PENERAPAN *SOCIAL NETWORK ANALYSIS* DAN *LATENT DIRICHLET ALLOCATION* UNTUK PEMETAAN PUBLIKASI PENELITIAN DOSEN POLITEKNIK STATISTIKA STIS

Muhammad Iqbal[1] dan Setia Pramana[1,2*]

[1]Politeknik Statistika STIS, [2] Direktorat Analisis dan Pengembangan Statistik, Badan Pusat Statistik
*e-mail: setia.pramana@stis.ac.id

Abstrak

Penelitian merupakan salah satu tugas utama institusi pendidikan, dimana dosen adalah motor dalam pengembangan penelitian. Pemetaan dosen berdasarkan pola kolaborasi serta topik yang dikaji dibutuhkan sebagai salah satu dasar untuk pengembangan kualitas dan kuantitas penelitian. Penelitian ini bertujuan untuk menganalisis pola kolaborasi antar dosen serta topik utama dari hasil penelitian dosen Politeknik Statistika STIS. Analisis penelitian ini menggunakan metode *Social Network Analysis* (SNA) dan *Latent Dirichlet Allocation* (LDA). Dari hasil analisis terlihat pola hubungan kolaborasi penelitian antar dosen serta dosen dengan posisi sentral pada proses kolaborasi tersebut. Kemudian dari hasil LDA terlihat beberapa topik-topik utama dari publikasi ilmiah seperti implementasi analisis regresi, kemiskinan, serta analisis data panel dalam bidang ekonomi.

Kata kunci: Publikasi ilmiah, *Social Network Analysis*, *Latent Dirichlet Allocation*

Abstract

Research is one of main function of an educational institution where lecturer plays important roles. Lecture mapping based on their collaboration and research topic is neede for the institution and stakeholders to improve the quantity and quality of the research. This study aims to analyze the scientific publications of Politeknik Statistika STIS lecturers to see lecturer collaboration and the themes of the research. The analysis uses the Social Network Analysis (SNA) and Latent Dirichlet Allocation (LDA) methods. The results shows a network graphic illustrating how research collaboration relationship between lecturers and also identify lectures with central position in the collaboration. The LDA graph shows several main topics from scientific publications such as regression analysis, poverty, and panel data analysis for economics.

***Keywords**: Scientific publications, Social Network Analysis, Latent Dirichlet Allocation*

PENDAHULUAN

Pendidikan Tinggi adalah jenjang pendidikan setelah pendidikan menengah yang mencakup program diploma, program sarjana, program magister, program doktor, dan program profesi, serta program spesialis, yang diselenggarakan oleh perguruan tinggi berdasarkan kebudayaan bangsa Indonesia. Satuan pendidikan yang melaksanakan pendidikan tinggi adalah Perguruan Tinggi. Berdasarkan Tridharma Perguruan Tinggi, tugas dosen sebagai Sivitas Akademika ada tiga, yaitu pendidikan, penelitian, dan pengabdian kepada masyarakat. Pembelajaran dan penelitian merupakan pilar utama dalam pendidikan tinggi. Pemerintah Indonesia telah mengeluarkan sejumlah kebijakan peraturan untuk meningkatkan minat para peneliti dalam mempublikasikan hasil penelitiannya, baik secara nasional maupun internasional. Melalui Peraturan Menteri Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Nomor 17 tahun 2013 tentang Jabatan Fungsional Dosen dan Angka Kreditnya, pemerintah mewajibkan dosen yang ingin memperoleh jabatan akademik Asisten Ahli, atau kenaikan jabatan dari Asisten Ahli ke Lektor, atau dari Lektor Kepala untuk melakukan penelitian dan publikasi ilmiah.

Pemetaan dosen berdasarkan hasil penelitiannya baik dari sisi topik serta bagaimana berkolaborasi antar dosen sangat diperlukan untuk meningkatkan kualitas serta kuantitas penelitian di suatu institusi pendidikan. Pemetaan ini sendiri masih belum tersedia di Politeknik Statistika STIS dan dibutuhkan oleh Pusat Penelitian dan Pengabdian Masyarakat (P3M) Politeknik Statistika STIS untuk mengelompokkan dosen berdasarkan kolaborasi penelitiannya dan memetakan kompetensi dosen berdasarkan konten tulisan, sehingga penelitian ini berfokus pada analisis kolaborasi penelitian dan tema dari penelitian yang dilakukan dosen Politeknik Statistika STIS. Metode yang digunakan untuk menentukan kolaborasi dosen dan tema dari penelitian tersebut adalah *Social Network Analysis* (SNA) dan *Latent Dirichlet Allocation* (LDA).

Social Network Analysis (SNA) adalah satu alat untuk memetakan hubungan pengetahuan penting antara individu (Pryke, 2004). Dengan menggunakan SNA terhadap hubungan sosial dapat diketahui ukuran hubungan antar aktor dalam hubungan tersebut, seperti seberapa sering mereka berinteraksi, siapa yang menjadi pusat dari hubungan tersebut, dan lain-lain. *Latent Dirichlet Allocation* (LDA) menurut David M. Blei, dkk (2003) adalah model probabilistik generatif dari korpus. LDA memungkinkan set pengamatan bisa dijelaskan oleh kelompok yang tidak teramati untuk menjelaskan kemiripan beberapa bagian data. Penggunaan model LDA memungkinkan untuk mendapatkan topik-topik umum dari suatu korpus dokumen.

Dengan menggunakan metode SNA pada publikasi ilmiah dosen bisa dihasilkan sebuah grafik jaringan yang memvisualisasikan kolaborasi penelitian antar dosen dan menampilkannya datanya secara informatif, dan didapatkan juga ukuran sentralistis dosen untuk mengetahui pengaruh dosen pada jaringan kolaborasi penelitian. Dan dengan menggunakan metode LDA pada publikasi ilmiah dosen akan didapatkan topik-topik umum dari publikasi ilmiah dosen yang dapat digunakan untuk mengelompokkan publikasi berdasarkan topik-topik tersebut. Hasil penelitian ini diharapkan bermanfaat untuk mempermudah P3M Politeknik Statistika STIS untuk memetakan kolaborasi dosen dan tema penelitian.

Dengan meningkatnya jumlah publikasi ilmiah dosen Politeknik Statistika STIS dari tahun ke tahun menyebabkan tidak memungkinkan untuk melakukan pemetaan dosen berdasarkan publikasi ilmiah secara manual. Pemetaan dosen berdasarkan kolaborasi penelitian dan tema publikasi ilmiah dengan tanpa pengawasan (*unsupervised*) masih belum dilakukan di Politeknik Statistika STIS. Tujuan dari penelitian ini adalah untuk menganalisis kolaborasi dosen dan tema publikasi tanpa

perlu adanya pengawasan menggunakan metode SNA dan LDA.

Terdapat beberapa penelitian terkait dengan penelitian ini. Penelitian terkait pertama adalah penelitian yang dilakukan oleh Hennie Tuhuteru dan Ade Iriani (2018) dengan judul "Analisis Kolaborasi Penelitian Dosen Fakultas X dengan Social Network Analysis (SNA)". Penelitian terkait selanjutnya adalah penelitian yang dilakukan oleh I Made Kusnanta Bramantya Putra (2017) dengan judul "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)".

Pada penelitian Henni (2018), dosen divisualisasikan berdasarkan jenis kelamin, *node* pria berbentuk lingkaran dan *node* perempuan berbentuk kotak. Pewarnaan *node* dilakukan berdasarkan atribut program studi dari masing-masing aktor. Dengan penghitungan didapatkan nilai *degree centrality*, *closeness centrality*, dan *betweenness centrality* dengan kesimpulan yang didapat adalah jurusan yang sering melakukan kolaborasi, jumlah kolaborasi, dan hubungan Jabatan Fungsional Akademik yang sering melakukan kolaborasi.

Sedangkan Putra (2017) melakukan penelitian pada pendapat publik terhadap stasiun radio dengan mengklasifikasikan pendapat-pendapat tersebut ke dalam beberapa topik. Dengan jumlah topik yang didapatkan adalah 4, diuji dengan mesin didapatkan nilai perplexity sebesar 231,4 dan diuji kemudahannya untuk diinterpretasi oleh manusia melalui uji koherensi topik yang terdiri dari *Word Intrusion Task* dan *Topic Intrusion Task.* Kesimpulan dari uji koherensi topik menyatakan bahwa model yang dihasilkan dengan metode LDA pada studi kasus ini dapat diinterpretasi manusia dengan baik.

METODOLOGI

1. Tinjauan Referensi

Social Network Analysis (SNA)

Menurut Lin Zhu, dkk (2019) *Social Network Analysis* (SNA) adalah sebuah proses yang digunakan untuk memeriksa struktur sosial dengan menggunakan teori jaringan dan grafik. Freeman (1989) menyatakan bahwa minimal basis data jaringan terdiri dari satu set objek (aktor atau *node*) yang dihubungkan oleh suatu hubungan pada suatu kejadian (*edge*). Representasi matriks yang umum digunakan untuk menggambarkan jaringan adalah matriks sosiometri atau matriks "dari siapa ke siapa". Jika jaringan sosial dimisalkan sebagai sktruktur sosial yang terdiri dari individu ataupun organisasi, Wasserman & Faust (1994) mengatakan bahwa *nodes* jaringan tersebut adalah individu atau organisasi tersebut, dimana *edgenya* bisa berupa hubungan seperti persahabatan, kekerabatan, kepentingan bersama, pertukaran uang, tidak sukaan, dan pengetahuan.

Berdasarkan Carrington, dkk (2005) SNA sudah digunakan sejak pertengahan tahun 1930-an dan mengalami berbagai macam perkembangan mencakup sosiometri, teori grafik, *dyads*, *triads*, *subsgroup*, blok model dan masih terus berkembang sampai tahun 2000-an. Ada berbagai macam metode untuk mengukur jaringan sosial, salah satunya metodenya adalah yang dikemukakan oleh Freeman (1979) yaitu dengan menggunakan metode analisis sentralistis. ia mengembangkan tiga konsep utama analisis sentralistis, yaitu *degree, closeness,* dan *betweenness.* Metode analisis sentralistis memungkinkan untuk menganalisis sentralistis berbagai jenis macam data, dari *one-mode* hingga *two-mode data*.

Pengukuran sentralistis yang digunakan adalah mengukur tiga kategori dasar sentralistis, yaitu *degree, closeness*, dan *betweenness*. Meskipun sudah banyak analisis sentralistis yang telah ditemukan, tiga ini mendominasi penggunaan empiris.

1. Degree Centrality

Shaw (1954) memperkenalkan ide untuk menggunakan *degree* sebagai indeks sentralitas. Misal pada jaringan komunikasi, seseorang yang berada pada posisi yang melakukan kontak langsung dengan orang banyak, maka orang tersebut bisa dianggap sebagai saluran informasi utama atau dengan kata lain titik fokus dari

komunikasi. Jadi *degree* melihat sentralitas berdasarkan banyaknya aktivitas yang dilakukan.

Nieminen (1974) merumuskan cara menghitung *degree* (C_D) suatu titik p_k:

$C_D(p_k) = \sum_{i=1}^{n} \quad a(p_i, p_k)$, dimana $a(p_i,p_k)$ = 1 jika dan hanya jika p_i dan p_k terhubung oleh garis, 0 jika tidak.

Semakin banyak jumlah titik yang terhubung maka semakin besar nilai *degree*, *degree* sama dengan 0 jika pk tidak terhubung dengan titik mana pun, jadi rumus ini berguna untuk menghitung jumlah aktivitas yang terhubung dari titik tersebut. Dengan menggunakan rumus *degree* bisa didapatkan *degree centrality* dengan:

$C_D'(p_k) = \frac{\sum_{i=1}^{n} \quad a(p_i,p_k)}{n-1}$, dimana n – 1 adalah nilai maksimal dari $C_D(p_k)$.

2. Betwenness Centrality

Bavelas (1948) dan Shaw (1954) mengemukakan bahwa ketika seseorang secara strategis berada di jalur yang menghubungkan sepasang orang lain, maka orang yang terletak pada posisi tersebut merupakan adalah pusat dari jalur tersebut. Misal pada jaringan komunikasi, orang yang berposisi sebagai penghubung komunikasi antara dua orang merupakan pusat karna orang tersebut bisa mengontrol komunikasi mereka. Jadi *betweenness* melihat sentralitas berdasarkan posisi pada jalur antara dua atau lebih *node* pada jaringan.

Cara mengukur *betweenness* dikemukakan oleh Anthonisse (1971) dan Freeman (1977) dengan cara menghitung jarak antara dua titik, dan melakukan perhitungan peluang jika jalurnya lebih dari satu. Rumus untuk menghitung *betweenness* yang dikemukan:

$B_{ij}(P_k) = \frac{g_{ij}(p_k)}{g_{ij}}$,

dimana

g_{ij} = banyak jalur terpendek antara titik p_i dan p_j

$g_{ij}(p_k)$ = banyaknya jalur terpendek antara titik p_i dan p_j yang mengandung titik p_k

dan rumus untuk menghitung *betweenness* secara keseluruhan:

$C_B(p_k) = \sum_{i=1}^{n} \quad \sum_{j=1}^{n} \quad B_{ij}(p_k)$,

dimana i<j, i ≠ j ≠ k, dan n banyaknya titik pada grafik. Setiap kali p_k menjadi titik yang hanya menghubungkan dua pasang titik maka nilai $C_B(P_k)$ bertambah 1.

Freeman (1977) menunjukkan bahwa rumus *betweenness* secara keseluruhan adalah:

$$C_B'(p_k) = \frac{2C_B(p_k)}{n^2 - 3n + 2}$$

Rumus $C_B`(P_k)$ pada suatu titik akan menghasilkan nilai 1 jika semua titik yang ada hanya terhubung ke titik tersebut, sehingga pada grafik itu $C_B`(P_k)$ akan menghasilkan nilai 0 jika dicoba pada titik yang lain.

3. Closeness Centrality

Closeness berkaitan dengan kontrol komunikasi untuk melihat sejauh mana pengaruh dari orang lain dapat dihindari, atau kata lainnya independensi. Berdasarkan Bavelas (1950), posisi non-sentral adalah orang yang harus menyampaikan informasinya melalui orang lain. Leavit (1951) berpendapat bahwa orang dalam posisi pusat dalam contoh kasus penyampaian informasi tidak terlalu bergantung ke orang lain dalam menyampaikan informasi, bahkan menurut Leavit kata sentralistis dan independen bisa ditukar.

Bermacam peneliti mencoba menjelaskan *closeness*, Bavelas (1948) menjelaskan bahwa informasi yang berasal dari titik pusat membutuhkan waktu yang minimal disampaikan ke seluruh titik pada jaringan. Beuchamp (1965) membandingkannya dengan organisasi seperti efisiensi, optimal, dll. dalam melakukan komunikasi. Hakimi (1965) dan Sabidussi (1966) mendefinisikannya secara umum yaitu titik tengah dalam melakukan komunikasi ke semua titik menggunakan biaya dan waktu yang paling minimum. Jadi secara berdasarkan uraian tersebut *closeness* adalah titik yang mempunyai jarak yang paling minimum untuk mencapai semua titik.

Sabidussi (1966) merumuskan cara menghitung *closeness* dengan cara menjumlahkan semua jarak terpendek yang menghubungkan suatu titik ke titik yang

lain. Rumus ini untuk mengukur desentralisasi titik atau invers *closeness:* $C_C(p_k)^{-1} = \sum_{i=1}^{n} d(p_i, p_k)$, dimana $d(p_i,p_k)$ = banyaknya jalur terpendek yang menghubungkan antara p_i dan p_k.

Semakin banyak jarak antara p_i dan p_k maka nilai $C_C(p_k)^{-1}$ semakin besar. $C_C(p_k)^{-1}$ akan bernilai tak hingga jika digunakan pada titik-titik yang tidak terhubung.

Beauchamp (1965) merumuskan *closeness* secara keseluruhan dengan rumus:

$$C'_C(p_k) = \frac{n-1}{\sum_{i=1}^{n} d(p_i, p_k)}$$

Pengaplikasian ketiga sentralistis tersebut tergantung dengan konteks aplikasi yang digunakan. Jika berkaitan dengan aktivitas komunikasi bisa dengan menggunakan pengukuran *degree*. Jika berkaitan dengan kontrol komunikasi bisa dengan menggunakan pengukuran *betweenness*. Jika berkaitan dengan independensi dan efisiensi bisa dengan menggunakan pengukuran *closeness.*

Latent Dirichlet Allocation (LDA)

Baeza-Yates dan Ribeiro-Neto (1999) mengusulkan metodologi dasar untuk pengambilan informasi dari korpus teks, sebuah metodologi yang sering digunakan dalam mesin pencarian, yaitu mereduksi setiap dokumen dalam korpus menjadi vektor bilangan real, yang masing-masing mewakili rasio jumlah dari kata.

Dalam skema *term frequency – invers document frequency* (tf – idf) (Salton dan McGill, 1983), kosaka yang digunakan untuk melambangkan unit dasar dari korpus adalah "kata" atau "istilah", dimana hitungannya dibentuk dari jumlah kemunculannya untuk setiap dokumen pada korpus. Setelah dinormalisasi, jumlah kata ini dibandingkan dengan invers jumlah dokumen, yang mengukur jumlah kemunculan kata di seluruh korpus. Hasil akhirnya adalah matrik X kata per dokumen yang kolomnya terdiri dari nilai tf-idf untuk masing-masing dokumen dalam korpus. Jadi dengan menggunakan skema tf-idf dokumen yang panjang mengecil menjadi daftar angka.

Meski skema tf-idf mempunyai beberapa kelebihan, pendekatan ini juga memiliki beberapa kelemahan, yaitu pendekatan ini hanya menyediakan relatif kecil pengurangan dan mengungkapkan sedikit dari struktur statistik antar/dalam dokumen. Untuk mengatasi kekurangan ini beberapa peneliti telah mengusulkan beberapa metode pengurangan dimensionalitas lainnya, dan kebanyakan metode ini berdasarkan pada asumsi "*bag of word*", yaitu urutan dari kata dalam dokumen diabaikan. Dan meski kurang dinyatakan secara formal, metode ini juga mengasumsikan urutan dokumen dalam korpus dapat diabaikan. Asumsi dapat dipertukarkan tidak berarti sama dengan asumsi variabel acak yang berdistribusi independen dan identik.

De Finetti (1990) menetapkan bahwa setiap kumpulan variabel acak yang dapat dipertukarkan memiliki representasi sebagai distribusi campuran, yang secara umum campuran yang tidak terhingga. Jadi jika ingin mempertimbangkan representasi untuk dokumen dan kata yang dapat dipertukarkan maka perlu dipertimbangkan model campuran yang dapat menangkap pertukaran kata dan dokumen tersebut, alasan inilah terbentuknya *Latent Dirichlet*, untuk menangkap struktur statistikal intra-dokumen yang signifikan melalui distribusi campuran.

Istilah-istilah yang digunakan berupa:

- Kata merupakan unit dasar dari data diskrit, yang didefinisikan sebagai item dari kosakata yang diindeks dengan $\{1,...,V\}$. Kata-kata diwakili oleh vektor satuan dasar yang memiliki satuan tunggal sama dengan satu dan semua komponen lainnya sama dengan nol.
- Dokumen adalah urutan dari N kata-kata yang dinotasikan dengan $w = (w_1, w_2, ..., w_N)$, dimana w_N adalah kata ke n dalam urutan.
- Korpus adalah kumpulan M dokumen yang dinotasikan dengan $D = \{w_1, w_2, ..., w_M\}$

Blei dkk (2003) menjelaskan LDA sebagai model probabilistik Bayesian tiga tingkat untuk sekumpulan data diskrit seperti korpus. Ide dasarnya adalah

dokumen merupakan campuran acak dari serangkaian topik yang tidak kelihatan, dimana setiap topik dibentuk dari distribusi berbagai kata tertentu.

LDA mengasumsikan proses generatif berikut untuk setiap *w* dokumen dengan *D* korpus:

1. Pilih $N \sim$ Poisson (ξ).
2. Pilih θ ~ Dir(α).
3. Untuk setiap *N* kata w_n :
   a. Pilih topik $z_n \sim$ Multinomial(θ).
   b. Pilih kata w_n dari $p(w_n | z_n, \beta)$, probabilitas multinomial yang dikondisikan pada topik z_n.

Simplistiknya rumus dilakukan dengan menggunakan asumsi:

1. Dimensionalitas k dari distribusi dirichlet dan dimensionalitas dari topik variabel z diketahui dan tetap.
2. Peluang dari kata diparameterkan dengan $k \times V\ matrix\ \beta$ yang diperlakukan sebagai jumlah yang tetap untuk diestimasi.
3. Asumsi Poisson dianggap tidak terlalu penting, karena N tersebut juga independen dari variabel penghasil data ($\theta\ dan\ z$).

Dengan menggunakan parameter α dan β, maka distribusi gabungan dari campuran topik ϕ, satu set N topik Z, dan satu set kata *W* adalah:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} \quad p(\theta)p(w_n|z_n, \beta),$$

dimana $p(z_n|\theta)$ sama dengan θ_i untuk i unik seperti z_n^i=1. Dengan mengintegrasikan θ dan menjumlahkan z, maka didapatkan distribusi marginal dari dokumen:

$$p(w|\alpha, \beta) = \int \quad p(\theta|\alpha)\left(\prod_{n=1}^{N} \quad \sum_{z_n} \quad p(\theta)p(w_n|z_n, \beta)\right)d\theta$$

Dengan mengambil hasil dari probabilitas marginal dari dokumen tunggal, maka didapatkan peluang korpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \quad \int \quad p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \quad \sum_{z_{dn}} \quad p(\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d$$

.

Perplexity

Perplexity adalah ukuran seberapa baik model probabilitas dalam menerima set data yang baru. Blei, dkk. (2003) menggunakan *perplexity* dari set uji untuk mengevaluasi model. Nilai *perplexity* secara monoton menurun dari kemungkinan uji data. Semakin rendah nilai *perplexity* menunjukkan kinerja generalisasi yang lebih baik. Berdasarkan penelitian tersebut, dengan M dokumen uji, *perplexity* adalah: $Perplexity(D_{test}) = exp\left\{\frac{\sum_{d=1}^{M} \quad logp(w_d)}{\sum_{d}^{M} \quad N_d}\right\}$.

Coherence

Sekumpulan pernyataan atau fakta dikatakan berkoherensi jika sekumpulan pernyataan atau fakta tersebut saling mendukung. Dalam filsafat ilmiah, pendekatan yang digunakan adalah dengan menggunakan gabungan fungsi dan probabilitas marginal yang terasosiasi dengan faktanya. Penggunaan koherensi pada topik model muncul karena pada topik model belum adanya jaminan interpretabilitas pada hasil topik model tersebut. Newman, dkk. (2010) mengusulkan langkah-langkah koherensi otomatis yang menilai topik terkait dengan pemahaman pada topik tersebut. Caranya dengan memperlakukan kata-kata pada topik sebagai fakta terus membatasi koherensi yang digunakan didasarkan pada perbandingan sepasang kata. Evaluasi pada penelitian tersebut adalah peringkat topik diukur berdasarkan pada co-occurence kata.

Aletra dan Stevenson (2013) memperkenalkan koherensi topik yang berdasar pada vektor konteks untuk setiap kata teratas pada topik. Vektor konteks dari kata w dibuat dengan menghitung co-occurence kata yang ditentukan menggunakan jendela konteks yang berisi semua kata yang terletak kurang lebih 5 token di sekitar kemunculan dari kata w. Element vektor didefinisikan menggunakan Normalized Pointwise Mutual Information (NPMI) untuk mendapatkan korelasi terbaik. Sehingga element ke-j pada vektor konteks v_i dari kata w_i mempunyai NPMI:

$$v_{ij} = NPMI\ (w_i, w_j)^{\gamma}$$
$$= \left(\frac{log\ \frac{P(w_i, w_j) + \epsilon}{P(w_i) + P(w_j)}}{-log\ (P(w_i, w_j) + \epsilon)} \right)^{\gamma}$$

Berdasarkan Roder, dkk. (2015) topik koherensi menggunakan vektor memiliki rata-rata hasil yang lebih baik dibandingkan topik koherensi yang lain.

Relevance

Bischof dan Ailordi (2012) mengajukan sebuah metode untuk memberikan peringkat terhadap istilah dari suatu topik tertentu baik dari segi frekuensi istilah di topik tersebut maupun eksklusivitas istilah tersebut pada topik, dengan memperhitungkan seberapa banyak istilah tersebut muncul pada topik tersebut dengan mengesampingkan topik lainnya. Siever dan Shirley (2014) mengajukan sebuah pengukuran yang sama dengan nama Relevance yang memungkinkan pengguna untuk menentukan peringkat istilah secara fleksibel untuk memudahkan pengguna dalam menafsirkan sebuah topik. Relevance diterapkan pada LDA karena secara umum topik yang dihasilkan susah untuk ditafsirkan manusia.

Berdasarkan penelitian Siever dan Shirley (2014), misalkan ϕ_{wk} menotasikan peluang istilah $w \in \{1, \dots, V\}$ untuk topik $k \in \{1, \dots, K\}$, dimana V menotasikan banyak istilah dalam kosakata, dan misalkan p_w menotasikan peluang marginal dari istilah w pada korpus. Rumus Relevance pada istilah w pada topik k dengan parameter bobot λ (dimana $0 \leq \lambda \leq 1$) adalah:

$$\Upsilon\ (w, k \mid \lambda) = \lambda\ log\ log\ (\phi_{kw}) + (1 - \lambda)\ log(\frac{\phi_{wk}}{p_w}),$$

dimana λ adalah bobot yang diberikan pada probabilitas suatu istilah w dalam suatu topik k relatif terhadap kenaikannya. Ketika λ = 1 maka istilah diberikan peringkat berdasarkan peluang istilah tersebut dalam topik, sedangkan ketika λ = 0 maka istilah tersebut hanya diperingkat berdasarkan liftnya, yaitu perbandingan antara probabilitas istilah pada topik tersebut dengan probabilitas marginal istilah pada korpus. Jadi ketika λ = 1 maka istilah yang umum pada topik peringkatnya akan tinggi, dan ketika λ = 0 maka istilah yang spesifik pada topik akan tinggi. Relevansi yang digunakan pada penelitian ini adalah 0,6.

2. Metode Analisis

Data yang digunakan dalam penelitian adalah data teks dari situs repositori Sistem Informasi Terpadu (SIPADU) Politeknik Statistika STIS. SIPADU merupakan sistem informasi yang dikembangkan dan diimplementasikan secara mandiri oleh para alumni STIS untuk mendukung kegiatan seluruh civitas akademik di STIS. Data yang diambil berupa nama, penulis, dan judul. Data tersebut dikumpulkan dengan metode *web scraping* menggunakan bahasa pemrograman Python dengan *package* ***selenium*** untuk mengambil elemen-elemen yang terdapat pada suatu halaman situs, sehingga dengan mengombinasikannya dengan algoritma pemrograman maka pemrograman bisa mengambil, mengolah, dan menyimpan data-data dari halaman repositori. Pendeteksian elemen pada situs menggunakan *full Xpath* agar bisa dilakukan algoritma pengulangan untuk mengambil data satu per satu seluruh data yang dibutuhkan. Data yang telah dikumpulkan sebanyak 652 data publikasi dengan data nama unik dosen penulis yang tersedia sebanyak 64 dosen Politeknik Statistika STIS. Dikarenakan tidak tersedianya data abstrak dan kata kunci untuk diambil menggunakan *scrapping* maka data yang digunakan untuk metode LDA hanya data judul publikasi, yang mungkin akan menyebabkan hasil yang didapat tidak sesuai dengan hasil yang diharapkan karena kekurangan data untuk diolah.

Data judul publikasi dan data nama penulis yang telah dikumpulkan kemudian dilakukan *preprocessing* seperti penghapusan tanda kurung, *case folding*, *remove punctuation*, *remove number*, *tokenization*, *remove stopword*, dan stemming. Untuk data nama penulis

dilakukan penghapusan tanda kurung, *case folding, remove punctuation, remove number, tokenization,* menghapus gelar menggunakan *stopword,* mengganti kata 'dan' dan 'and', dan terakhir dibuat variabel *source* menggunakan nama unik penulis dan variabel *target* berupa data nama yang telah dilakukan *preprocessing* untuk membentuk variabel baru yang dinamakan Concat yang merupakan gabungan dari dua variabel tersebut. Data yang telah diolah disimpan ke tabel DBMS MySQL menggunakan *package* ***PyMySQL***.

Untuk melakukan SNA digunakan *package* ***networkx*** metode *adjency list* pada variabel Concat akan menghasilkan data dalam bentuk SNA, yaitu tersedianya *node, edge, degree*, data ini dinotasikan dengan simbol G. Karena G masih mengandung nama-nama dosen di luar Politeknik Statistika STIS maka nama dosen selain dari PolStat STIS dikeluarkan. Variable dari G yang digunakan adalah *edge, degree, degree centrality, betweeness centrality,* dan *closeness centrality.* Untuk keperluan visualisasi agar lebih informatif dilakukan penghitungan per nama sebagai informasi jumlah publikasi.

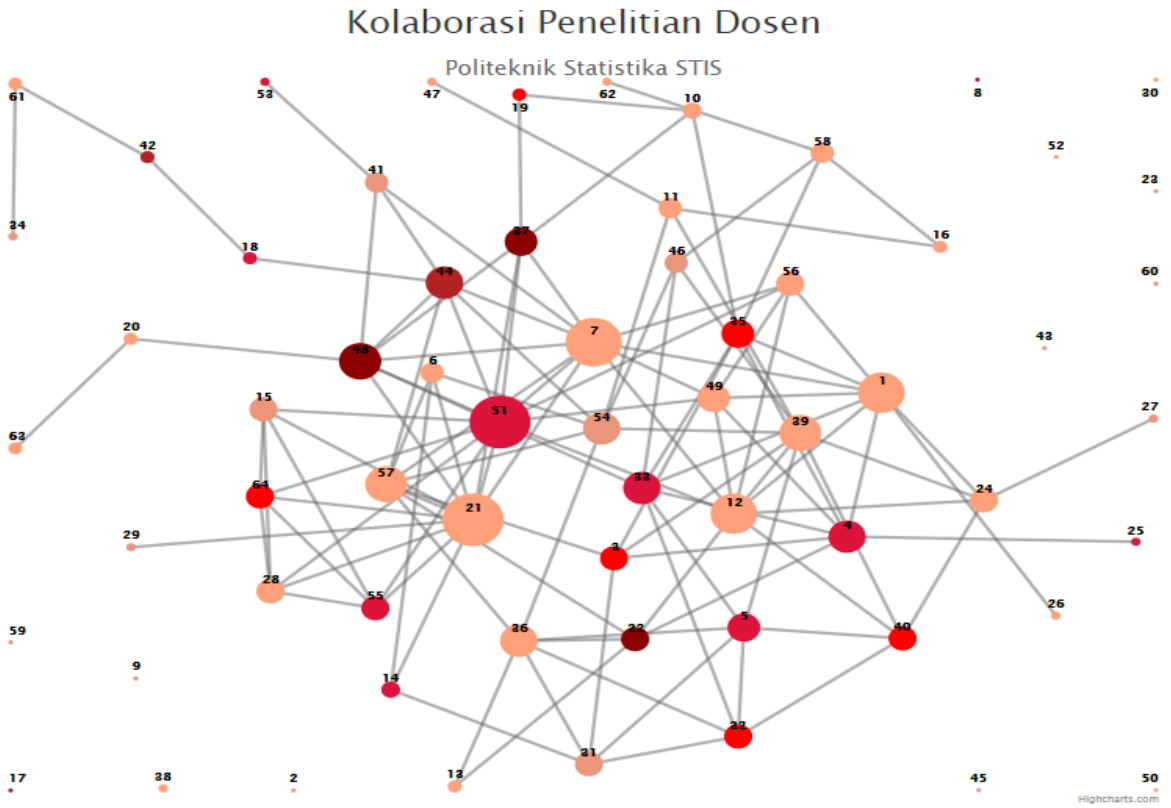
Kemudian LDA diimplementasikan pada data judul publikasi. Dari data judul publikasi tersebut dilakukan pembentukan *dictionary* dan *corpus* menggunakan *package* ***gensim***. *Dictionary* dan *corpus* akan digunakan untuk membentuk model LDA dengan jumlah topik yang ditentukan. P*ackage* ***pyLDAvis*** digunakan pada model untuk menghasilkan visualisasi grafik LDA.

Untuk menguji ketepatan model SNA dilihat dari ukuran sentralitasnya beruba *degree centrality, betweeness centrality,* dan *closeness centrality*. Visualisasi SNA dilakukan dengan menggunakan *JavaScripts Highcharts* untuk menghasilkan *network graph.*

Pengujian terhadap model LDA dilakukan dengan menentukan jumlah topik yang tepat untuk model. Pengujian ini dilakukan dengan menggunakan uji *perplexity* dan uji *coherence.* Jumlah topik yang baik adalah jumlah topik yang memiliki nilai uji yang tinggi dan efisien.

HASIL DAN PEMBAHASAN

Implementasi SNA

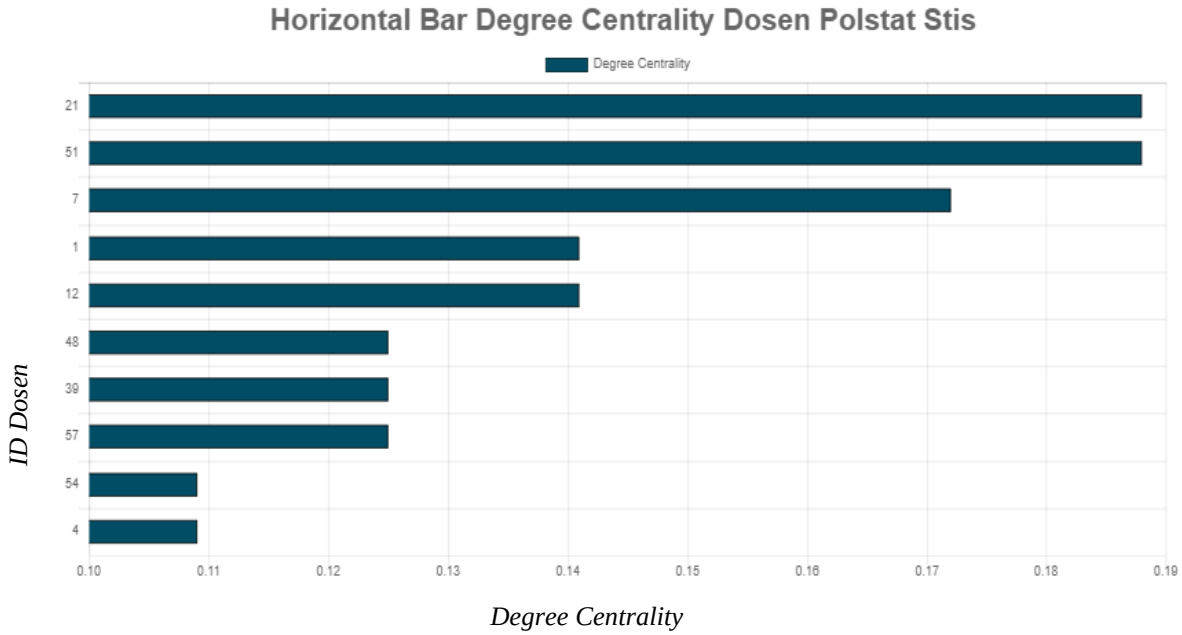


Gambar 1. Grafik jaringan kolaborasi penelitian dosen Politeknik Statistika STIS

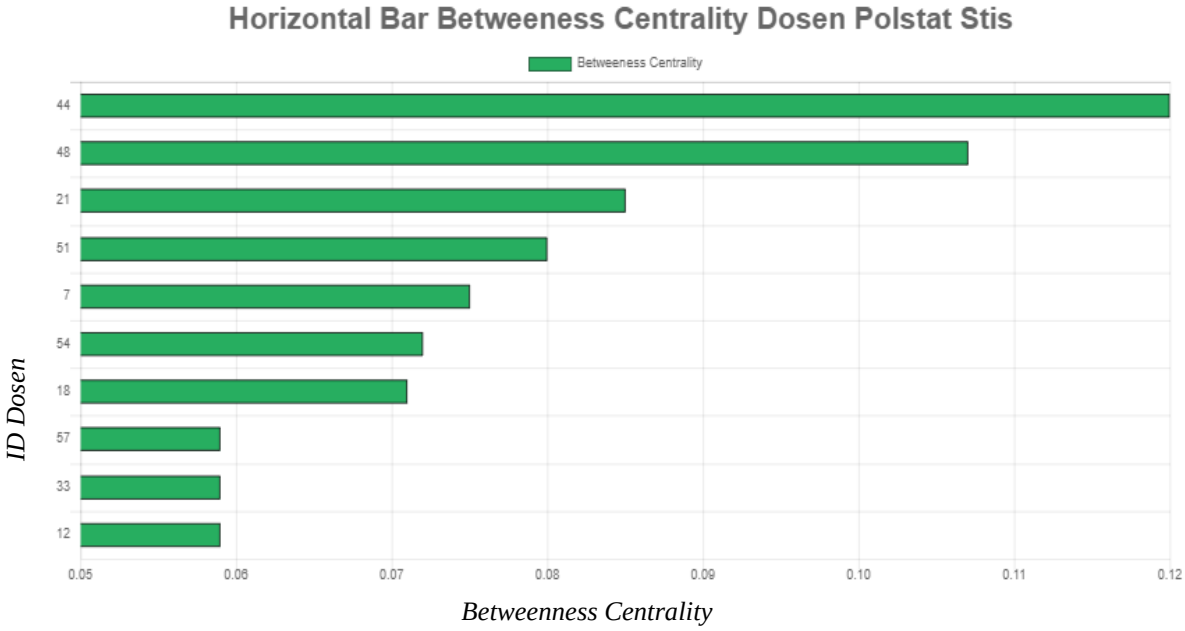
Dari hasil scraping didapat 652 publikasi ilmiah yang terambil dari SIPADU PolStat STIS. Implementasi SNA menggunakan package Python networkx untuk menghasilkan data dalam bentuk *node* dan *edge* menggunakan data dosen Politeknik STIS dari database yang telah melaporkan publikasinya. Data yang digunakan adalah nama dan penulis yang telah dilakukan preprocessing. Visualisasi SNA pada website menggunakan JavaScripts Highcharts untuk menghasilkan network graph seperti pada gambar 1.

Model SNA yang dihasilkan ditampilkan dalam Gambar 1. *Node* menunjukkan data dosen Politeknik Statistika STIS yang telah melaporkan publikasinya dan *edge* menunjukkan kolaborasi penelitian dengan dosen Politeknik Statistika STIS lainnya.

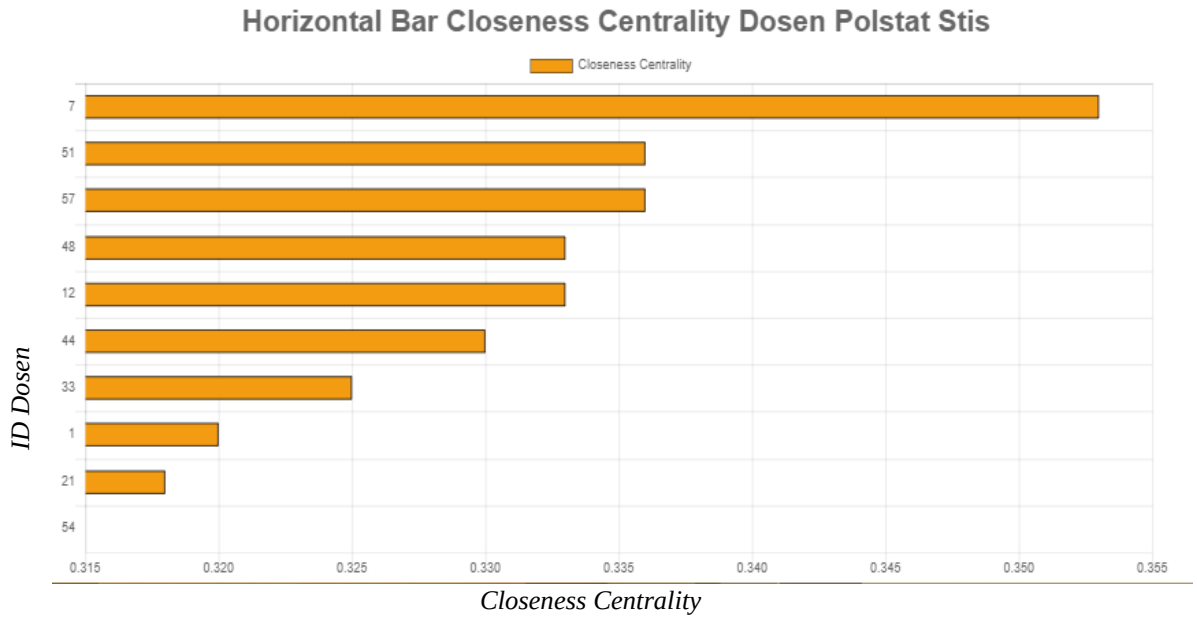
Semakin gelap warna pada *node* maka akan semakin tinggi rentang jumlah publikasinya. Ukuran lingkaran pada *node* menunjukkan seberapa banyak dosen yang pernah diikut sertakan dalam kolaborasi penelitian. Terlihat bahwa dosen dengan ID



Gambar 2. Horizontal bar *degree centrality* dosen Politeknik Statistika STIS



Gambar 3. Horizontal bar *betweenness centrality* dosen Politeknik Statistika STIS



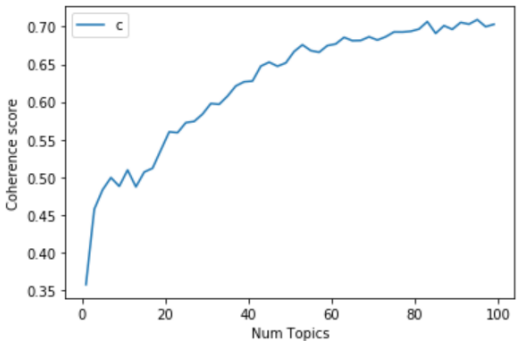
Gambar 4. Horizontal bar *closeness centrality* dosen Politeknik Statistika STIS

51 memiliki jumlah publikasi antara 21-25, dan memiliki jumlah koneksi/kolaborasi dengan dosen lain di lingkungan PolStat STIS paling banyak. Sedangkan dosen dengan jumlah publikasi terbanyak adalah dosen ID 48. Dosen ini terlihat tidak terlalu banyak memiliki jaringan/kolaborasi dengan internal dosen, kemungkinan dosen tersebut berkolaborasi dengan pihak luar Polstat STIS. Dosen ID 21 berkolaborasi cukup banyak dengan dosen lainnya, namun masih belum banyak memiliki jumlah publikasi.

Gambar 2, 3, dan 4 merupakan hasil dari metode evaluasi SNA menggunakan *package* ***Python.*** Gambar 2 menunjukkan bahwa dosen dengan kolaborasi terbanyak adalah dosen dengan ID 21 dan 51, yang kemudian terbanyak kedua adalah dosen ID 7. Sedangkan *betweenness centrality* yang ditampilkan di gambar 4 menunjukkan bahwa dosen yang menjadi penghubung terbanyak adalah dosen ID 44 diikuti oleh dosen ID 48. Berikutnya gambar 5 menunjukkan bahwa dosen yang menjadi pusat dari grafik adalah dosen dengan ID 7.

Implementasi LDA

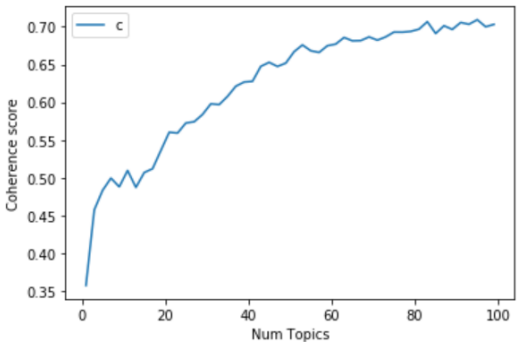
Evaluasi LDA dilakukan dengan cara mengecek nilai *perplexity* dan *coherence* dari model untuk menentukan jumlah topik yang sesuai.



Gambar 5. Uji *coherence* model LDA

Gambar 5 adalah hasil uji *coherence* dengan menggunakan modul ***coherenceModel*** dari *package* ***gensim***. Metode *coherence* yang digunakan adalah *coherence* vektor. Berdasarkan gambar, terjadi kenaikan yang signifikan jumlah titik 1 ke titik 11 yaitu sebesar 42,6%, dari *coherence* 0,3574 menjadi 0,5096. Meski dari titik 11 nilai *coherence* terus mengalami kenaikan, agar terjadi kenaikan yang signifikan harus meningkatkan jumlah topik yang secara signifikan juga. Sehingga jumlah topik minimal dengan *coherence* terbaik adalah 11. *Coherence* menunjukkan nilai interpretabilitas, semakin tinggi maka akan semakin gampang dipahami manusia topik, sehingga *coherence* dengan nilai 0,5 masih dikatakan belum baik.

Gambar 6 merupakan grafik uji *perplexity* menggunakan fungsi *log perplexity* dari modul ***LdaModel***. Semakin

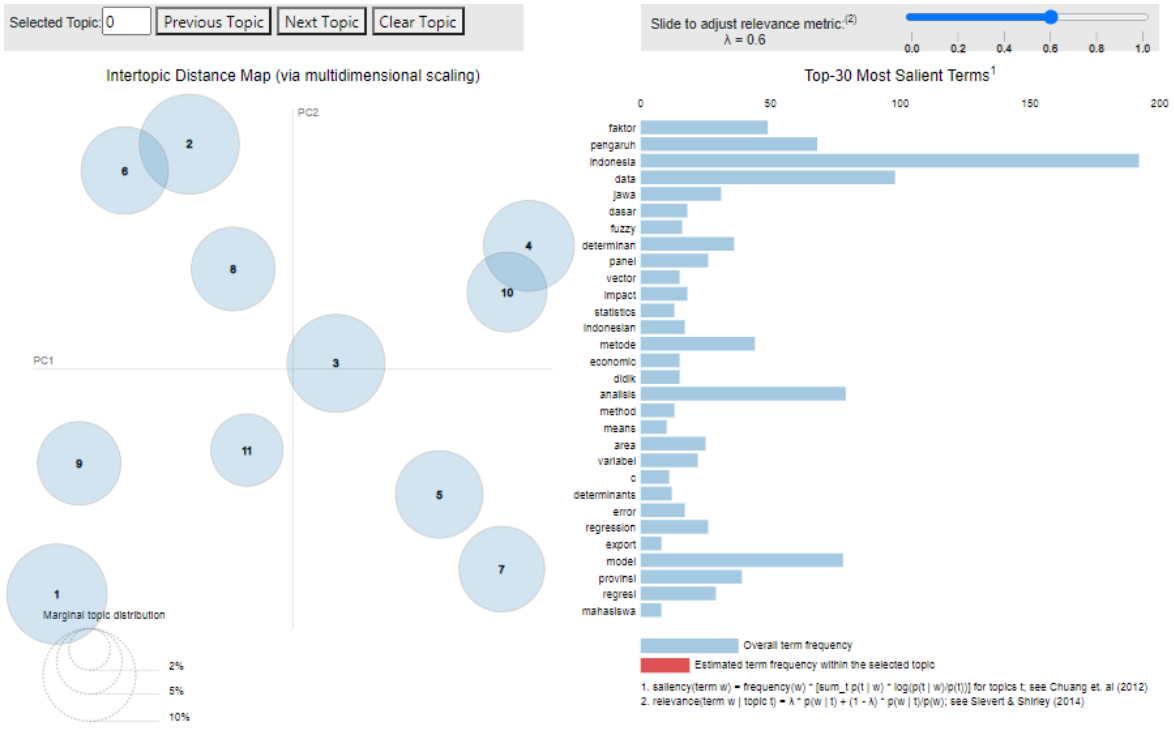


Gambar 6. Uji *perplexity* LDA model

kecil nilai *perplexity* menunjukkan bahwa model semakin baik dalam menangkap data. Dari gambar di atas terjadi penurunan signifikan hingga topik 17 dengan nilai perplexity sebesar -6.8179. Jumlah topik 15 dengan nilai *perplexity* sebesar -6.717 kurang bagus digunakan karena nilainya naik dan jumlah topik 19 dengan nilai *perplexity* sebesar -6.8173 juga kurang bagus karena nilainya tidak berbeda dengan jumlah topik 17. Tetapi dengan acuan uji *coherence* maka jumlah topik 11 dengan perplexity sebesar -6.6784 masih untuk digunakan.

Gambar 7 menampilkan grafik LDA judul publikasi dosen Politeknik Statistika STIS. Panel kiri membahas tentang topik yang dihasilkan. Terlihat pada gambar terdapat 11 lingkaran yang menggambarkan 11 topik umum judul publikasi. Jarak antar lingkaran menggambarkan hubungan antar topik, semakin dekat jaraknya maka topiknya memiliki kesamaan. Ukuran lingkaran menunjukkan prevalensi topik atau ukuran dari topik, semakin besar ukuran lingkaran maka semakin banyaknya *token* yang dicakupnya, urutan topik sendiri diurutkan berdasarkan prevalensi tersebut. Tiga topik utama yang didapatkan dari penelitian dosen Politeknik Statistika STIS adalah faktor-faktor tertentu yang memengaruhi variabel tertentu dalam analisis regresi, pembahasan mengenai data rumah tangga, pangan, dan kemiskinan di daerah Jawa Timur dan Nusa Tenggara Timur, dan penggunaan data panel untuk mengestimasi ekonomi di Indonesia.

Kontribusi penelitian ini berupa penerapan metode SNA dan LDA menggunakan *python* untuk mengetahui perkembangan penelitian yang dilakukan oleh dosen Politeknik Statistika STIS, baik dari topik maupun kolaborasi penelitian mereka. Untuk pendekatan dengan SNA merupakan penyederhanaan kolaborasi nama penulis sehingga hanya nama dosen dari Instansi tertentu yang tercakup dan



Gambar 7. Grafik LDA judul publikasi dosen Politeknik Statistika STIS

penambahan informasi tertentu pada grafik jaring seperti jumlah publikasi yang dilakukan. Sedangkan dalam implementasi LDA, terdapat keterbatasan dikarenakan kekurangan data untuk diolah menyebabkan pendefinisian topik tidak dapat dilakukan karena topik yang dihasilkan tidak membentuk sebuah kalimat atau topik.

KESIMPULAN

Berdasarkan pembahasan yang telah disampaikan sebelumnya, maka peneliti menyimpulkan sebagai berikut:

1. Pengimplementasian SNA terhadap data publikasi ilmiah dosen Politeknik Statistika STIS berhasil menghasilkan dan menampilkan kolaborasi penelitian. Dengan menggunakan pengukuran sentralitas dapat diidentifikasi dosen-dosen yang berposisi sebagai sentral dalam kolaborasi penelitian antar dosen.
2. Pengimplementasian LDA terhadap publikasi ilmiah dosen Politeknik Statistika STIS telah menampilkan beberapa topik utama pada publikasi ilmiah dosen berdasarkan distribusi kata pada judul publikasi.

SARAN

Disebabkan keterbatasan penelitian ini, peneliti mengusulkan beberapa saran sebagai berikut untuk pengembangan berikutnya:

1. Melakukan pengecekan manual yang lebih mendalam terhadap nama penulis untuk menemukan nama dosen yang salah di dalam nama penulis, dengan menggunakan metode *replace* nama yang salah tersebut bisa digantikan dengan nama yang sesuai standar yang digunakan.
2. Melakukan penerjemahan terhadap judul yang berbahasa Inggris ke dalam Bahasa Indonesia agar tidak terdapat topik yang terbentuk karena kesamaan bahasa yang digunakan.
3. Melakukan penurunan jumlah topik jika model yang digunakan masih sama, karena berdasarkan hasil penelitian ini penggunaan jumlah topik 11 menghasilkan beberapa topik yang tidak bagus.
4. Menggunakan data abstrak dan kata kunci dari publikasi ilmiah dosen untuk diterapkan metode LDA agar datanya lebih banyak dan variatif

DAFTAR PUSTAKA

Avasarala, S. 2014. *Selenium Webdriver Practical Guide.* Birmingham: Packt Publishing ltd.

Bavelas, A. 1948. A Mathematical Model for Group Structures. Human Organization.

Beauchamp, M.A. 1965. An improved index of centrality. Syst. Res.

Blei, D. M, Ng, A.Y, & Jordan, M.I. 2003. *Latent Dirichlet Allocation*. Berkeley: University of California.

Carrington, P., Scott, J., & S. Wasserman. (Eds.). 2005. *Models and Methods in Social Network Analysis* (Structural Analysis in the Social Sciences). Cambridge: Cambridge University Press.

de Finetti, B. 1990. Theory of Probability (A critical introductory treatment). Wiley, New York.

Dvorski, D. D. 2007. Installing, Configuring, and Developing with Xampp. Ontario: Skills Canada.

Everett, M and Borgatti, S.P. 2003. *Extending Centrality.* London: University of Westminster.

Feng, Jun, dkk. 2019. Product Feature Extraction via Topic Model and Synonym Recognition Approach. Nanjing: Hohai University.

Freeman, L. C. 1979. Centrality in Social Networks Conceptual Clarification. Lehigh University.

Haghberg, A. A. dkk. 2008. *Exploring Network Structure, Dynamics, and Function using NetworkX.* Los Alamos: Los Alamos National Laboratory.

Hakimi, S. 1965. Optimal Distribution of Switching Centers in a Communication Network and Some Related Theoretic Graph Theoretic Problems. Evanston, Illinois

Hunt, J. 2019. *Advanced Guide to Python 3 Programming.* Cham: Springer Nature Switzerland AG.

Anthonisse, J.M. 1971. The Rush in a Directed Graph , penerbit Stichting Mathematisch Centrum. Mathematische Besliskunde. Amsterdam, Netherlands

Leavitt, H. J. 1951. Some effects of certain communication patterns on group performance. The Journal of Abnormal and Social Psychology

Loper, E & Bird, S. 2002. *NLTK: The Natural Language Toolkit.* Philadelphia: University of Pennsylvania.

Putra, I. M. K. B. 2017. Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). Surabaya: Institut Teknologi Sepuluh Nopember.

Baeza-Yates, R., dan Ribeiro-Neto, B. 1999. Modern Information Retrieval: The Concepts and Technology behind Search

Roder, M., Both, A., dan Hinneburg, A.. 2015. *Exploring the Space of Topic Coherence Measures.* Leipzig: Leipzig University.

Sabidussi, G. 1996. The centrality index of a graph. *Psychometrika* **31.**.

Shaw, M. E. 1954. Some effects of unequal distribution of information upon group performance in various communication nets.

Sievert, C & Shirley K.E. 2014. *LDAvis: A method for visualizing and interpreting topics*. Ames: Iowa State University.

Suehring, S. 2002. *MySQL Bible.* New York: Wiley Publishing, Inc.

Pryke, S.D .2004. Analysing construction project coalitions: exploring the application of social network analysis, Construction Management and Economics, 22:8, 787-797, DOI:

Tuhuteru, H., dan Iriane, A. 2018. Analisis Kolaborasi Penelitian Ilmiah Dosen Fakultas X dengan Social Network Analysis (SNA). Salatiga: Universitas Kristen Satya Wacana.

Wasserman, S., dan Faust, K. 1994. Structural analysis in the social sciences.Social network analysis: Methods and applications. Cambridge University Press.

Zhu, L., Menzies, N.A., Wang, J. *et al.* 2020. Estimation and correction of bias in network simulations based on respondent-driven sampling data. *Sci Rep* 10.