLSA. Jumlah kata dari 9 tingkat korpus seperti pada Tabel 2.

Tabel 2.
Jumlah Kata Dari 9 Level Korpus Tkt

No	Level Kata Korpus	Jumlah Kata
1.	TKT level 1	31
2.	TKT level 2	23
3.	TKT level 3	32
4.	TKT level 4	15
5.	TKT level 5	32
6.	TKT level 6	24
7.	TKT level 7	26
8.	TKT level 8	30
9.	TKT level 9	34
Total Kata		248

Berdasarkan Tabel 2 dapat dilihat bahwa jumlah kata dalam masing-masing level sangat terbatas dan perbandingan jumlah keyword dalam corpus dengan jumlah kata sangat tinggi. Hal ini perlu dikembangkan lagi dengan memperkaya corpus keyword agar corpus dapat menampung banyak pengetahuan dari ragam kata.

Penggunaan kata sinonim menjadi cara yang efektif untuk menambahkan daftar kata kunci baru [5]. Mencari kata sinonim menggunakan WordNet library dapat dikembangkan sesuai dengan pencarian sinonim pada setiap kata dalam 9 tingkat kesiapterapan teknologi kata kunci korpus. Sinonim akan ditambahkan ke korpus *default* di setiap tingkat kesiapterapan teknologi. Peningkatan kata kunci korpus dapat memperluas jumlah kata korpus di TKT. Tingkat sinonim telah terlihat dari tingkat kemiripan tertinggi kata sinonim dan kata dasar [12]. Jumlah sinonim di setiap korpus TKT sangat bervariasi, begitu juga jumlah kata yang meningkat di korpus TKT. Tabel 3 menunjukkan peningkatan kata korpus menggunakan sinonim di setiap tingkat.

Tabel 3.
Jumlah Kata Pada Setiap Level Synonym

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

Tabel 3 menunjukkan jumlah kata di setiap tingkat dengan menggunakan sinonim 4 tingkat. Hasil dari langkah ini akan menjadi kata kunci akhir korpus baru.

E. Menentukan Asumsi Label Awal

Proses penentuan asumsi label awal merupakan tahap yang harus dilakukan untuk menentukan kelas topik yang menjadi lingkup topik dalam metode PLSA. Penentuan asumsi label awal pada umumnya ditentukan dengan cara manual yaitu dengan menganalisa topik apa saja yang mencakup konten dalam dataset abstrak dokumen. Analisa ini dilakukan oleh expert. Untuk mengurangi waktu dan biaya kerja dalam menganalisa dalam abstrak dokumen ini dilakukan dengan metode pencocokan kata berdasarkan term frequency.

Pada abstrak dokumen, hasil tahap teks preprocessing akan dihitung frekuensi kemunculan kata yang sesuai dengan salah satu level corpus keyword TKT final hasil update dari tahap memperkaya corpus TKT seperti pada persamaan (1).

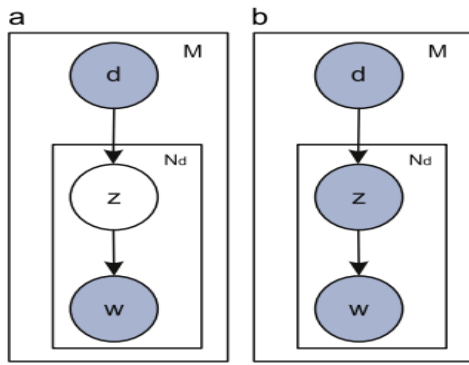
$$Tf = tf_{dt} \quad (1)$$

Dalam satu dokumen abstrak paper penelitian memungkinkan untuk terdapat ke-9 level TKT karena konten kata yang ada dapat muncul disemua level corpus keyword TKT. Jumlah frekuensi tertinggi akan menjadi asumsi topik awal dari dokumen abstrak paper penelitian yang akan digunakan untuk masukan tahap proses PLSA. Dikarenakan asumsi topik awal diharuskan lebih dari satu, karena prinsip utama dari metode PLSA adalah mempunyai label jamak dan supervised maka dalam abstrak dokumen ini ditentukan bahwa asumsi topik awal dari setiap dataset berjumlah 3, yang didapatkan dari jumlah 3 frekuensi tertinggi. Jika asumsi topik awal ditentukan 1 tertinggi maka tahap metode PLSA tidak akan berpengaruh pada probabilitas topik final karena hasil topik sudah pasti dapat ditentukan.

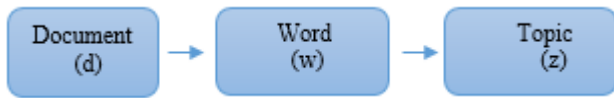
F. Supervised Probabilistic Latent Semantic Analysis (sPLSA)

Probabilistic Latent Semantic Analysis (PLSA) merupakan pengembangan dari LSA dengan menggunakan pendekatan baru untuk pengindeksan dokumen secara otomatis yang didasarkan pada model kelas laten statistik untuk analisis faktor data yang terhitung. PLSA dilengkapi dari korpus pelatihan dokumen teks dengan generalisasi dari algoritma *Expectation Maximization*, dimana model yang digunakan mampu menangani kesinoniman domain-spesifik serta menangani kata - kata polisemi.

PLSA digunakan untuk menghitung probabilitas pada sebuah kata-kata dan dokumen. PLSA dapat digunakan untuk mengidentifikasi kata-kata dengan beberapa arti dan pemetaan kata-kata dalam berbagai topik. Hubungan antara dokumen, topik, dan kata. sPLSA adalah Supervised Topic Modeling dimana topiknya dapat diketahui.



Gambar 1. Model grafik dari pLSA (a) and SpLSA (b). Node menggambarkan variabelnya. Node gelap adalah variable yang dapat diketahui dan node yang terang adalah variable yang tidak dapat diketahui.



Gambar 2. Hubungan antara dokumen, topik, dan kata

PLSA juga disebut model statistik untuk menemukan pola dalam dokumen sehingga lebih mudah untuk menghubungkan konteks dengan setiap kata yang muncul pada dokumen. Dengan proses pemodelan, topik atau konteks akan diperoleh dari teks asli dokumen tanpa deskripsi dokumen sebelumnya [13]

sPLSA digunakan untuk mengklasifikasikan kata-kata menjadi topik yang dapat diamati. Jadi, setiap dokumen dikelompokkan berdasarkan topik. Algoritma dari PLSA sebagai berikut: menentukan jumlah topik yang telah ditentukan (z) dan menginisialisasi parameter probabilitas. P(z) adalah probabilitas topik, P (d | z) adalah dokumen probabilitas yang berisi topik, P (w | z) adalah Kemungkinan kata-kata yang terkandung dalam topik itu secara acak.

1. Untuk semua k dan j, hitung

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

Sebagai inialisasi dari P(w|z) dan inialisasi secara acak dari P(z|d).

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

Perhitungan kata dalam dokumen dijelaskan pada (2). Langkah selanjutnya adalah menghitung probabilitas masing-masing parameter menggunakan Expectation Maximization dengan dua langkah, Yaitu E step dan M step. Langkah E digunakan untuk menghitung Kemungkinan topik dalam dokumen dan dapat dilihat pada (3)

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

Langkah M digunakan untuk memperbarui nilai parameter dan dapat dilihat pada (4) dan (5). Hasil perhitungan sPLSA adalah probabilitas kata dalam suatu topik dan probabilitas topik dalam sebuah dokumen.

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

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

G. Menentukan Bobot Final Reputasi Akademik

Perhitungan metode pLSA menjadi dasar dalam tahap penentuan bobot final reputasi akademik. Hasil probabilitas topik terhadap dokumn dari abstraksi dokumen dikonversi menjadi skor reputasi akademik. Skor ini didapatkan dari ketentuan bobot seperti pada Tabel 4.

Tabel 4. Perhitungan Bobot Topik

No.	Topik Akhir	Bobot level
1.	TKT level 1	10
2.	TKT level 2	20
3.	TKT level 3	30
4.	TKT level 4	40
5.	TKT level 5	50
6.	TKT level 6	60
7.	TKT level 7	70
8.	TKT level 8	80
9.	TKT level 9	90

Untuk menghitung bobot reputasi akademik dilakukan dengan persamaan (8) berikut.

$$Bobot\ final = \frac{\sum (\theta^{(d_i)} \times bobot\ level(i))}{\sum\ topik\ keseluruhan} \quad (8)$$

Variabel $\theta^{(d_i)}$ merupakan probabilitas distribusi dari topik terhadap dokumen ke-i. Jumlah topik keseluruhan dalam penelitian ini adalah 9 topik berasal dari 9 level TKT. Setelah mendapatkan nilai bobot final maka nilai ini yang mencerminkan kualitas reputasi akademik perguruan tinggi. Untuk membandingkan hasil reputasi dan melihat peringkat akhir, perlu dilakukan pengurutan terhadap bobot final yang merepresentasikan kualitas reputasi akademik dari masing-masing perguruan tinggi.

IV. PRINSIP-PRINSIP PUBLIKASI

Percobaan dalam paper ini dilakukan dengan 450 paper penelitian. Abstrak paper ini diperoleh dari Google Scholar. Kami menggunakan sembilan TKT sebagai pedoman. Langkah pertama adalah melakukan tahap preprocessing teks ke dataset

abstrak. Pengurangan kata dalam dataset paper ditunjukkan pada Tabel 5.

Tabel 5.
Pengurangan Jumlah Kata Dataset

Nama Perguruan Tinggi	Kata Asli	Kata Hasil Preprocess
Institut Pertanian Bogor	10539	6293
Institut Teknologi Bandung	9287	5675
Institut Teknologi Sepuluh Nopember	9630	6015
Univeritas 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

Langkah kedua dengan memperbaharui ke 9 keyword corpus menggunakan kata-kata sinonim. Hasil langkah ini seperti pada Tabel 3. Langkah ketiga adalah mencocokkan kata-kata dalam setiap dataset dengan kata korpus berdasarkan semua tingkat sinonim. Tabel 6 menunjukkan jumlah kata dalam kata-kata yang cocok dengan dataset dengan kata kunci korpus berdasarkan tingkat semua sinonim.

Tabel 6.
Frekuensi Term dalam Corpus Keyword TKT

Perguruan Tinggi	Frekuensi Term dalam corpus TKT (kata)								
	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

Berdasarkan Tabel 6, frekuensi terjadinya kata-kata tertinggi berada pada 3 kelas teratas. Langkah selanjutnya setelah menentukan label topik ke semua dokumen maka akan dilakukan proses sPLSA dengan nilai probabilitas. Hasil dari probabilitas topik terhadap dokumen ditunjukkan pada Tabel 7.

Tabel 7.
Probabilitas Topik Terhadap Dokumen (PLSA)

Unive rsity	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
-----	---	---	---	---	-----------	---	-----------	---	----------

Tabel 8.

Pemerinkatan Universitas Berdasarkan 2 Hasil Percobaan dengan Indikator Reputasi Akademik					
Ra nk	Ground Truth	Experiment I		Experiment II	
		University	Score	Univer sity	Score
1.	ITB	ITB	2.974926	ITB	3.68801482
2.	UI	UI	2.978581	UI	3.68798061
3.	UGM	UI	2.978678	UGM	3.68795244
4.	UNAIR	UGM	2.978760	UMS	3.68795218
5.	IPB	UNPAD	2.978736	UNDIP	3.68795216
6.	UNDIP	ITS	2.978836	UB	3.68793780
7.	ITS	UNAIR	2.978934	ITS	3.68793103
8.	UMS	UPH	2.979156	UNAI	3.68790003
9.	UB	BINUS	2.978880	IPB	3.68771307

Pada percobaan pertama berdasarkan reputasi akademik perguruan tinggi didapatkan similarity ranking dengan ground truth 80,00%, perbedaan gap yaitu 10, dan toleransi perbedaan peringkat yaitu 88,88%. Pada percobaan kedua berdasarkan reputasi akademik perguruan tinggi didapatkan similarity ranking dengan ground truth 51%, perbedaan gap yaitu 16, dan toleransi perbedaan peringkat yaitu 66,66%.

Berdasarkan kedua percobaan diatas, didapatkan bahwa percobaan pertama memiliki nilai yang lebih bagus yang bisa dibandingkan dengan ground truth asli yaitu peringkat reputasi akademik pada QS World University Rankings 2016-2017.

Tabel 9.

Hasil Pemerinkatan Universitas Berdasarkan Keseluruhan Indikator		
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

Berdasarkan tabel IX diatas, hasil pemerinkatan dengan keseluruhan indikator dengan menggunakan sPLSA didapatkan similarity ranking dengan ground truth 63,33%, perbedaan gap yaitu 16, dan toleransi perbedaan peringkat yaitu 88,88%.

V. KESIMPULAN/RINGKASAN

Pada makalah ini, kami mengusulkan pendekatan baru untuk meningkatkan indikator penilaian di peringkat reputasi akademik perguruan tinggi. Pendekatan baru yang diusulkan yaitu dengan menggunakan metode sPLSA untuk pemodelan topik. Penggunaan metode sPLSA memerlukan ground truth analysis untuk dilakukan secara manual. Hal ini berguna sebagai patokan kinerja sistem. Berdasarkan hasil 2 percobaan yang telah dilakukan, persentase kesamaan peringkat dengan ground truth QS World University berada pada peringkat 78,00%. Hasil ini menunjukkan bahwa penggunaan metode sPLSA berhasil mencocokkan ranking dari QS World University Rankings.

DAFTAR PUSTAKA

- [1] E. X. and Q. T. K. S. Reddy, "Higher education, high-impact research, and world university rankings: A Case of India and comparison with China," *Pacific Sci. Rev. B Humanit. Soc. Sci.*, vol. 2, no. 1, pp. 1–21, 2016.
- [2] T. G. Tatiana Sidorenko, "Efficiency of Russian Education Through The Scale of World University Rankings," *Procedia - Soc. Behav. Sci.*, vol. 166, pp. 464–467, 2015.
- [3] R. van der W. Anne Wil K. Harzing, "Google Scholar As a New Source for Citation Analysis," *Ethics Sci. Environmental Polit.*, vol. 8, no. 1, pp. 61–73, 2008.
- [4] P. Serdyukov, "Department of Teacher Education, School of Education," *J. Res.*, vol. 320, no. 1, pp. 10–11, 2014.
- [5] dan P. T. R. I. Menteri Riset, Teknologi, *Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 42 Tahun 2016 Tentang Pengukuran dan Penetapan Tingkat Kesiaapterapan Teknologi*. Jakarta: Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia, 2016.
- [6] A.-P. Pavel, "Global University Rankings - A Comparative Analysis," *Procedia Econ. Financ.*, vol. 26, pp. 54–63, 2015.
- [7] M.-H. Huang, "Opening The Black Box of QS World University Rankings," *Res. Eval.*, vol. 21, no. 1, pp. 71–78, 2012.
- [8] S. N. Jin Wang, Ping Liu, Mary F.H. She, Abbas Kouzani, "Supervised learning probabilistic Latent Semantic Analysis," vol. 1, no. 1, p. 139, 2012.
- [9] e. W. F. and a. L. h. K. Landauer, "An Introduction to Latent Semantic Analysis," vol. 1, no. 1, pp. 259–284.
- [10] R. S. Bagus Setya Rintyarna, "Adapted weighted graph for Word Sense Disambiguation," *Int. Conf. Inf. Commun. Technol.*, 2016.
- [11] F. E. and D. G. Ferilli.S., "Automatic learning of linguistic resources for stopword removal and stemming from Text," *Procedia Comput. Sci.*, vol. 38, no. 1, pp. 116–123, 2014.
- [12] T. Verma, "Tokenization and Filtering Process in RapidMiner," *Int. J. Appl. Inf. Syst. – ISSN 2249-0868 Found. Comput. Sci. FCS*, vol. 7, no. 2, pp. 16–18, 2014.
- [13] L. C. and D. D. D. Blei, "Probabilistic topic models," *IEEE*.