TESIS – KI142502

EKSTRAKSI FITUR BERDASARKAN TOPIK PADA ARTIKEL ILMIAH UNTUK PENGELOMPOKAN POTENSI PENULIS DALAM JARINGAN KOLABORASI DINAMIS

Amelia Sahira Rahma
NRP. 5116201024

DOSEN PEMBIMBING
Dr.Eng. Chastine Fatichah, S.Kom., M.Kom.
NIP. 197512202001122002

PROGRAM MAGISTER
DEPARTEMEN INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

LEMBAR PENGESAHAN

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M.Kom.)
di
Institut Teknologi Sepuluh Nopember Surabaya

oleh:
AMELIA SAHIRA RAHMA
NRP. 5116201024

Dengan judul :
Ekstraksi Fitur Berdasarkan Topik pada Artikel Ilmiah untuk Pengelompokan Potensi Penulis dalam Jaringan Kolaborasi Dinamis

Tanggal Ujian : 24 Juli 2018
Periode Wisuda : September 2018

Disetujui oleh:

Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
NIP. 197512202001122002 ..... (Pembimbing 1)

Prof. Ir. Handayani Tjandrasa, M.Sc., Ph.D.
NIP. 194908231976032001 ..... (Penguji 1)

Dr. Agus Zainal Arifin, S.Kom., M.Kom.
NIP. 197208091995121001 ..... (Penguji 2)

Dr. Eng. Nanik Suciati, S.Kom., M.Kom.
NIP. 197712172003121001 ..... (Penguji 3)

Dekan Fakultas Teknologi Informasi dan Komunikasi,

Dr. Agus Zainal Arifin, S.Kom., M.Kom.
NIP. 197208091995121001

EKSTRAKSI FITUR BERDASARKAN TOPIK PADA ARTIKEL ILMIAH UNTUK PENGELOMPOKAN POTENSI PENULIS DALAM JARINGAN KOLABORASI DINAMIS

Nama mahasiswa	:	Amelia Sahira Rahma
NRP	:	5116201024
Pembimbing	:	Dr. Chastine Fatichah S.Kom., M.Kom.

ABSTRAK

Artikel ilmiah merupakan dokumentasi yang memaparkan hasil suatu metode usulan untuk menyelesaikan sebuah permasalahan dalam bidang tertentu. Publikasi artikel ilmiah memiliki kontribusi penting bagi penulis atau *author*, baik dalam hal karir, kemampuan bekerjasama, maupun tolak ukur kinerja, dan potensi *author*. Terdapat banyak metode untuk pengelompokan potensi *author* yang telah diusulkan sebelumnya, diantaranya perangkingan, klasifikasi, dan *clustering author*. Setiap metode mempertimbangkan beberapa aspek penilaian yang berbeda, yaitu jumlah produktifitas *author*, peringkat tempat publikasi artikel, kemampuan kolaborasi *author*, dan jumlah sitasi artikel. Mengingat seorang *author* mampu melakukan penelitian dalam beberapa bidang yang berbeda, sehingga terdapat dinamika kegiatan penelitian *author* yang terus berubah, dibutuhkan suatu metode yang dapat mempertimbangkan unsur topik penelitian untuk membedakan potensi penulis. Penelitian ini mengusulkan pendekatan baru dalam mengekstraksi beberapa fitur dengan menambahkan unsur topik pada artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis. Metode ini mengekstraksi informasi *author* dari beberapa aspek, diantaranya informasi produktifitas dan kolaborasi berdasarkan topik, serta informasi dinamika *author* untuk mengetahui perubahan kegiatan penelitian pada setiap periode. Selanjutnya fitur-fitur tersebut digunakan sebagai inputan dalam algoritma k-means++ *clustering* untuk pengelompokan *author* berdasarkan potensi dan bidang penelitiannya. Uji coba dilakukan pada 3.481 *author* dengan 296.341 artikel ilmiah. Berdasarkan hasil uji coba, penelitian ini menghasilkan 10 kelompok potensi *author* dengan 9 topik penelitian. Analisis empiris menunjukkan bahwa metode yang diusulkan telah membedakan penulis dengan baik, masing-masing kelompok mempertahankan perbedaan substansial satu sama lain dalam jumlah publikasi dan sitasi pada semua periode. Dapat disimpulkan bahwa penggunaan ekstraksi fitur berdasarkan topik yang diusulkan mampu mengelompokkan penulis dalam beberapa kelompok menurut potensi pada masing-masing topik penelitian.

Kata kunci : artikel ilmiah, ekstraksi fitur berdasarkan topik, pengelompokan potensi *author*, jaringan kolaborasi dinamis.

FEATURES EXTRACTION BASED ON TOPICS IN SCIENTIFIC ARTICLES FOR AUTHOR'S POTENTIAL CLUSTERING IN DYNAMIC COLLABORATION NETWORKS

Nama mahasiswa	:	Amelia Sahira Rahma
NRP	:	5116201024
Pembimbing	:	Dr.Eng. Chastine Fatichah S.Kom., M.Kom.

ABSTRACT

Scientific article is a documentation that describes the results of a proposed method to solve a problem in a particular field. Publication of scientific articles has an important contribution to the author, both in terms of career, ability to cooperate, as well as benchmark performance and potential author. There are many methods for grouping of potential authors that have been proposed before, such as ranking, classification and clustering author. Each method considers several different aspects of assessment, namely the amount of author's productivity, the ranking of the article publication, the author's collaboration capabilities and the number of article citations. Given that an author is able to conduct research in several fields, so that there is a dynamic research activity of the author that is constantly changing, it takes a method that can consider the topic of research to differentiate the potential of the author. This research proposes a new approach by doing feature extraction based on topic of scientific articles for grouping of potential authors in a dynamic collaboration network. This method extracts author information from several aspects, such as productivity and collaboration information by topic, as well as author's dynamic information to know the change of research activities in each period. Furthermore, these features are used as input in the k-means clustering algorithm for grouping the author based on the potential and field of research. Extensive experiments conducted on 3,481 authors with 296,341 scientific articles from AMiner dataset. The empirical analysis demonstrates that proposed method has distinguished authors well, each cluster maintain substantial differences with each other in number of publications and citations over time. It could be concluded that the used of proposed author potential detection is capable to detect several author potentials in respective topic of research

Keywords : scientific articles, feature extraction based on topic, author’s potential clustering, dynamic collaboration network.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT, atas segala rizki, berkah, nikmat serta karunia-Nya yang terlimpahkan kepada penulis, sehingga penulis akhirnya dapat menyelesaikan penelitian dengan judul "Ekstraksi Fitur Berdasarkan Topik pada Artikel Ilmiah untuk Pengelompokan Potensi Penulis dalam Jaringan Kolaborasi Dinamis".

Penulis juga ingin mengucapkan banyak terimakasih sepenuhnya kepada berbagai pihak, yang mana tanpa bantuan dari mereka penelitian ini tidak akan terselesaikan dengan hasil seperti sekarang ini. Oleh karena itu pada kesempatan ini penulis ingin menyampaikan terima kasih yang sebesar-besarnya, kepada :

1. Allah SWT atas limpahan nikmat iman, islam, kesehatan, waktu, serta berbagai kemudahan dari arah yang tidak pernah diduga sebelumnya, sehingga penulis dapat menyelesaikan penelitian ini dengan baik.
2. Ibu Failun Indrawati, Bapak Rindi Sudjono, Mbak Putri, Mbak Rati, Adik Aulia, dan segenap keluarga tercinta yang tidak hentinya memberikan dukungan materil, do'a tulus yang tidak ada henti-hentinya, serta semangat membara agar penulis harus terus senantiasa menjadi yang terbaik dan bermanfaat untuk orang sekitar dimanapun penulis berada.
3. Ibu Dr. Eng. Chastine Fatichah, S.Kom., M.Kom. dan Ibu Diana Purwitasari, S.Kom., M.Sc. selaku pembimbing yang telah banyak meluangkan waktu untuk dengan sangat sabar mendidik, membimbing, mengarahkan, menjadi teman diskusi, mendengar curhatan dan keluh kesah pengerjaan, serta mengajak penulis untuk berpikir lebih keras dalam menyelesaikan penelitian.
4. Ibu Prof. Ir. Handayani Tjandrasa, M.Sc., Ph.D., Bapak Prof. Dr. Agus Zainal Arifin, S.Kom., M.Kom., dan ibu Dr. Eng. Nanik Suciati, S.Kom., M.Kom. selaku dosen penguji yang telah memberikan banyak saran dan arahan agar penulis mampu lebih baik dalam menyelesaikan penelitian.
5. Bapak Waskitho Wibisono, S.Kom., M.Eng., Ph.D., selaku Kaprodi S2 Teknik Informatika ITS Surabaya yang memfasilitasi mahasiswanya untuk belajar di Lab S2 hingga larut dalam rangka menyelesaikan penelitian.

6. Seluruh staf dosen, staf tata usaha dan karyawan perpustakaan Departemen Informatika, Institut Teknologi Sepuluh Nopember (ITS), Surabaya.
7. Tim Penyelamat; Mbak Alif dan Mas Adhi yang bersedia mengerahkan segenap tenaga dan pikiran, serta senantiasa siap siaga 24 jam dan rela menginap selama beberapa hari di kampus hanya untuk membantu penulis dalam menyelesaikan penelitian ini dalam bentuk apapun.
8. Genk jajan asal Sidoarjo; Mbak Eva dan Ilmi yang senantiasa mengingatkan makan dan melupakan diet, memberikan inovasi dan masukkan akan makanan baru diberbagai wilayah setiap harinya, sehingga penulis tidak pernah kekurangan gizi dalam mengerjakan penelitian ini.
9. Sahabat tercintah; Ozzy, Kak Yaya, Mak Ita, Mbak Vynska, Mbak Pipit, Mbak Ulum, Mbak Myrna, Kak Herna, Kak Udis, Mbak Dian, yang senantiasa kuat untuk menguatkan, bahagia untuk membahagiakan, serta semangat untuk menyemangati dalam menyegerakan tesis, ibadah, juga makan tepat waktu.
10. Kawan-kawan S2 Departemen Informatika, Institut Teknologi Sepuluh Nopember (ITS), Surabaya atas bantuan dan diskusi selama penelitian.
11. Semua pihak yang tidak dapat dituliskan satu per satu oleh penulis, terima kasih banyak atas doa dan dukungannya.

Semoga Allah SWT senantiasa menyayangi, menguatkan, memampukan, dan menunjukan jalan yang terbaik atas semua kebaikan yang telah diberikan. Penulis menyadari bahwa laporan penelitian ini tentunya masih jauh dari kesempurnaan. Oleh sebab itu, saran dan kritik sangat diharapkan untuk perbaikan dimasa yang akan datang. Semoga laporan penelitian ini dapat bermanfaat bagi penulis dan pembaca pada umumnya.

” *Jangan pernah berhenti berharap ditengah ketidakmungkinan*”

Surabaya, Agustus 2018

Amelia Sahira Rahma

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

Pada bab pendahuluan ini akan dijelaskan mengenai beberapa hal dasar dalam pembuatan proposal penelitian yang meliputi latar belakang, perumusan masalah, tujuan, manfaat, kontribusi penelitian, dan batasan masalah.

1.1 Latar Belakang

Artikel ilmiah merupakan hasil dokumentasi dari proses pengumpulan data, perbandingan hingga pengevaluasian suatu metode usulan terhadap penelitian, dan pengamatan sebuah objek dalam bidang tertentu [1]. Terdapat berbagai macam artikel ilmiah baik yang telah terpublikasi maupun tidak terpublikasi, bertaraf nasional hingga internasional yang telah diciptakan, diantaranya berupa: resensi, referat, skripsi, *thesis*, disertasi, *paper conference* hingga jurnal internasional [2].

Penulis artikel ilmiah atau sering disebut *author* tentu saja memiliki satu atau beberapa bidang penelitian yang berbeda-beda, bergantung pada latar belakang pendidikan atau pekerjaan [2]. Selain itu, tidak menutup kemungkinan bahwa seorang *author* dapat bekerja sama dengan *author* lain dari latar belakang pendidikan yang berbeda, sehingga mereka mampu memperluas pengalaman ilmiah dan memperkaya bidang penelitian yang dimiliki sebelumnya [3]. Hal tersebut menyebabkan seorang *author* mampu menulis jurnal penelitian atau artikel ilmiah dalam bidang keilmuan yang beragam.

Pembuatan dan publikasi artikel ilmiah memiliki kontribusi yang sangat penting bagi penulisnya, diantaranya sebagai tolak ukur kinerja dan kualitas penelitian, meningkatkan kemampuan bekerja sama dengan *author* lain, bahkan berpengaruh terhadap perkembangan karir penulis. Dalam perkembangan karirnya, sangatlah penting bagi *author* untuk menunjukkan profil yang mencakup informasi tentang ringkasan riwayat kegiatan penelitian untuk memperlihatkan tingkat kesuksesan dan potensi dalam *author* tersebut.

Secara umum terdapat beberapa aspek yang dapat dipertimbangkan untuk mengukur potensi *author*, yaitu aspek produktifitas, aspek kolaborasi, aspek dinamika *author*, dan aspek topik. Aspek produktifitas didapatkan dari informasi jumlah publikasi artikel ilmiah, jumlah sitasi artikel, tahun publikasi, dan tempat publikasi artikel. Aspek kolaborasi didapatakan dari informasi *co-authorship* atau daftar *author* pada masing-masing artikel. Aspek dinamika *author* didapatkan dari hasil informasi pada aspek produktifitas dan aspek kolaborasi disetiap periode. Sedangakan aspek topik didapatkan dari informasi bidang penelitian yang terkandung pada masing-masing artikel ilmiah [4][5].

Penulis berpotensi atau sering disebut dengan *rising stars* adalah seorang *author* yang saat ini memiliki profil penelitian yang relatif rendah, namun memungkinkan pada akhirnya *autor* tersebut akan muncul sebagai peneliti yang terkemuka [6]. Sumber yang lain juga menyatakan bahwa *rising strars* adalah *author* yang menunjukkan peningkatan di dalam produktifitas dan dampak dari publikasi ilmiahnya sepanjang waktu [4].

Tantangan dalam mengidentifikasi *rising stars* adalah pendekatan prediksi yang dilakukan dalam jangka waktu yang panjang. Sangat penting untuk mengumpulkan data *author* dari awal karir dan mengamati perjalanan penelitian *author* selama beberapa periode, agar dapat melihat potensi *author* secara dini dan menyatakan bahwa mereka akan menjadi peneliti yang hebat dimasa depan [7]. Hal tersebut diharapkan dapat membantu departemen dalam pemilihan pengajar baru untuk meningkatkan produktifitas penelitian [8].

Terdapat banyak penelitian dalam mengidentifikasi *rising stars* pada jaringan akademik yang telah dikembangkan sebelumnya. Penelitian tersebut diantaranya berupa perangkingan *author* berdasarkan nilai dari sekumpulan indikator yang menunjukkan potensi *rising stars*, mengklasifikasikan *author* kedalam kategori *rising stars* dan *non-rising stars*, atau mengelompokkan *author* yang memiliki karakteristik serupa dengan metode *clustering*.

Pada penelitiannya, [6] mengusulkan metode untuk mengidentifikasi *rising stars* dengan perangkingan *author*, yang disebut algoritma PubRank. Penelitian ini mempertimbangkan 3 aspek dalam metodenya, yaitu 1) pengaruh antar *author* di dalam jaringan kolaborasi, menggunakan pembobotan dan *link* yang berarah untuk

setiap kolaborasi, 2) nilai rata-rata pada dampak publikasi penelitian, menggunakan pembobotan pada *node* dalam model *graph*, 3) perubahan yang terjadi pada jaringan kolaborasi secara berurutan.

Penelitian tersebut disanggah dalam penelitian [8] yang menyoroti dua kekurangan utama pada PubRank, yaitu mengabaikan kontribusi *author* di dalam publikasi dan perangkingan tempat publikasi yang dilakukan secara statis (dengan mengambil nilai rata-rata). Penelitian ini mencoba untuk menyelesaikan kedua kekurangan tersebut dengan metode usulannya, yaitu algoritma StarRank. Metode ini mempertimbangkan pengurutan nama *author* dalam artikel ilmiah dan menyesuaikan perangkingan tempat publikasi berdasarkan waktu, menggunakan entropi pada topik di dalam publikasi tersebut.

Kedua aspek tersebut tidak diterapkan pada semua disiplin. Pada aspek pertama, dalam beberapa kasus *co-author* mengurutkan nama mereka berdasarkan abjad, tetapi dikasus lain *author* senior selalu diletakkan di akhir urutan *co-author*. Aspek kedua juga sangat lemah karena algoritma tersebut memperlakukan perhitungan yang sama antara *paper conference* dan jurnal, yang seharusnya jurnal memiliki bobot yang lebih besar bagi karir *author* [7].

Mengingat seorang *author* mampu melakukan penelitian dalam beberapa bidang yang berbeda serta dinamika kegiatan penelitian *author* yang terus berubah, dibutuhkan suatu metode yang dapat mempertimbangkan unsur topik penelitian untuk membedakan potensi penulis. Pada penelitian tesis ini, penulis mengusulkan suatu pendekatan baru dengan melakukan ektraksi fitur berdasarkan topik pada artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis dengan mempertimbangkan aspek produktifitas, aspek kolaborasi antar *author*, dan aspek dinamika *author* setiap periode.

Di dalam metode ini akan dilakukan ekstraksi topik pada artikel ilmiah dengan cara *clustering* yang diolah dari judul dan abstrak. Selanjutnya fitur lainnya akan diekstraksi berdasarkan topik yang telah didapatkan sebelumnya, yaitu mengekstraksi informasi pada indikator *bibliometrical* berupa produktifitas *author* (jumlah publikasi *author*) dan sitasi yang diterima pada masing-masing artikel. Kemudian informasi tersebut akan digunakan untuk membuat *graph* kolaborasi atar *author (co-author)* dalam pembuatan artikel ilmiah. Langkah terakhir fitur yang

telah didapatkan dari ketiga aspek tersebut akan diamati perubahan setiap periodenya dalam jumlah periode tertentu untuk mengetahui dinamika kinerja kegiatan penelitian *author*, yang akan berfungsi sebagai fitur pada proses pengelompokan *author* dalam pengelompokan potensi penulis.

1.2 Perumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah sebagai berikut :

1. Bagaimana cara melakukan ekstraksi fitur berdasarkan topik pada artikel ilmiah yang mewakili informasi penting dari seorang penulis?
2. Bagaimana cara mengelompokkan potensi penulis dengan mempertimbangkan kombinasi dari aspek topik, produktifitas, kolaborasi, dan dinamika penulis?

1.3 Tujuan Penelitian

Tujuan yang akan dicapai dalam penelitian ini adalah melakukan ekstraksi fitur berdasarkan topik pada artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini diantaranya adalah menyajikan rangkuman evaluasi kinerja ilmiah penulis dan mengelompokkan penulis yang berpotensi dalam bidang penelitiannya melalui artikel ilmiah yang telah terpublikasi. Penelitian ini diharapkan dapat membantu menemukan seorang peneliti baru dengan potensi untuk menjadi peneliti handal dimasa depan serta membantu dalam pemilihan pengajar baru dalam rangka meningkatkan produktifitas penelitian departemen.

1.5 Kontribusi

Kontribusi pada penelitian ini adalah sebuah pendekatan baru dalam mengekstraksi beberapa fitur yang merepresentasikan performa penulis dengan menambahkan unsur topik penelitian pada publikasi artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis.

Tahap awal dan yang paling fundamental pada penelitian tesis ini adalah ekstraksi topik pada publikasi artikel ilmiah untuk mendapatkan bidang keahlian pada masing-masing *author*. Kemudian hasil ekstraksi topik penelitian tersebut digunakan sebagai acuan dalam mengekstraksi informasi *author* pada tiga aspek besar. Aspek pertama adalah mengambil informasi tentang jumlah publikasi dan sitasi yang telah dicapai oleh *author*, dimana penelitian ini mempertimbangkan dampak kontribusi publikasi *author* dari jumlah sitasi yang didapatkan pada masing-masing artikel ilmiah, bukan dari dampak kualitas tempat publikasi artikel tersebut. Aspek kedua adalah mempertimbangkan kemampuan *author* dalam bekerja sama dengan *author* lain, dimana faktor tersebut juga sangat berpengaruh terhadap penilaian potensi seorang penulis. Aspek ketiga adalah memonitor dan mengamati perubahan atau kedinamisan dari pekerjaan ilmiah *author* yang sangat penting untuk mengatasi kebiasan dalam *self-citation* dan jumlah kumulatif produktifitas. Sehingga dengan mengkombinasikan keempat aspek tersebut, metode usulan diharapkan dapat dengan baik membedakan *author* yang berpotensi dengan kelompok *author* yang lainnya.

1.6 Batasan Masalah

Mengingat permasalahan pengelompokan potensi penulis merupakan hal yang kompleks dan luas, penelitian ini memiliki beberapa batasan sebagai berikut :

1. Artikel ilmiah yang digunakan sebagai data uji coba merupakan artikel dalam bahasa inggris yang didapatkan dari arnetminer.
2. Data yang digunakan dalam penelitian ini adalah data artikel, produktifitas *author*, sitasi artikel, *co-author* (kolaborasi penulis), dan periode publikasi.

[Halaman ini sengaja dikosongkan]

BAB 2
KAJIAN PUSTAKA DAN DASAR TEORI

Pada bab kajian pustaka dan dasar teori ini akan dipaparkan teori-teori yang digunakan sebagai bahan acuan dalam melakukan penelitian. Adapun materi yang akan dibahas, meliputi konsep indikator *bibliometric*, teori *graph*, *preprosesing* teks, *term weighting*, *clustering*, metode evaluasi kluster, dan pelabelan kluster.

2.1 Indikator *Bibliometrics*

Indikator *Bibliometrics* adalah seperangkat metode matematika dan statistik yang digunakan untuk menganalisis dan mengukur kuantitas dan kualitas buku, artikel, dan bentuk publikasi lainnya. Ada tiga jenis indikator bibliometrik, yaitu :

a. **Indikator kuantitas**, yang mengukur produktivitas seorang peneliti.

b. **Indikator kualitas**, yang mengukur kualitas atau kinerja kegiatan penelitian dari output seorang peneliti.

c. **Indikator struktural**, yang mengukur hubungan antara publikasi, penulis, dan bidang penelitian yang sedang dikerjakan.

Indikator *Bibliometrics* sangat penting bagi para peneliti dan organisasi karena pengukuran ini sering digunakan dalam keputusan pendanaan dan promosi para peneliti. Semakin banyak penemuan-penemuan ilmiah dilakukan, kemudian hasil penelitian yang dipublikasikan dibaca atau dikutip oleh peneliti lain, indikator bibliometrik menjadi semakin penting bagi penulis tersebut [5].

2.2 Teori *Graph*

Pada bagian ini penulis akan membahas mengenai bentuk *graph* dalam memodelkan kerjasama kegiatan penelitian antar *author* yang digunakan pada penelitian ini, yaitu *Co-authorship Graph.*

Co-authorship graph adalah sebuah *graph* yang merepresentasikan hubungan antar penulis, dimana *node* merepresentasikan penulis dan *edge* sebagai hubungan antar penulis. Secara umum *co-authorship graph* digunakan untuk menggambarkan kolaborasi atau kerjasama antar penulis, apabila terdapat dua penulis yang pernah menulis bersama, maka pada *co-authorship graph* keduanya

akan dihubungkan. Begitu juga sebaliknya jika *author* tersebut tidak pernah bekerjasama, maka keduanya tidak dihubungkan [9].

Di dalam penelitiannya [10], memanfaatkan *co-authorship graph* untuk memprediksi topik dari sebuah makalah. Penelitian tersebut memiliki asumsi bahwa makalah yang bertetanggaan pada *co-authorship graph* memiliki topik yang sama dan topik makalah yang akan dipresiksi bergantung pada topik-topik makalah yang terhubung dengan makalah tersebut. Menggunakan data ILPnet2 yang berisi tentang informasi makalah dari ILP (*Inductive Logic Programming*) tahun 1970 sampai dengan 2003. Dari *co-authorship graph* yang terbentuk diketahui adanya komunitas ilmiah atau grup riset dari penulis makalah tersebut, pasangan penulis yang produktif. Tetapi keberhasilan metode *Fast Algorithm* ini sangat dipengaruhi oleh tingkat kepadatan ketetanggaan pada *co-authorship graph*.

2.3 Preprosesing Teks

Sebelum data teks diolah, preprosesing dilakukan dengan tujuan agar teks pada dokumen menjadi seragam dan mudah untuk dibaca oleh sistem. Preprosesing terbagi menjadi empat tahap, yaitu Tokenizing, Normalisasi, *Stopwords Removal* ,dan *Stemming*. Dokumen yang diolah adalah judul dan abstrak artikel.

a. Tokenizing

Tokenizing merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi. Token seringkali disebut sebagai istilah (term) atau kata, sebagai contoh sebuah token merupakan suatu urutan karakter dari dokumen tertentu yang dikelompokan sebagai unit semantik yang berguna untuk diproses (Salton, 1989).

b. Normalisasi

Tahapan normalisasi pada pengolahan teks memproses teks agar memiliki format yang sama sehingga memudahkan untuk analisis. Contoh normalisasi adalah *case folding* yaitu mengubah teks menjadi huruf kecil, penghilangan tanda baca, dll. Normalisasi juga dapat dilakukan untuk membersihkan tags yang biasa muncul pada *tweet, seperti hashtag, mentioned,* dan *link.*

c. Stopwords Removal

Kata-kata yang terlalu sering muncul dalam dokumen-dokumen bukanlah pembeda yang baik. Bahkan kata-kata yang muncul 80% dalam dokumen tidak berguna dalam proses text mining. Kandidat umum *stopword* adalah *article*, preposisi, dan konjungsi. *Stopwords removal* menyebabkan pengurangan ukuran strukur index hingga 40%. Karena pengurangan ukuran index, beberapa kata kerja, kata sifat, dan kata keterangan lainnya dapat juga dapat dimasukkan ke dalam daftar stopwords.

d. Stemming

Stemming merupakan sebuah proses yang melakukan mapping berbagai variasi morfologikal suatu kata menjadi bentuk dasar kata tersebut. Proses ini disebut juga dengan istilah *conflation*. Berdasarkan pada asumsi bahwa term-term yang memiliki bentuk dasar (stem) yang sama pada umumnya memiliki makna yang mirip.

2.4 Pembentukan Vector Space Model

Model ruang vektor merepresentasikan teks, dokumen, atau kueri sebagai sebuah vektor dalam sebuah ruang term. Ruang ini memiliki dimensi sebanyak jumlah term atau dengan kata lain untuk dokumen teks yang memiliki N kata maka dibutuhkan N dimensi. Setiap vektor direpresentasikan sesuai bobot dari term yang ada. Term-term yang ada menjadi sumbu-sumbu koordinat sebanyak jumlah term, sedangkan vektornya adalah sebuah titik yang posisinya berdasarkan nilai dari sumbu-sumbu tersebut. Pembuatan ruang vektor adalah dengan cara membuat sebuah matriks dua dimensi dengan kolom sebagai term dan dokumen teks sebagai baris. Isi dari matriks tersebut merupakan nilai bobot term dari masing-masing term terhadap masing-masing dokumen teks.

Salah satu *vector space model* di dalam word embeddings adalah Word2Vec. Word2Vec merupakan perangkat yang menyediakan implementasi efisien dalan *continous bag-of-words*. *Word embeddings* ini merupakan bagian dari Google News dataset yang mengandung 300 dimensi vektor untuk tiga juta kata dan frasa.

Proses penemuan nilai untuk masing-masing vektor diperoleh dari pembandingan antara suatu kata terhadap seluruh kata dengan menggunakan persamaan (2.1) :

$$P(\text{context} \mid w_t) \tag{2.1}$$

Selanjutnya diikuti oleh nilai loss function pada persamaan (2.2) :

$$J = 1 - p(w_{-t} \mid w_t) \tag{2.2}$$

Nilai dari $p(w_{t+j} \mid w_t)$ didefinisikan dengan persamaan (2.3) :

$$P(o \mid c) = \frac{\exp({u_o}^T V^c)}{\sum_{w=1}^{v} \exp({u_w}^T v_c)} \tag{2.3}$$

Menjelaskan bahwa *o* adalah output dari indeks kata, *c* adalah pusat dari indeks kata, v_c dan u_o adalah nilai vektor "tengah" dan "luar" dari kata *c* dan *o*. Nilai softmax dengan menggunakan *c* untuk memperoleh nilai probability dari *o*. Nilai yang sama cenderung memiliki nilai vektor yang sama. Hasil yang muncul merupakan suatu kumpulan vektor-vektor untuk setiap jenis tipe kata. Hal ini membuat lebih mudah untuk memprediksi kata yang muncul sebagai *context words.*

2.5 *Clustering* Dokumen

Clustering dokumen adalah pembagian dokumen ke dalam kelompok dari objek-objek yang serupa (similar). Setiap kelompok yang disebut kluster terdiri dari objek-objek yang serupa satu dengan yang lainya dan tidak serupa (dissimilar) dengan objek-objek pada *cluster* lain. Penelitian ini akan menggunakan K-Means untuk metode *clustering* dokumen.

K-Means++

K-Means++ merupakan metode pengembangan dari metode K-Means untuk mengelompokkan dokumen. Pada umumnya metode K-Means memilih data titik pusat klaster secara acak. Hal tersebut menyebabkan metode K-Means dapat menghasilkan solusi yang sub-optimal. Oleh karena itu, K-Means++ mengatasi permasalahan tersebut dengan mengoptimasi pemilihan titik pusat klaster awal sebelum metode K-Means dijalankan. Pada umumnya K-Means dan K-Means++

menggunakan pendekatan VSM (*Vector Space Model*), dimana dokumen dimodelkan dalam vektor yang memiliki kata sebagai fitur. Setelah kumpulan dokumen telah melewati tahap pra-proses teks, seluruh kata pada kumpulan dokumen diekstrak untuk dijadikan fitur dokumen, pendekatan ini disebut juga *"bag-of-words model"*. Vektor dokumen dibentuk dengan menggunakan pembobotan Tf-Idf (*Term Frequency – Inverse Document Frequency*) pada kata.

Cosine Similarity

Salah satu faktor penting dari setiap teknik *clustering* adalah bagaimana untuk menghitung similarity antara dua objek. *Cosine similarity* adalah metode pengukuran yang sering digunakan pada proses *clustering* dan peringkasan. Selain itu, *cosine similarity* adalah metode pengukuran yang mendefinisikan setiap dokumen atau term yang memiliki jarak diantaranya berdasarkan atas makna atau arti secara semantik. Terdapat dua jenis perhitungan similaritas, yaitu berdasarkan sumber daya yang telah ada, seperti thesaurus dan berdasarkan pada penyebaran kata pada suatu corpus.

$$\boldsymbol{Similarity}(\boldsymbol{A},\boldsymbol{B}) = \frac{\sum_{i=1}^{n} \boldsymbol{A}_i \,.\, \boldsymbol{B}_i}{\sqrt{\sum_{i=1}^{n} \boldsymbol{A}_i^2} \,.\, \sqrt{\sum_{i=1}^{n} \boldsymbol{B}_i^2}} \tag{2.4}$$

dimana similarity mengukur jarak kedekatan antara kata 1 dan kata2. Nilai maksimal dari jarak similariti adalah 1 yang berarti benar-benar sama, dan nilai minimal adalah -1 yang berarti benar-benar berbeda.

Proses *Clustering*

Inputan dari metode K-Means++ *clustering* adalah kumpulan dokumen *D* dan parameter jumlah *cluster* k. Setelah vektor dokumen terbentuk, algoritma K-Means++ dilakukan seperti berikut :

1. Pilih sebuah data dokumen secara acak sebagai *centroid* awal klaster 1 yang dinotasikan sebagai C_1.
2. Untuk setiap data dokumen d_i hitung $D(d_i)$ yang merupakan jarak d_i ke *centroid* terdekat yang sudah terpilih.

3. Pilih *centroid* klaster selanjutnya menggunakan probabilitas $\frac{D(d_i)}{\sum D(d_i)}$
4. Ulangi langkah 2-3 sampai *centroid* untuk seluruh klaster telah terpilih.
5. Hitung similaritas dokumen d_i dengan tiap *centroid* pada C menggunakan rumus (2.4).
6. Dokumen d_i akan menjadi anggota klaster yang memiliki nilai similaritas *centroid* tertinggi.
7. Ulangi langkah 2-3 untuk setiap dokumen dalam D.
8. Hitung ulang *centroid* untuk setiap $C_i \in Centroids$.
9. Ulangi langkah 5-8, sampai tidak ada dokumen yang berpindah klaster.

2.6 Evaluasi Jumlah *Cluster* (*Silhouette*)

Silhouette yang diperkenalkan oleh (Rousseeuw, 1987) merupakan salah satu cara untuk mengevaluasi kualitas *cluster* yang dihasilkan. Selain itu silhouette juga mengindikasikan derajat kepemilikan derajat kempemilikan setiap objek yang berada di dalam *cluster*. Objek *Oj* yang berada pada *cluster* memiliki rentang nilai *Silhouette* antara -1 sampai 1. Semakin dekat nilai *silhouette* ke 1 maka semakin tinggi derajat *Oj* di dalam *cluster*. Pada persamaan (2.5) dan (2.6) merupakan perhitungan nilai *Silhouette (s(i))* untuk setiap dokumen.

$$\boldsymbol{b(i) = max_{cj \neq a}\, d(i, c_j)} \tag{2.5}$$

$$\boldsymbol{s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}} \tag{2.6}$$

dimana *a(i)* adalah jarak kedekatan dokumen *i* terhadap seluruh dokumen yang ada di *cluster* tempat *i* berada disebut juga *cluster* internal. Sedangkan *b* adalah jarak kedekatan dokumen *i* terhadap seluruh dokumen yang ada *cluster* selain *cluster* internal (*cluster* external). Selanjutnya setiap *cluster* yang telah dihitung nilai *s(i)* akan dihitung nilai rata-rata dari *s(i)*. Perhitungan ini lebih dikenal dengan nama *Average Silhouette Width* (ASW). Range nilai ASW dapat dibagi menjadi empat kriteria yaitu :

1. Sangat baik : range ($0.71 \leq ASW < 1$)
2. Baik : range ($0.51 \leq ASW < 0.71$)
3. Cukup baik : range ($0.26 \leq ASW < 0.51$)
4. Kurang baik : range ($ASW < 0.26$)

2.7 Pelabelan *Cluster* (*TextRank*)

Cara paling umum untuk merepresentasikan teks adalah dengan pendekatan VSM (*Vector Space Model*). Pendekatan VSM pada umumnya menggunakan frekuensi kata sebagai fitur teks. Namun, cara tersebut tidak mempertimbangkan mengenai informasi semantik dan struktur dari sebuah teks. Model graf dapat merepresentasikan teks secara matematis dengan tetap mempertahankan semua informasi semantik dan struktur teks.

Sumber teks dapat berupa satu atau banyak dokumen. Setelah melalui tahap pra-pemrosesan teks, kata – kata pada teks tersebut akan menjadi *node*. Relasi kemunculan bersama (*co-occurrence*) antar kata digunakan untuk membentuk *edge* pada graf. Relasi kemunculan bersama telah banyak digunakan untuk menggambarkan hubungan kontekstual antar kata pada teks. Namun belakangan *Word2Vec* banyak diusulkan untuk mencari similaritas antar kata seperti pada kasus pencarian sinonim. Oleh karena itu, pembentukan relasi antar kata pada graf ditentukan dengan 2 cara yaitu : a) Relasi kata menggunakan *Word2Vec*, b) Relasi kata menggunakan *Co-occurrence*.

TextRank [11] merupakan metode ekstraksi kata atau frase yang terinspirasi oleh algoritma penentuan peringkat PageRank [12]. TextRank menerima masukkan teks yang telah dimodelkan dalam bentuk *graph*. Teks dapat bersumber dari satu atau banyak dokumen di dalam korpus.

Graph masukkan tersebut memiliki kata sebagai *node* dan *edge* berupa relasi antar kata. Pada penelitian ini, *edge* pada *graph* masukkan berupa kemunculan bersama antar kata (*word co-occurrence*) dan jenis *graph* merupakan *undirected graph*. Pada dasarnya metode TextRank menentukan tingkat kepentingan sebuah node berdasarkan informasi global pada struktur graph secara iteratif.

Bobot pada *node* akan selalu diperbaharui pada setiap iterasi sampai konvergensi tercapai. Jika sebuah *graph* dilambangkan sebagai $G = (V, E, W)$, dimana V adalah kumpulan *vertex/node*, E adalah kumpulan *edge*, dan W adalah kumpulan bobot *edge*. Untuk setiap *node* $V_i \in V$, notasi $In(V_i)$ merupakan kumpulan *node* yang mengarah ke *node* V_i, dan notasi $Out(V_i)$ merupakan kumpulan *node* tujuan *node* V_i. Relasi antara *node* V_i dan V_j memiliki bobot *edge* yang dilambangkan dengan $w_{ij} \in W$. Maka, bobot *node* V_i, dinotasikan dengan $WS(V_i)$, akan diperbaharui setiap iterasi dengan Persamaan (2.7).

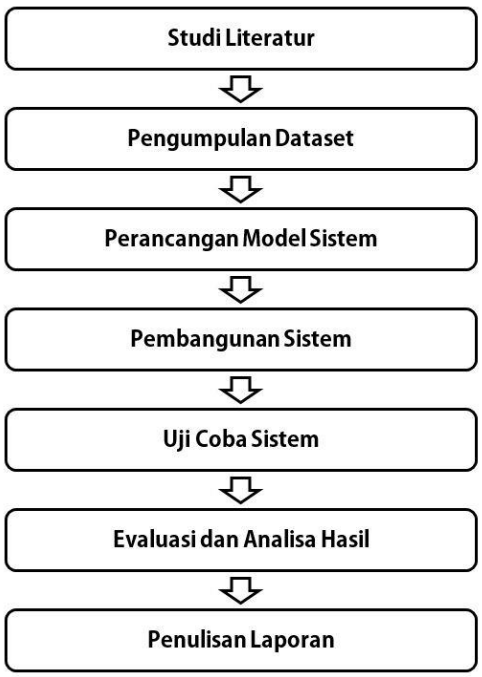
$$WS(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (2.7)$$

Pada persamaan 2.7, d merupakan koofisien *damping*, yang memiliki nilai antara 0 dan 1. Koefisien *damping* mewakili kemungkinan loncatan dari sebuah *node* ke *node* acak. Dalam konteks penelusuran web, koefisien *damping* menggambarkan probabilitas pengguna untuk memilih *link* yang tersedia pada halaman tersebut sebesar d serta probabilitas pengguna untuk pergi ke halaman web yang benar-benar acak sebesar $(1 - d)$. Implementasi koefisien *damping* dapat disebut juga *"Random Surfer Model"*.

Pasca-pemrosesan dilakukan setelah konvergensi tercapai, yaitu jika nilai bobot *node* $WS(V_i)$ sudah tidak banyak mengalami perubahan. Pada pasca-pemrosesan, sejumlah *n* kata yang memiliki skor bobot *node* terbesar akan dipilih. Setiap *n* kata akan diperiksa kumpulan *in-degree node* dan *out-degree node* milik *node* kata tersebut, untuk mencari kata lain yang terletak bersebelahan pada dokumen asal. Jika ditemukan, maka kata tersebut akan digabung menjadi sebuah frase. Keluaran tahap pasca-pemrosesan adalah kumpulan kata dan frase yang dianggap merepresentasikan korpus sumber.

BAB 3
METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai tahapan metodologi penelitian yang akan dikerjakan dalam penelitian tesis ini. Secara umum, penelitian ini diawali dengan studi literatur untuk mempelajari permasalahan dalam penelitian terkait, pengumpulan dataset, perancangan model sistem, kemudian dilanjutkan dengan pembuatan dan implementasi sistem, uji coba sistem, evaluasi dan analisa hasil percobaan, selanjutnya diakhiri dengan penulisan laporan. Keseluruhan tahapan dalam proses penelitian ini digambarkan pada Gambar 3.1.



Gambar 3.1. Alur metodologi penelitian

3.1 Studi Literatur

Dalam melakukan suatu penelitian, tahapan studi literatur dan analisa awal permasalahan merupakan hal yang sangat dibutuhkan. Hal ini berkaitan dengan pemahaman detail baik dari sisi konsep dan dasar teori yang digunakan maupun teknis dari setiap tahapan suatu penelitian. Studi literatur dilakukan untuk mendapatkan segala informasi dan sumber pustaka yang berkaitan dengan lingkup

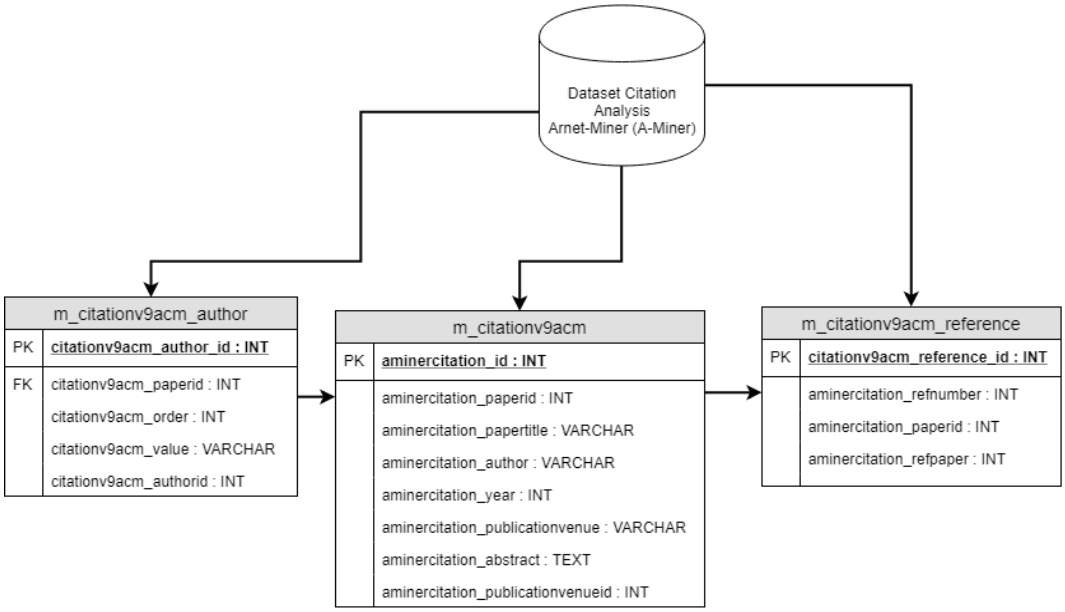
pembahasan, perkembangan keilmuan dalam penelitian, serta metode yang telah tercipta sebelumnya. Dalam penelitian ini, literatur yang dikaji secara garis besar meliputi konsep dasar yang berkaitan dengan metode pendeteksian penulis berpotensi atau biasa disebut *rising stars* dan berbagai perkembangan penelitiannya. Studi literartur yang dilakukan mencakup pencarian atau mempelajari referensi-referensi yang terkait, seperti :

1. *Key Performance Indicator* (KPI) yang menunjukkan ringkasan riwayat dari kinerja kegiatan penelitian *author*.
2. *Text Preprocessing* (tokenisasi, normalisasi, stopwords removal, stemming) dan pembobotan term untuk merepresentasikan suatu dokumen.
3. Metode *clustering* untuk proses ekstraksi topik pada sekumpulan dokumen.
4. Teori *Graph* (*Co-Authorship Graph* dan *Power Graph*) untuk memodelkan kemampuan kolaborasi *author* dengan *author* yang lain.
5. Metode evaluasi yang dapat menunjukkan keunggulan dari usulan metode.

3.2 Pengambilan Dataset

Data artikel ilmiah yang digunakan dalam penelitian ini bersumber dari ACM-Citation-networkV9 di dalam AMiner dataset. Secara keseluruhan dataset tersebut memiliki jumlah *author* dan *co-authorship* sebesar 4.858.661 dengan 2.385.066 artikel ilmiah, data sitasi artikel sejumlah 9.671.893, dan semua data tersebut menyediakan informasi dari tahun 1936 hingga tahun 2016. AMiner dataset terdiri dari berbagai macam artikel ilmiah dalam bahasa inggris seperti *proceedings*, *journal papers*, dan tesis yang berkaitan dengan ilmu komputer.

Secara keseluruhan terdapat 28 tabel dengan 12 macam versi data di dalam AMiner dataset, tetapi penelitian ini hanya menggunakan 3 tabel besar di dalam versi ACM-Citation-networkV9. Gambar 3.2 menjelaskan secara detail mengenai skema basis data dan semua tabel yang digunakan dalam penelitian ini. Tabel *m_citationv9acm_author* menyimpan semua data *author* yang memiliki artikel ilmiah pada tabel *m_citationv9acm*, seperti data id *author*, nama *author*, dan urutan penulis atau *author* dalam artikel ilmiah.



Gambar 3.2. Skema database pada AMiner dataset

Tabel *m_citationv9acm* tersebut berisi informasi lengkap tentang artikel ilmiah seperti, id *paper*, judul, abstrak, daftar *author*, tahun publikasi, dan tempat publikasi artikel ilmiah. Setiap artikel ilmiah memiliki beberapa referensi yang dijabarkan dalam bentuk id *paper*, id referensi, dan urutan referensi, semua informasi tersebut tersimpan pada tabel *m_citationv9acm_reference*.

Tabel 3.1. Data table m_citationv9acm

Nama Kolom	Tipe Data	Deskripsi
aminercitation_id	int (11)	id baris
aminercitation_paperid	int (11)	id artikel
aminercitation_papertitle	varchar (500)	judul artikel
aminercitation_author	varchar (500)	daftar *author*
aminercitation_year	int (11)	tahun publikasi
aminercitation_publicationvenue	varchar (500)	tempat publikasi
aminercitation_abstract	text	abstrak artikel
aminercitation_publicationvenueid	int (11)	id tempat publikasi

Tabel 3.2. Contoh data pada table m_citationv9acm

Nama Kolom	Contoh Data
aminercitation_id	134
aminercitation_paperid	134
aminercitation_papertitle	Logical, internal, and physical reference behavior in CODASYL database systems
aminercitation_author	Wolfgang Effelsberg, Mary E. S. Loomis
aminercitation_year	1984
aminercitation_publicationvenue	ACM Transactions on Database Systems (TODS)
aminercitation_abstract	This work investigates one aspect of the performance of CODASYL database systems: the data reference behavior. We introduce a model of database traversals at three levels: the logical, internal, and physical levels……
aminercitation_publicationvenueid	-

Tabel 3.1 dan Tabel 3.2 berturut-turut menjelaskan secara detail tentang semua nama kolom, tipe data, deskripsi singkat, dan contoh data pada table *m_citationv9acm*. Tabel *m_citationv9acm* sendiri memiliki 8 kolom dengan *aminercitation*_id bertindak sebagai *primary key* pada tabel tersebut, dan kolom *aminercitation_paperid* berfungsi untuk menghubungkan tabel *m_citationv9acm* ke tabel *m_citationv9acm_author* dan tabel *m_citationv9acm_reference*.

Sedangkan di dalam Tabel 3.3 menjelaskan tentang nama kolom, tipe data, deskripsi singkat, dan contoh data pada table *m_citationv9acm_author*. Tabel *m_citationv9acm_author* memiliki 5 kolom dengan *citationv9acm_author_id* sebagai *primary key* pada tabel tersebut, dan kolom *citationv9acm_paperid* berfungsi untuk menghubungkan tabel *m_citationv9acm_author* ke tabel *m_citationv9acm* dan tabel *m_citationv9acm_reference*.

Tabel 3.3. Data table m_citationv9acm_author

Nama Kolom	Tipe Data	Deskripsi	Cotoh Data
citationv9acm_author_id	int (11)	id baris	213
citationv9acm_paperid	int (11)	id artikel	134
citationv9acm_order	int (11)	urutan *author*	1
citationv9acm_value	varchar (125)	nama *author*	Wolfgang Effelsberg
citationv9acm_authorid	int (11)	id *author*	1607399

Kemudian Tabel 3.4 menjelaskan secara detail tentang nama kolom, tipe data, deskripsi singkat, dan contoh data pada table *m_citationv9acm_reference*. Tabel *m_citationv9acm_reference* memiliki 4 kolom, *aminercitation_reference_id* bertindak sebagai *primary key* dan kolom *aminercitation_paperid* bertindak sebagai penghubung antara tabel *m_citationv9acm_reference* dengan tabel *m_citationv9acm* dan tabel *m_citationv9acm_author*.

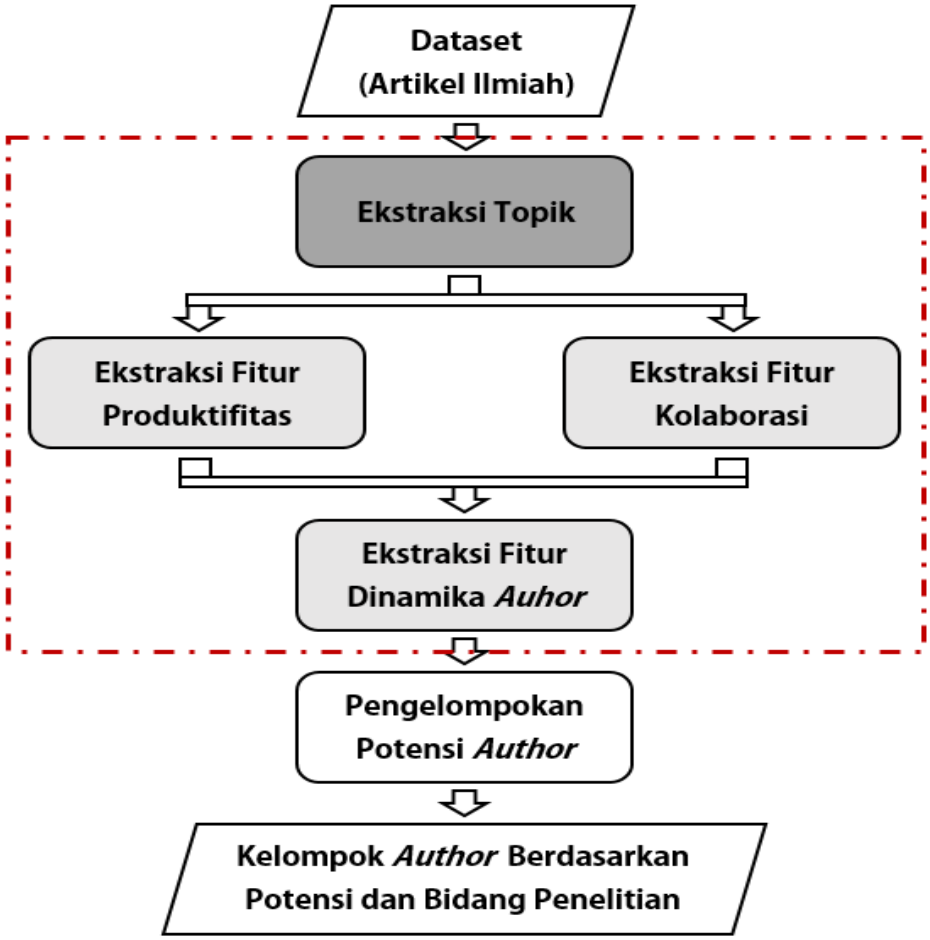
Tabel 3.4. Data table m_citationv9acm_reference

Nama Kolom	Tipe Data	Deskripsi	Cotoh Data
aminercitation_reference_id	int (11)	id baris	127
aminercitation_refnumber	int (11)	urutan referensi	9
aminercitation_paperid	int (11)	id artikel	134
aminercitation_refpaper	int (11)	id referensi	319472

Bagian dari artikel ilmiah yang digunakan dalam proses ekstraksi topik yaitu id artikel, judul, dan abstrak. Selanjutnya data yang digunakan dalam proses ekstraksi fitur produktifitas berdasarkan topik adalah id *author*, nama *author*, id artikel, tahun publikasi, id sitasi, dan tahun sitasi. Pada proses berikutnya, ekstraksi fitur kolaborasi berdasarkan topik membutuhkan semua informasi pada fitur produktifitas ditambahkan dengan informasi *co-authorship* setiap *author*. Proses terakhir dalam ekstraksi fitur yaitu ekstraksi fitur dinamika *author*, membutuhkan informasi tahun pada semua fitur yang telah terbentuk sebelumnya.

3.3 Perancangan Model Sistem

Secara garis besar penelitian ini melakukan 5 tahap besar dalam proses pengelompokan potensi *author* berdasarkan bidang penelitian. Pada Gambar 3.3 menunjukkan alur usulan metode penelitian dan urutan proses dari kelima tahap utama pada penelitian ini. Tahap pertama merupakan proses usulan pada penelitian ini, yaitu proses ekstraksi topik yang diperoleh dari proses *clustering* pada judul dan abstrak dalam publikasi artikel ilmiah yang dimiliki *author*.



Gambar 3.3. Alur usulan metode penelitian

Tahap kedua adalah ekstraksi fitur produktifitas setiap *author* berdasarkan topik penelitian yang telah diekstraksi sebelumnya. Pada penelitian ini produktifitas *author* diambil dari informasi frekuensi jumlah artikel masing-masing *author* dan sitasi yang didapatkan pada setiap publikasi per tahunnya.

Selanjutnya tahap ketiga adalah pembuatan *graph* untuk menggambarkan kemampuan *author* dalam berkolaborasi dengan *author* lain dalam melakukan kegiatan penelitian berdasarkan topik penelitiannya. Untuk menggambarkan kemampuan kolaborasi setiap *author*, metode usulan membentuk *Co-Authorship Graph* dengan dua macam pembobotan edge yaitu dengan mempertimbangkan produktifitas *author* (jumlah publikasi artikel ilmiah) dan kontribusi dari publikasi tersebut (jumlah sitasi yang diterima pada setiap artikel).

Tahap yang keempat adalah mengekstraksi fitur dinamika *author* yang mengamati perubahan dalam kegiatan penelitian *author* pada setiap periode berdasarkan fitur-fitur yang telah didapatkan dari kedua proses ekstraksi fitur sebelumnya. Pada tahap kelima atau tahap terakhir dalam penelitian ini adalah melakukan proses *clustering* pada *author* dengan menggunakan keempat aspek utama tersebut sebagai fitur dalam rangka menemukan sekelompok *author* yang berpotensi dalam bidang penelitian yang dilakukan.

Dalam penelitian ini, penulis menggunakan beberapa notasi dalam proses pembuatan metode usulan. Adapun notasi dan definisi yang digunakan akan dijelaskan secara terperinci pada Tabel 3.5.

Tabel 3.5. Notasi dan Definisi dalam Metode Usulan

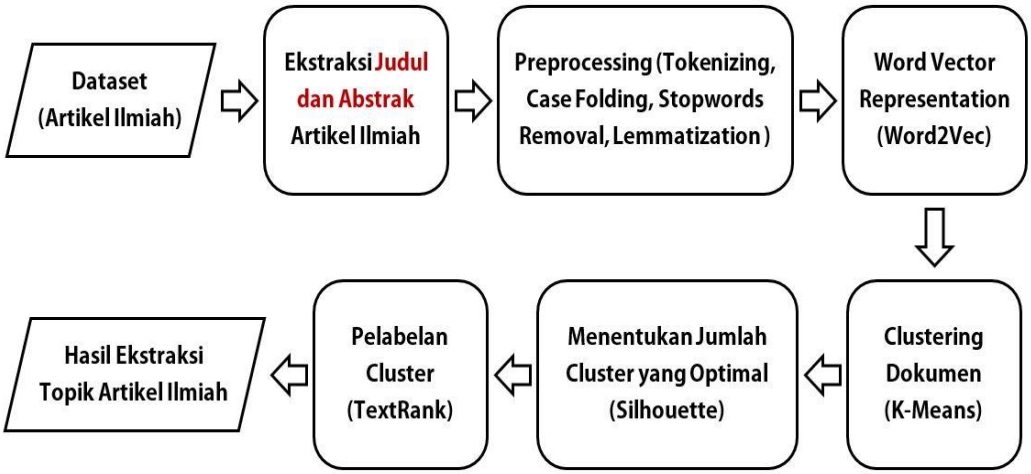
NO	Notasi	Keterangan
Ekstraksi Topik		
1.	D	sekumpulan artikel ilmiah, $D = \{d_1, d_2, d_3, \dots, d_n\}$
2.	W	sekumpulan term dalam artikel, $W = \{w_1, w_2, w_3, \dots, w_n\}$
3.	Z	hasil ekstraksi topik, $Z = \{z_1, z_2, z_3, \dots, z_n\}$
Ekstraksi Fitur Produktifitas		
4.	x atau y	penulis atau *author*
5.	i atau j	artikel ilmiah
6.	t_k	periode waktu, dalam tahun
7.	t_0	periode *author* pertama kali mempublikasikan artikel ilmiah

8.	t_n	periode pengujian
9.	P_x	sekumpulan artikel *author* x pada periode t_0 sampai t_n
10.	$per\ (i)$	periode artikel i dipublikasikan
11.	$pub\ (x, z, t)$	jumlah artikel *author* x dalam topik z pada periode t
12.	$cit\ (i, t)$	jumlah sitasi yang diterima artikel i pada periode t
Ekstraksi Fitur Kolaborasi		
13.	C_x	sekumpulan *author* yang memiliki *co-author* dengan *author* x
14.	$CoAut\ (x, z)$	jumlah *co-author* dari *author* x dalam topik z
15.	$aut\ (i)$	jumlah *author* pada artikel i
16.	$w\ (t_k)$	bobot periode t_k terhadap periode t_n
17.	$WE_z\ (x, y)$	bobot *edge* yang menghubungkan kolaborasi penelitian dalam topik z antara *author* x dan *author* y
18.	λ	nilai bobot *edge* maksimum dalam *graph*
19.	α	bobot *impact* artikel pada $WE_{imp,z}\ (x, y)$, α = 0.7
20.	β	bobot *quantity* artikel pada $WE_{imp,z}\ (x, y)$, β = 0.3
Ekstraksi Fitur Dinamika *Author*		
26.	n	jumlah periode perubahan dari t_0 sampai t_n
27.	f_t	nilai dari fitur f pada periode t
28.	f, a	fitur dari nilai pada periode pengujian (f_1 dan f_4)
29.	f, b	fitur dari nilai kumulatif hingga periode pengujian (f_2 dan f_5)
30.	f, c	fitur dari nilai pembobotan yang mempertimbangkan dampak penurunan waktu (f_3 , f_6 , f_7 , f_8 , f_9 , f_{10} , f_{11} , f_{12} dan f_{13})

3.3.1 Ekstraksi Topik dalam Artikel Ilmiah

Proses awal pada metode yang diusulkan adalah ekstraksi topik dalam artikel ilmiah. Serangkaian proses dari ekstraksi topik dalam artikel ilmiah digambarkan secara detail pada Gambar 3.4. Ekstraksi topik didapatkan dari proses *clustering* terhadap dataset yang telah diolah sebelumnya. Adapun data yang digunakan untuk proses ekstraksi topik adalah informasi judul dan abstrak dalam semua artikel

ilmiah yang dimiliki oleh *author*. Selanjutnya, beberapa tahap preprosesing diterapkan pada dataset agar lebih mudah untuk dibaca dan diolah oleh sistem komputer. Berikut ini adalah penjelasan dan contoh dari *preprocesing* teks, serta ringkasan bentuk input teks dan hasil output dari serangkaian tahap preprosesing yang dijelaskan pada Tabel 3.6.



Gambar 3.4. Proses ekstraksi topik dalam artikel ilmiah

1. Tokenizing

   Teks pada artikel dipecah menjadi unit terkecil kata-kata atau term.

   "In today' s Complex Academic Environment the process of Performance Evaluation of Scholars is becoming increasingly difficult."

   akan berubah menjadi

   {In, today' s, Complex, Academic, Environment, the, process, of, Performance, Evaluation, of, Scholars, is, becoming, increasingly, difficult}

2. Case Folding

   Merubah term dalam artikel menjadi huruf kecil (lowercase).

   {In, today' s, Complex, Academic, Environment, the, process, of, Performance, Evaluation, of, Scholars, is, becoming, increasingly, difficult}

   akan berubah menjadi

   {in, today' s, complex, academic, environment, the, process, of, performance, evaluation, of, scholars, is, becoming, increasingly, difficult}

3. Stopwords Removal

   Menghapus kata henti dalam bahasa inggris.

   {in, today’ s, complex, academic, environment, the, process, of, performance, evaluation, of, scholars, is, becoming, increasingly, difficult}

   akan berubah menjadi

   {today’ s, complex, academic, environment, process, performance, evaluation, scholars, becoming, increasingly, difficult}

4. Lemmatization

   Merubah term dalam artikel menjadi bentuk dasar.

   {today’ s, complex, academic, environment, process, performance, evaluation, scholars, becoming, increasingly, difficult}

   akan berubah menjadi

   {today, complex, academic, environment, process, performance, evaluation, scholar, become, increase, difficult}

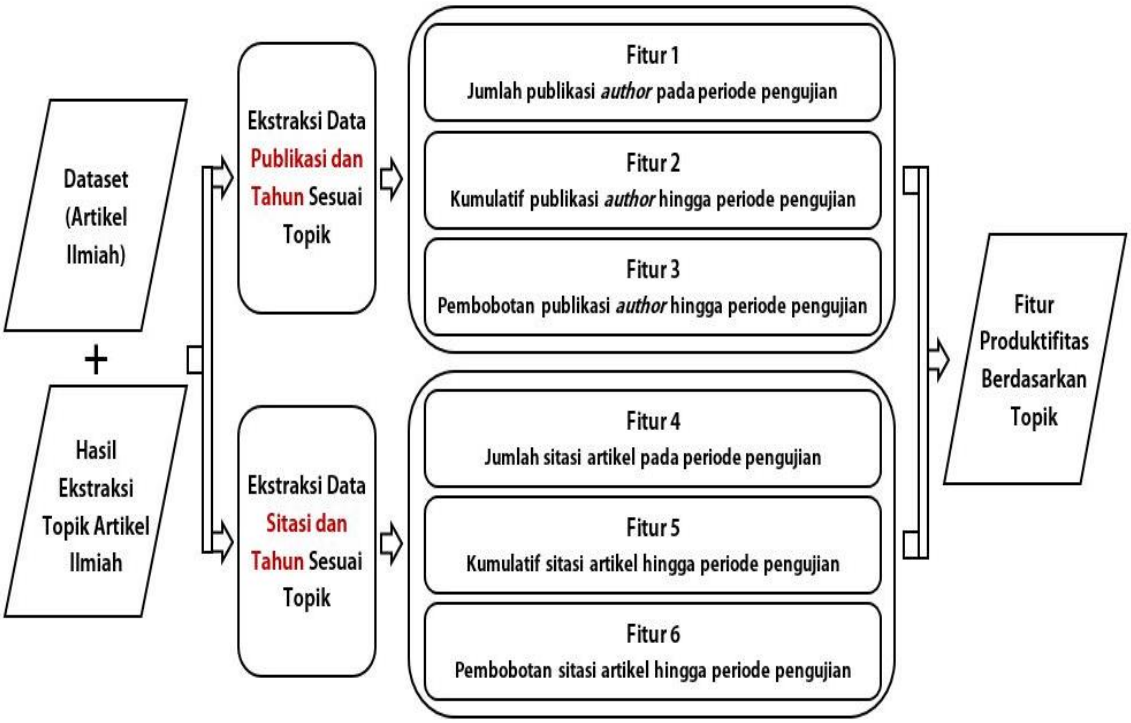
Tabel 3.6. Input dan Output pada Tahap Preprosesing

Input	In today’s Complex Academic Environment, the process of Performance Evaluation of Scholars is becoming increasingly difficult.
Output	{today, complex, academic, environment, process, performance, evaluation, scholar, become, increase, difficult}

Setelah dilakukan beberapa tahap preprosesing teks di dalam artikel ilmiah, tahap selanjutnya yaitu dilakukan proses representasi kata kedalam bentuk vektor dengan menggunakan Word2Vec. Setelah didapatkan vektor yang berisi pembobotan dari setiap term pada artikel, vektor tersebut dijadikan fitur dalam proses *clustering* dengan menggunakan metode K-Means++. Proses *clustering* dilakukan pada jumlah *cluster* yang berbeda, yaitu 2 sampai dengan 25 *cluster*.

Selain itu proses *clustering* menggunakan cosnine similarity untuk pengukuran jarak antar artikel ilmiah. Kemudian hasil *cluster* tersebut dianalisa dengan menggunakan *Silhouette* analisis untuk menentukan jumlah *cluster* yang optimal terhadap kumpulan artikel tersebut. Proses terakhir yaitu pelabelan *cluster* dengan menggunakan metode pemodelan TopicRank.

3.3.2 Ekstraksi Fitur Produktifitas *Author* Berdasarkan Topik



Gambar 3**Error! No text of specified style in document.**.5. Proses ekstraksi fitur produktifitas *author* berdasarkan topik

Setiap *author* memiliki beberapa fitur yang menggambarkan kinerjanya secara individual. Pada proses ekstraksi fitur produktifitas, dataset yang berisi informasi artikel setiap *author* beserta tahun publikasi dan sitasi setiap artikel per tahun diproses bersama dengan hasil ekstraksi topik artikel ilmiah yang telah didapatkan pada proses sebelumnya. Selanjutnya pada proses ini, 6 macam fitur produktifitas *author* dihasilkan berdasarkan topik penelitian sesuai yang digambarkan secara detail pada Gambar 3.5.

Penelitian ini mengkutip cara pengambilan fitur dari penelitian yang dilakukan oleh George Panagopoulos pada tahun 2017 [7], selanjutnya penelitian ini mengembangkan fitur tersebut dengan menambahkan unsur topik di dalam pengekstraksiannya. Berikut adalah penjelasan secara detail dalam ekstraksi masing-masing fitur dan contoh penerapannya.

Fitur 1 :

Menunjukkan jumlah publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z pada periode pengujian t_n, ditunjukkan dengan persamaan (3.1).

$$\boldsymbol{f}_{\mathbf{1},\boldsymbol{z}}(\boldsymbol{x}) = \boldsymbol{pub}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{t_n}) \tag{3.1}$$

Sebagai contohnya, diasumsikan pendekatan ini melakukan pengujian di tahun 2015 (t_n = 2015) pada seorang *author* x yang memiliki publikasi artikel ilmiah dalam topik z sebanyak 1 artikel pada tahun 2013 dan 2 artikel pada tahun 2015, Sehingga akan diperoleh : $f_1(x) = 2$

Fitur 2 :

Menunjukkan jumlah kumulatif publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z dari periode pertama kali *author* mempublikasikan artikel ilmiah t_0 hingga periode pengujian t_n, ditunjukkan dengan persamaan (3.2).

$$\boldsymbol{f}_{\mathbf{2},\boldsymbol{z}}(\boldsymbol{x}) = \sum_{\boldsymbol{t_k}=\boldsymbol{t_0}}^{\boldsymbol{t_n}} \boldsymbol{pub}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{t_k}) \tag{3.2}$$

Menggunakan contoh kasus yang sama, maka diperoleh fitur 2 sebesar :

$f_2(x) = 1 + 0 + 2 = 3$

Fitur 3 :

Menunjukkan pembobotan pada publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z dari periode pertama kali *author* mempublikasikan artikel ilmiah t_0 hingga periode pengujian t_n. Pembobotan yang dimaksud adalah penilaian artikel yang mempertimbangkan unsur pinalti terhadap tahun publikasi artikel,

sehingga artikel akan mengalami penurunan nilai seiring dengan bertambahnya periode pengujian, yang ditunjukkan dengan persamaan (3.3).

$$\boldsymbol{f}_{\boldsymbol{3,z}}(\boldsymbol{x}) = \sum_{\boldsymbol{t_k = t_0}}^{\boldsymbol{t_n}} \frac{\boldsymbol{pub}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{t_k})}{\boldsymbol{t_n - t_k + 1}} \tag{3.3}$$

Menggunakan contoh kasus yang sama, maka diperoleh fitur 3 sebesar :

$$f_3(x) = \frac{1}{2015 - 2013 + 1} + \frac{0}{2015 - 2014 + 1} + \frac{2}{2015 - 2015 + 1} = 2.333$$

Berikut contoh perbedaan nilai bobot fitur 3 (f_3) pada 3 artikel yang diterbitkan dalam tahun 2013 (t_k = 2013) jika dilakukan pengujian pada tahun 2013, 2014, dan 2015 (t_n = 2013, 2014, 2015) :

$$f_3(t_n\ 2013) = \frac{3}{2013 - 2013 + 1} = 3$$

$$f_3(t_n\ 2014) = \frac{3}{2014 - 2013 + 1} = 1.5$$

$$f_3(t_n\ 2015) = \frac{3}{2015 - 2013 + 1} = 1$$

Saat dilakukan pengujian pada tahun 2013 bobot artikel bernilai 3, jika dilakukan pengujian pada tahun 2014 bobot artikel bernilai 1.5, jika dilakukan pengujian pada tahun 2015 bobot artikel tersebut bernilai 1. Begitu juga seterusnya, nilai bobot artikel akan terus menurun seiring dengan bertambahnya periode pengujian.

Fitur 4 :

Menunjukkan jumlah sitasi yang diterima oleh masing-masing publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z pada periode pengujian t_n, ditunjukkan dengan persamaan (3.4).

$$\boldsymbol{f}_{\boldsymbol{4,z}}(\boldsymbol{x}) = \sum_{\forall \boldsymbol{i} \in (\boldsymbol{z} \cap \boldsymbol{P_x})} \boldsymbol{cit}(\boldsymbol{i}, \boldsymbol{t_n}) \tag{3.4}$$

Melanjutkan dari contoh sebelumnya, diasumsikan artikel pertama menerima sitasi sebanyak 11 ditahun 2013, 20 ditahun 2014, dan 25 ditahun 2015. Artikel kedua

menerima sitasi sebanyak 27 ditahun 2015, sedangkan artikel ketiga menerima sitasi sebanyak 17 ditahun 2015. Sehingga akan diperoleh fitur 4 sebesar :

$f_4\,(x)$ = 25 + 27 + 17 = 69

Fitur 5 :

Menunjukkan jumlah kumulatif sitasi yang diterima oleh masing-masing publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z dari periode pertama kali *author* mempublikasikan artikel ilmiah t_0 hingga periode pengujian t_n, ditunjukkan dengan persamaan (3.5).

$$\boldsymbol{f_{5,z}\,(x)} = \sum_{\forall i \in (z \cap P_x)} \sum_{t_k = t_0}^{t_n} \boldsymbol{cit\,(i, t_k)} \tag{3.5}$$

Menggunakan contoh kasus yang sama, maka diperoleh fitur 5 sebesar : $f_5\,(x)$ = (11 + 20 + 25) + (27) + (17) = 100

Fitur 6 :

Menunjukkan pembobotan pada sitasi yang diterima oleh masing-masing publikasi artikel ilmiah yang dimiliki oleh *author* x dalam topik z dari periode pertama kali *author* mempublikasikan artikel ilmiah t_0 hingga periode pengujian t_n. Pembobotan yang dimaksud yaitu penilaian sitasi artikel yang mempertimbangkan unsur pinalti terhadap tahun publikasi artikel, sehingga bobot sitasi artikel akan mengalami penurunan nilai seiring dengan bertambahnya periode pengujian seperti yang telah dijelaskan pada fitur 3, yang ditunjukkan dengan persamaan (3.6).

$$\boldsymbol{f_{6,z}\,(x)} = \sum_{\forall i \in (z \cap P_x)} \sum_{t_k = t_0}^{t_n} \frac{\boldsymbol{cit\,(i, t_k)}}{\boldsymbol{t_n - t_k + 1}} \tag{3.6}$$

Menggunakan contoh kasus yang sama, maka diperoleh fitur 6 sebesar :

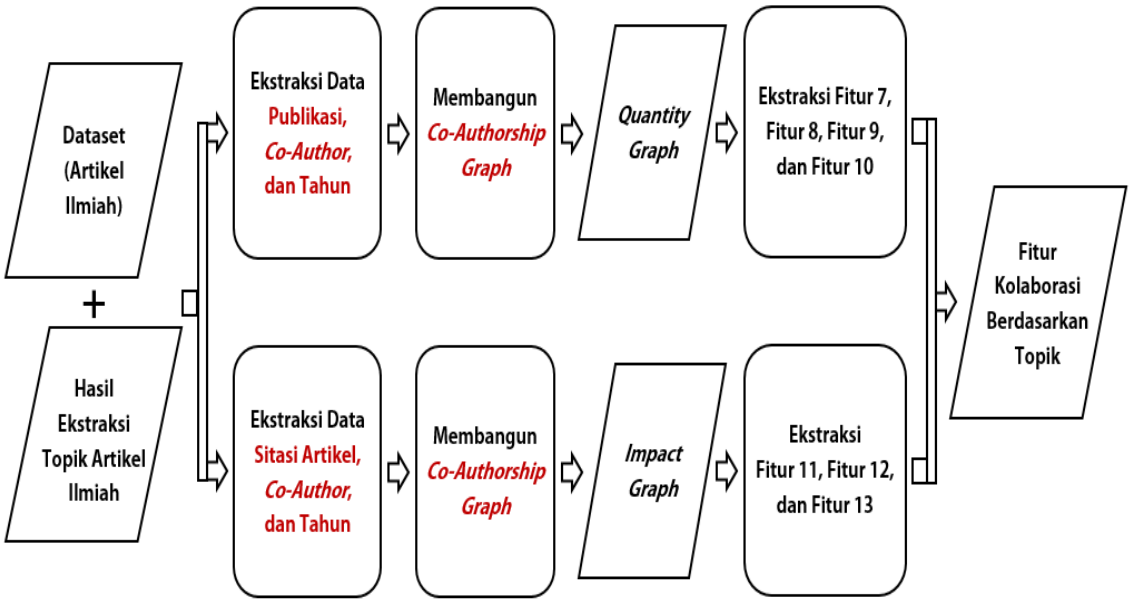
$$f_6\,(x) = \left(\left(\frac{11}{2015 - 2013 + 1} + \frac{20}{2015 - 2014 + 1} + \frac{25}{2015 - 2015 + 1}\right) + \left(\frac{27}{2015 - 2015 + 1}\right) + \left(\frac{17}{2015 - 2015 + 1}\right)\right) = 82.667$$

Dari beberapa tahapan proses ekstraksi fitur produktifitas diatas, didapatkan 6 fitur untuk *author* x yang memiliki publikasi artikel dalam topik z, yang akan disimpulkan dalam Tabel 3.7 :

Tabel 3.7. Output pada Proses Ekstraksi Fitur Produktifitas

	$f_{1,z}$	$f_{2,z}$	$f_{3,z}$	$f_{4,z}$	$f_{5,z}$	$f_{6,z}$
x_1	2	3	2.333	69	100	82.667
.....						
x_n						

3.3.3 Ekstraksi Fitur Kolaborasi *Author* Berdasarkan Topik



Gambar 3.6. Proses ekstraksi fitur kolaborasi *author* berdasarkan topik

Selain fitur produktifitas, sangatlah penting dalam proses pengelompokan potensi penulis untuk mengamati kemampuan kolaborasi *author* dengan *author* yang lain. Ekstraksi fitur kolaborasi *author* diproses berdasarkan bidang penelitian *author* yang telah dididapat dari tahap sebelumnya, yaitu ekstraksi topik. Dalam penelitian ini menggunakan *co-authorship graph* untuk memodelkan kemampuan kolaborasi *author*, seperti yang telah digambarkan pada Gambar 3.6.

Pada penelitian ini *co-authorship graph* memiliki pembobotan edge dengan dua cara yang berbeda, yaitu pembobotan edge berdasarkan *quantity* atau frekuensi publikasi artikel ilmiah *author*, kemudian *co-authorship graph* tersebut dinamakan *Quantity Graph*. Selanjutnya pembobotan edge dilakukan berdasarkan *impact* dari artikel atau sitasi yang diterima oleh setiap artikel, dan *co-authorship graph* tersebut dinamakan *Impact Graph*. Hasil *Quantity Graph* akan digunakan sebagai inputan dalam ekstraksi fitur 7, fitur 8, fitur 9, dan fitur 10. Sedangkan hasil dari *Impact Graph* akan digunakan untuk mengekstraksi fitur 11, fitur 12, dan fitur 13.

Membangun *Co-Authorship Graph*

Co-authorship graph merupakan tipe *graph* dalam ukuran besar, berisi *edge* tunggal yang menghubungkan *author* satu dengan *author* lain yang telah menulis setidaknya satu artikel ilmiah secara bersama. Di dalam *graph* ini, terdapat ***node*** yang menggambarkan individu yang berkolaborasi satu sama lain, dan ***edge*** yang menunjukkan kolaborasi atau *co-authorship* terhadap satu atau banyak artikel.

Bobot *edge* pada *co-authorship graph* menggambarkan kekuatan atau kesuksesan dari kolaborasi *author*, yang berisi gabungan informasi dari seluruh artikel di dalam periode tertentu berdasarkan jumlah publikasi, jumlah kumulatif sitasi, dan informasi lainnya. Terdapat dua bentuk *graph* dalam *co-authorship graph*, yaitu *quantity graph* dan *impact graph*, masing-masing *graph* tersebut memiliki cara yang berbeda dalam pembentukan bobot *edge*. Berikut merupakan persamaan untuk mendapatkan bobot *edge* dalam *quantity graph* (3.7) :

$$\boldsymbol{WE}_{\boldsymbol{quan,z}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\forall \boldsymbol{i} \in (\boldsymbol{z} \cap (\boldsymbol{P_x} \cap \boldsymbol{P_y}))} \frac{\mathbf{1}}{\boldsymbol{t_n} - \boldsymbol{per}(\boldsymbol{i}) + \mathbf{1}} \tag{3.7}$$

Persamaan tersebut menggambarkan jumlah bobot semua artikel yang ditulis oleh *author* x dan *author* y dalam topik z hingga periode t_n. Adapun persamaan untuk mendapatkan bobot *edge* dalam i*mpact graph* yang menggambarkan dampak dari kolaborasi *author* x dan *author* y dalam topik z, yang berupa jumlah bobot sitasi artikel hingga periode t_n. Didapatkan dengan persamaan (3.8).

$$\boldsymbol{WE}_{\boldsymbol{imp,z}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\forall \boldsymbol{i} \in (\boldsymbol{z} \cap (\boldsymbol{P_x} \cap \boldsymbol{P_y}))} \frac{\boldsymbol{\alpha} \,.\, \sum_{\boldsymbol{t_k}=\boldsymbol{t_0}}^{\boldsymbol{t_n}} (\boldsymbol{cit}(\boldsymbol{i},\boldsymbol{t_k}) \,.\, \boldsymbol{w}(\boldsymbol{t_k})) + \boldsymbol{\beta}}{\boldsymbol{aut}(\boldsymbol{i}) \,.\, (\boldsymbol{t_n} - \boldsymbol{per}(\boldsymbol{i}) + \boldsymbol{1})} \quad (3.8)$$

Nilai α merupakan bobot prioritas pada *impact* artikel dan β merupakan bobot prioritas pada *quantity* artikel. Beberapa penelitian sebelumnya telah mengusulkan teknik untuk menyeimbangkan antara nilai *impact* dan *quantity*. Pada penelitian ini penulis mengikuti skema pembobotan [7] yang mengatur nilai α = 0,7 dan β = 0,3. Adapun bobot periode t_k terhadap t_n diperoleh dengan persamaan (3.9).

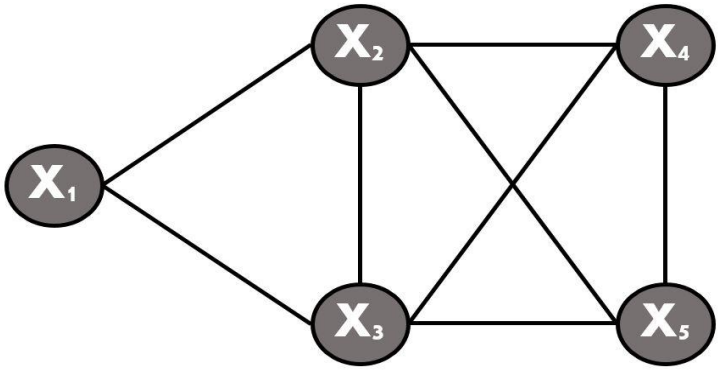
$$\boldsymbol{w}(\boldsymbol{t_k}) = \frac{\boldsymbol{1}}{\boldsymbol{t_n} - \boldsymbol{t_k} + \boldsymbol{1}} \quad (3.9)$$

Sebagai contohnya, diasumsikan pendekatan ini melakukan pengujian di tahun 2015 (t_n = 2015) pada lima orang *author* $\{x_1, x_2, x_3, x_4, x_5\}$ yang memiliki publikasi artikel ilmiah dalam topik z sebanyak lima artikel $\{i_1, i_2, i_3, i_4, i_5\}$. Diketahui *author* x_1 menulis artikel ilmiah $\{i_2, i_4, i_5\}$, selanjutnya *author* x_2 menulis artikel ilmiah $\{i_1, i_2, i_3, i_4, i_5\}$, kemudian *author* x_3 menulis artikel ilmiah $\{i_1, i_2, i_3, i_4, i_5\}$, sedangkan *author* x_4 menulis artikel ilmiah $\{i_1, i_3\}$, dan *author* x_5 menulis artikel ilmiah $\{i_1, i_3\}$. Tahun publikasi dan jumlah sitasi per tahun pada masing-masing artikel akan ditunjukkan secara detail pada Tabel 3.8.

Tabel 3.8. Tahun publikasi dan sitasi artikel

Artikel	**Tahun Publikasi**	**Jumlah Sitasi Artikel**					**Total Sitasi**
		2011	**2012**	**2013**	**2014**	**2015**	
i_1	2011	15	7	17	25	10	74
i_2	2013	-	-	11	20	25	56
i_3	2013	-	-	27	17	23	67
i_4	2015	-	-	-	-	27	27
i_5	2015	-	-	-	-	17	17

Dalam pembangunan *co-authorship graph*, langkah awal yang dilakukan adalah mencari *co-author* dari masing-masing *author*. Dari contoh studi kasus diatas, didapatkan bahwa *author* x_1 berkolaborasi dengan *author* $\{x_2, x_3\}$, selanjutnya *author* x_2 berkolaborasi dengan *author* $\{x_1, x_3, x_4, x_5\}$, sedangkan *author* x_3 berkolaborasi dengan *author* $\{x_1, x_2, x_4, x_5\}$, kemudian *author* x_4 berkolaborasi dengan *author* $\{x_2, x_3, x_5\}$, dan *author* x_5 berkolaborasi dengan *author* $\{x_2, x_3, x_4\}$. Berikut pada Gambar 3.7 menggambarkan bentuk dari hasil *co-authorship graph* yang terbentuk dari studi kasus tersebut.



Gambar 3.7. Hasil *co-authorship graph* dari studi kasus

Dalam penelitian ini, penulis menggunakan dua macam bentuk *co-authorship graph*, yaitu *quantity graph* dan *impact graph*. Kedua bentuk *graph* tersebut memiliki struktur yang sama, tetapi memiliki cara pembentukan bobot *edge* yang berbeda. Bobot *edge* dalam *quantity graph* didapatkan dengan persamaan (7), yaitu mencerminkan jumlah publikasi dari *author* dan kolaborasinya. Sedangkan untuk mendapatkan bobot *edge* dalam *impact graph* menggunakan persaman (8) dan (9), yaitu mencerminkan jumlah sitasi yang didapatkan oleh masing-masing artikel ilmiah yang dimiliki oleh *author* dan kolaborasinya. Berikut merupakan hasil dari bobot *edge* pada *quantity graph* dan *impact graph*, yang akan dijelaskan secara detai pada Tabel 3.9.

Setelah terbentuk dua macam *co-authorship graph* dengan proses tersebut, *quantity graph* akan digunakan untuk mengekstraksi fitur 7, fitur 8, fitur 9, dan fitur 10. Berikut adalah penjelasan dalam ekstraksi masing-masing fitur dan contoh penerapannya pada studi kasus yang telah dijelaskan sebelumnya.

Tabel 3.9. Bobot *edge* pada *quantity graph* dan *impact graph*

Bobot *Edge* Pada *Quantity Graph*						**Bobot *Edge* Pada *Impact Graph***					
	x_1	x_2	x_3	x_4	x_5		x_1	x_2	x_3	x_4	x_5
x_1	-	2,333	2,333	-	-	x_1	-	13,508	13,508	-	-
x_2	2,333	-	2,867	0,533	0,533	x_2	13,508	-	17,062	3,554	3,554
x_3	2,333	2,867	-	0,533	0,533	x_3	13,508	17,062	-	3,554	3,554
x_4	-	0,533	0,533	-	0,533	x_4	-	3,554	3,554	-	3,554
x_5	-	0,533	0,533	0,533	-	x_5	-	3,554	3,554	3,554	-

Fitur 7 :

Merupakan nilai sosiabilitas dari seorang *author* x dalam topik z pada periode pengujian t_n berdasarkan *quantitative graph*, dengan persamaan (3.10).

$$f_{7,z}(x) = \boldsymbol{CoAut}(x, z) \tag{3.10}$$

Menggunakan contoh studi kasus, maka diperoleh fitur 7 sebesar :

$f_7(x_1) = 2$ $\quad$ $f_7(x_3) = 4$ $\quad$ $f_7(x_5) = 3$

$f_7(x_2) = 4$ $\quad$ $f_7(x_4) = 3$

Fitur 8 :

Merupakan bobot dampak kolaborasi, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author* x dalam topik z pada periode pengujian t_n dalam *quantitative graph*, yang ditunjukkan dengan persamaan (3.11).

$$f_{8,z}(x) = \sum_{\forall y \in C_x} \boldsymbol{WE}_{quan,z}(x, y) \tag{3.11}$$

Menggunakan contoh studi kasus, maka diperoleh fitur 8 sebesar :

$f_8(x_1) = 13{,}508 + 13{,}508 = 27{,}016$

$f_8(x_2) = 13{,}508 + 17{,}062 + 3{,}554 + 3{,}554 = 37{,}678$

$f_8(x_3) = 13{,}508 + 17{,}062 + 3{,}554 + 3{,}554 = 37{,}678$

$f_8(x_4) = 3{,}554 + 3{,}554 + 3{,}554 = 10{,}662$

$f_8(x_5) = 3{,}554 + 3{,}554 + 3{,}554 = 10{,}662$

Fitur 9 :

Merupakan nilai sentralitas dari seorang *author*, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author* x dalam topik z pada periode pengujian t_n, dan dinormalisasikan dengan bobot *edge* maksimum λ dalam *quantitative graph*, yang ditunjukkan dengan persamaan (3.12).

$$\boldsymbol{f}_{\boldsymbol{9,z}}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{\lambda}} \cdot \sum_{\forall \boldsymbol{y} \in \boldsymbol{C_x}} \boldsymbol{WE}_{\boldsymbol{quan,z}}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.12}$$

Menggunakan contoh studi kasus, maka diperoleh fitur 9 sebesar :

$f_9(x_1) = \frac{1}{17{,}062} \times (13{,}508 + 13{,}508) = 1{,}583$

$f_9(x_2) = \frac{1}{17{,}062} \times (13{,}508 + 17{,}062 + 3{,}554 + 3{,}554) = 2{,}208$

$f_9(x_3) = \frac{1}{17{,}062} \times (13{,}508 + 17{,}062 + 3{,}554 + 3{,}554) = 2{,}208$

$f_9(x_4) = \frac{1}{17{,}062} \times (3{,}554 + 3{,}554 + 3{,}554) = 0{,}625$

$f_9(x_5) = \frac{1}{17{,}062} \times (3{,}554 + 3{,}554 + 3{,}554) = 0{,}625$

Fitur 10 :

Merupakan bobot dampak kolaborasi, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author* x dalam topik z pada periode pengujian t_n dalam *quantitative graph*, yang ditunjukkan dengan persamaan (3.13).

$$\boldsymbol{f}_{\boldsymbol{10,z}}(\boldsymbol{x}) = \boldsymbol{CoAut}(\boldsymbol{x}, \boldsymbol{z}) \cdot \sum_{\forall \boldsymbol{y} \in \boldsymbol{C_x}} \boldsymbol{WE}_{\boldsymbol{quan,z}}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.13}$$

Fitur 11 :

Merupakan bobot dampak kolaborasi, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author* x dalam topik z pada periode pengujian t_n dalam *impact graph*, yang ditunjukkan dengan persamaan (3.14).

$$f_{11,z}(x) = \sum_{\forall y \in C_x} WE_{imp,z}(x,y) \quad (3.14)$$

Fitur 12 :

Merupakan nilai sentralitas dari seorang *author*, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author x* dalam topik *z* pada periode pengujian t_n, dan dinormalisasikan dengan bobot *edge* maksimum λ dalam *impact graph*, yang ditunjukkan dengan persamaan (3.15) :

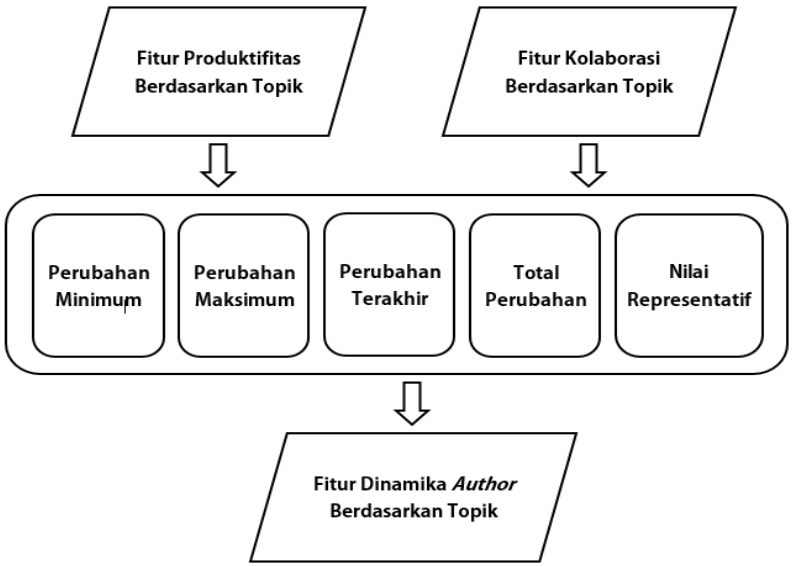
$$f_{12,z}(x) = \frac{1}{\lambda} . \sum_{\forall y \in C_x} WE_{imp,z}(x,y) \quad (3.15)$$

Fitur 13 :

Merupakan bobot dampak kolaborasi, yang diukur berdasarkan jumlah bobot *edge* dari semua kolaborasi yang dilakukan oleh *author x* dalam topik *z* pada periode pengujian t_n dalam *impact graph*, yang ditunjukkan dengan persamaan (3.16).

$$f_{13,z}(x) = CoAut(x,z) . \sum_{\forall y \in C_x} WE_{imp,z}(x,y) \quad (3.16)$$

3.3.4 Ekstraksi Fitur Dinamika *Author* Berdasarkan Topik



Gambar 3.8. Proses ekstraksi fitur dinamika *author* berdasarkan topik

Untuk menangkap kedinamisan dalam kinerja penelitian *author*, penelitian ini mendefinisikan beberapa indikator perubahan yang diterapkan pada masing-masing fitur yang telah dijelaskan sebelumnya. Seperti yang ditunjukan pada Gambar 3.8, terdapat lima macam fitur dinamika *author*, diantaranya : perubahan minimum, perumahan maksimum, perubahan terakhir, total perubahan, dan nilai representative dari setiap fitur yang terbentuk.

Semua fitur yang dihasilkan dari proses ekstraksi fitur produktifitas berdasarkan topik dan ekstraksi fitur kolaborasi berdasarkan topik menghasilkan nilai pertahun dari setiap fiturnya sebanyak tahun pengujian yang dilakukan. Jika dilakukan pengujian dari tahun 2011 hingga tahun 2015 pada masing-masing *author*, maka *author* tersebut memiliki 5 baris nilai pada setiap fitur sesuai tahun pengujiannya. Oleh karena itu untuk menangkap nilai dinamika pada setiap *author* dilakukan lima macam pengujian sebagai berikut :

1. Perubahan Minimum

   Perhitungan perubahan minimum dari semua tahun pengujian pada setiap fitur yang dihasilkan bertujuan untuk menangkap nilai minimum dari kinerja seorang *author* sepanjang tahun, yang ditunjukkan pada persamaan (3.22).

$$\boldsymbol{PerMin_f = min_{t_k \in (t_1,t_n)} \{(ft_k - ft_{k-1}), \dots, (ft_n - ft_{n-1})\}} \quad (3.22)$$

2. Perubahan Maksimum

   Perhitungan perubahan maksimum dari semua tahun pengujian pada setiap fitur yang dihasilkan bertujuan untuk menangkap nilai maksimum dari kinerja seorang *author* sepanjang tahun, yang ditunjukkan pada persamaan (3.23).

$$\boldsymbol{PerMaks_f = max_{t_k \in (t_1,t_n)} \{(ft_k - ft_{k-1}), \dots, (ft_n - ft_{n-1})\}} \quad (3.23)$$

3. Perubahan Terakhir :

   Perhitungan perubahan terakhir dari semua tahun pengujian pada setiap fitur yang dihasilkan bertujuan untuk menangkap kinerja terakhir yang dilakukan *author* sepanjang tahun pengujian, yang ditunjukkan pada persamaan (3.24).

$$\boldsymbol{PerAkhir_f = ft_n - ft_{n-1}} \quad (3.24)$$

4. Total Perubahan :

   Perhitungan total perubahan dari semua tahun pengujian pada setiap fitur yang dihasilkan bertujuan untuk menangkap kinerja yang dilakukan *author* secara keseluruhan selama periode pengujian, yang ditunjukkan pada persamaan (3.25).

$$\boldsymbol{PerTotal}_{\boldsymbol{f}} = \sum_{\boldsymbol{t_k = t_1}}^{\boldsymbol{t_n}} (\boldsymbol{ft_k} - \boldsymbol{ft_{k-1}}) + \ldots + (\boldsymbol{ft_n} - \boldsymbol{ft_{n-1}}) \tag{3.25}$$

5. Nilai Representatif :

   Perhitungan nilai representative pada setiap fitur yang dihasilkan bertujuan untuk menangkap kinerja yang dilakukan *author* secara keseluruhan selama. Nilai representatif dibedakan menurut jenis fitur yang telah dihasilkan pada tahap sebelumnya. fitur dari nilai pada periode pengujian yaitu fitur 1 dan fitur 4 dihitung dengan persamaan (3.26), fitur dari nilai kumulatif hingga periode pengujian yaitu fitur 2 dan fitur 5 dihitung dengan persamaan (3.27), begitu juga dengan fitur dari nilai pembobotan yang mempertimbangkan dampak penurunan waktu yaitu fitur 3, fitur 6, fitur 7, fitur 8, fitur 9, fitur 10, fitur 11, fitur 12, dan fitur 13 yang dihitung dengan persamaan (3.28),

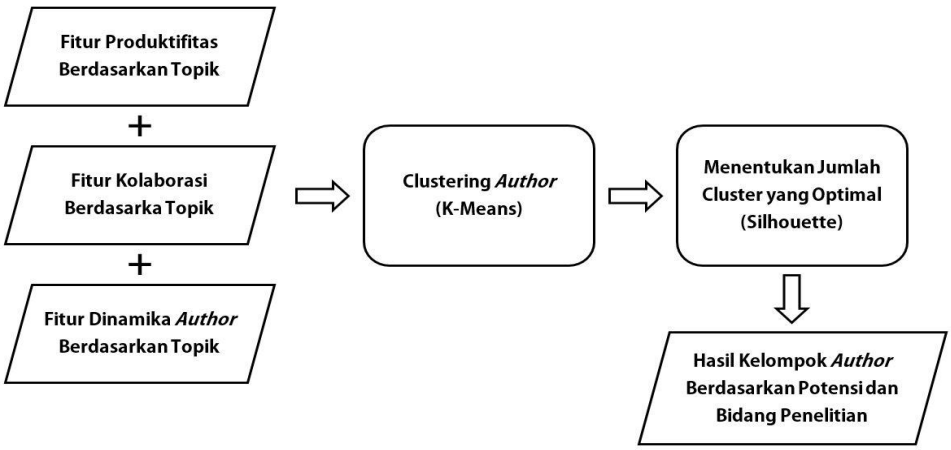
$$RepVal_{f,a} = \sum_{t_k = t_0}^{t_n} \frac{ft_k}{n} \tag{3.26}$$

$$RepVal_{f,b} = \frac{ft_n}{n} \tag{3.27}$$

$$RepVal_{f,c} = ft_n \tag{3.28}$$

3.3.5 Pengelompokan Potensi *Author*

Setelah didapatkan fitur-fitur dari proses ekstraksi fitur produktifitas *author* berdasarkan topik, ekstraksi fitur kolaborasi *author* berdasarkan topik, dan ekstraksi fitur dinamika *author* berdasarkan topik, seperti yang dijelaskan pada Gambar 3.9, semua fitur tersebut digunakan untuk pengelompokan potensi penulis dengan metode unsupervised yaitu *clustering*.



Gambar 3.9. Proses pengelompokan potensi *author* berdasarkan topik

Algoritma *clustering* yang digunakan adalah K-Means++. *Cosine similarity* digunakan untuk pengukuran jarak antar *author* dalam penentuan *cluster author*. Kemudian langkah terakhir, untuk mendapatkan hasil kelompok *author* berdasarkan potensi dan bidang penelitian, hasil *cluster* tersebut dianalisa dengan *Silhouette* yang dijelaskan pada bab 2.6 untuk mendapatkan jumlah *cluster* yang optimal dari hasil percobaan tersebut.

3.4 Pembuatan dan Implementasi Sistem

Pada tahapan ini dilakukan pembuatan dan implementasi sistem yang telah dibentuk sebelumnya ke dalam kode program sehingga dapat dijalankan pada komputer. Sistem aplikasi dibangun dalam bahasa pemrograman phyton dan penyimpanan data informasi *author* menggunakan database mysql.

3.5 Uji Coba Sistem

Pengujian akan dilakukan dengan beberapa sekenario untuk menunjukkan kelebihan dan kekurangan dari metode yang diusulkan :

1. Proses ekstraksi topik dalam artikel ilmiah

   Proses ekstraksi topik dilakukan dengan cara mengkluster artikel ilmiah dengan menerapkan jumlah *cluster* 2 sampai 35 pada masing-masing metode *clustering* sebagai berikut :

a. K-Means++ Clustering
b. Birch Clustering
c. AverageLink Agglomerative Clustering
d. CompleteLink Agglomerative Clustering
e. SingleLink Agglomerative Clustering

2. Proses pengelompokan potensi penulis

   Proses pengelompokan potensi penulis dilakukan dengan metode *clustering* K-Means ++ menggunakan fitur sebagai berikut :

   a. Fitur produktivitas *author* berdasarkan topik
   b. Fitur kolaborasi *author* berdasarkan topik
   c. Fitur dinamika *author* berdasarkan topik

3.6 Evaluasi dan Analisa Hasil

Mengingat metode yang penulis gunakan untuk mengelompokkan potensi penulis adalah metode *clustering*, maka evaluasi dan analisa terhadap hasil uji coba metode penelitian ini dilakukan dengan menghitung koherensi terhadap intra dan inter *cluster* yang terbentuk. Untuk menghitung kekonsistenan hasil kelompok *author*, penulis malakukan sekenario uji coba pada data *author* dari tahun 2000 hingga 2009 untuk pengelompokan potensi *author* dengan semua hasil ekstraksi fitur dan serangkaian proses yang diusulkan. Serta membandingkan hasil potensi kelompok *author* tersebut dengan data *author* pada tahun 2010 hingga 2016.

3.7 Penulisan Laporan

Pada tahap ini dilakukan pendokumentasian berupa laporan akhir dari proses yang telah diimplementasikan dan diuji coba berikut perancangan sistem dan teori penunjang selama proses penelitian. Laporan akhir yang disusun bertujuan untuk memberikan penjelasan terstruktur dan gambaran dari pengerjaan penelitian, serta sebagai publikasi sehingga dapat menunjang penelitian selanjutnya.

3.8 Jadwal Kegiatan Penelitian

Pada tahap dokumentasi sistem ini akan dilakukan penulisan laporan hasil penelitian dari setiap tahapan yang dilakukan. Tujuan dari tahapan ini adalah menghasilkan dokumentasi tertulis dari penelitian yang dilakukan. Jadwal penelitian yang dilakukan dapat dilihat pada Tabel 3.10.

Tabel 3.10. Jadwal Rencana Kegiatan Penelitian

Kegiatan	Bulan-1				Bulan-2				Bulan-3				Bulan-4			
Studi Literatur	■	■	■	■	■	■	■	■								
Perancangan Sistem				■	■	■	■	■								
Pembuatan Perangkat Lunak				■	■	■	■	■	■	■	■	■	■			
Uji Coba dan Analisa Hasil									■	■	■	■	■	■	■	■
Penulisan Laporan					■	■	■	■	■	■	■	■	■	■	■	■

BAB 4
UJI COBA DAN ANALISA HASIL

Bab ini akan membahas tentang implementasi, pengujian, dan pembahasan terkait metodologi penelitian yang diusulkan. Tahapan implementasi yang dilakukan sesuai dengan alur pada Gambar 3.2 yang terdiri dari lima proses utama, yaitu ekstraksi topik pada artikel ilmiah, ekstraksi fitur produktifitas berdasarkan topik, ekstraksi fitur kolaborasi berdasarkan topik, ekstraksi fitur dinamika *author* berdasarkan topik, dan pengelompokan potensi penulis dengan cara *clustering*.

Tahap berikutnya adalah pengujian dan analisis hasil dari implementasi yang telah dilakukan. Skenario pengujian akan dilakukan pada beberapa skenario sesuai dengan yang telah direncanakan pada sub-bab 3.5 tentang perancangan pengujian. Pembahasan tekahir dalam Bab ini adalah diskusi tentang hasil dan evaluasi pada metode usulan, yaitu hasil dan kinerja yang dihasilkan oleh proses ekstraksi fitur berdasarkan topik pada artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis.

4.1 Perangkat Pengujian

Untuk melakukan implementasi dan pengujian Hypergraph Partition untuk pengelompokan penulis, penulis menggunakan beberapa perangkat yang terdiri dari perangkat keras dan perangkat lunak.

1. Perangkat Keras

   Perangkat keras yang digunakan untuk implementasi dan pengujian adalah tiga buah PC dengan spesifikasi processor Intel Core i5-7400 3.00 GHz, RAM 8.00 Gb, kapasitas *harddisk* 1 Tb, dan menggunakan sistem operasi Windows 10 64-bit.

2. Perangkat Lunak

   Perangkat lunak yang digunakan pada tahapan implementasi dan evaluasi pengujian adalah aplikasi Anaconda-Spyder dengan Python versi 3.6, basis data *MySQL Workbench* 6.3 CE.

4.2 Implementasi Sistem

4.2.1 Persiapan Dataset

Dataset yang digunakan dalam penelitian ini adalah AMiner dataset, yang memiliki data artikel sejumlah 2.385.066, dengan jumlah *author* beserta *co-author* sebanyak 4.858.661, data sitasi artikel sejumlah 9.671.893, dan tahun publikasi *author* dari tahun 1936 hingga tahun 2016. Observasi awal dilakukan pada AMiner dataset dengan melakukan analisis jumlah publikasi masing-masing *author*. Dikarenakan penelitian ini bertujuan untuk mengelompokkan *author* berdasarkan topik penelitian dan potensinya, untuk mendapatkan variasi dan distribusi yang seimbang pada potensi *author*, penulis melakukan 4 proses dalam persiapan dataset. Proses pertama yaitu, memilih data berdasarkan jumlah publikasi artikel dengan range jumlah publikasi *author* sebagai berikut :

1. *Author* yang memiliki publikasi sebanyak 25-50 artikel
2. *Author* yang memiliki publikasi sebanyak 51-75 artikel
3. *Author* yang memiliki publikasi sebanyak 76-100 artikel
4. *Author* yang memiliki publikasi sebanyak 101-125 artikel
5. *Author* yang memiliki publikasi sebanyak 126-150 artikel
6. *Author* yang memiliki publikasi sebanyak 151-175 artikel
7. *Author* yang memiliki publikasi sebanyak lebih dari 175 artikel

Berikut adalah hasil analisa dari jumlah publikasi *author* dan jumlah *author* pada masing-masing range yang telah disebutkan diatas, dalam AMiner dataset, yang ditunjukkan pada Table 4.1.

Dari hasil analisis data *author* yang telah dilakukan, penelitian ini memilih 525 *author* yang memiliki publikasi 25-50 artikel, 525 *author* yang memiliki publikasi 51-75 artikel, 525 *author* yang memiliki publikasi 76-100 artikel, 525 *author* yang memiliki publikasi 101-125 artikel, 525 *author* yang memiliki publikasi 126-150 artikel, 331 *author* yang memiliki publikasi 151-175 artikel, dan 525 *author* yang memiliki publikasi lebih dari 175 artikel disemua tahun, dengan total estimasi 3.481 *author* dan 296.341 artikel.

Tabel 4.1. Analisis *author* berdasarkan jumlah publikasi artikel

Jumlah Publikasi *Author*	**Jumlah *Author***	**Jumlah *Author* yang digunakan**
25 - 50	18.795	525
51 - 75	5.132	525
76 - 100	2.119	525
101 - 125	1.109	525
126 - 150	594	525
151 - 175	331	331
> 175	686	525
Total	**28.766**	**3.481**

Terdapat kekurangan informasi mengenai judul artikel, abstrak, tahun publikasi, dan *id author* di dalam AMiner dataset. Oleh karena itu, tiga tahap pemrosesan selanjutnya bertujuan untuk mendapatkan semua informasi yang lengkap tentang data artikel, sitasi, dan *co-authorship* pada masing-masing *author* yang telah ditentukan dalam proses pertama. Proses kedua dalam pemilihan dataset adalah mencari artikel ilmiah yang dipublikasikan oleh masing-masing *author* yang memiliki informasi *id paper*, judul artikel, dan abstrak secara lengkap, dengan total artikel ilmiah sebanyak **296.341 artikel**, dimana semua artikel tersebut akan digunakan dalam proses ekstraksi topik dalam artikel ilmiah. Berikut contoh beberapa data artikel ilmiah yang ditunjukkan pada Table 4.2.

Proses ketiga dalam pemilihan dataset adalah mencari publikasi beserta *co-authorship* pada masing-masing *author*. Dalam proses ini hanya menyimpan data yang memiliki informasi *id paper*, tahun publikasi artikel, dan daftar *id author* yang menulis masing-masing artikel ilmiah secara lengkap. Dari proses seleksi dataset ketiga meghasilkan data *author* sebanyak **3.465 *author*** dengan **296.054 artikel ilmiah** dan **186.444 *co-authorship***, semua data tersebut akan digunakan dalam proses ekstraksi fitur produktivitas *author* berdasarkan topik dan proses ekstraksi fitur kolaborasi *author* berdasarkan topik. Berikut contoh beberapa data *author paper* yang ditunjukkan pada Tabel 4.3.

Tabel 4.2. Contoh data artikel ilmiah

id_paper	Judul	Abstrak
553	On the optimal nesting order for computing N relational joins	Using the nested loops method, this paper addresses the problem of minimizing the number of page fetches necessary to evaluate a given query to a relational database. We first propose a data structure whereby the number of page fetches required for query evaluation is substantially reduced and then derive a formula for the expected number of page fetches. An optimal solution to our problem is the nesting order of relations in the evaluation program, which minimizes the number of page fetches. Since the minimization of the formula is NP-hard, as shown in the Appendix, we propose a heuristic algorithm which produces a good suboptimal solution in polynomial time. For the special case where the input query is a “tree query,” we present an efficient algorithm for finding an optimal nesting order.
249	A heuristic for deriving loop functions	The problem of analyzing an initialized loop and verifying that the program computes some particular function of its inputs is addressed. A heuristic technique for solving these problems is proposed that appears to work well in many commonly occurring cases. The use of the technique is illustrated with a number of applications. An attribute of initialized loops is identified that corresponds to the ``effort" required to apply this method in a deterministic (i.e., guaranteed to succeed) manner. It is explained that in any case, the success of the proposed heuristic relies on the loop exhibiting a ``reasonable" form of behavior.
320	Highly available systems for database applications	As users entrust more and more of their applications to computersystems, the need for systems that are continuously operational (24hours per day) has become even greater. This paper presents asurvey and analysis of representative architectures and techniquesthat have been developed for constructing highly available systemsfor database applications. It then proposes a design of adistributed software subsystem that can serve as a unifiedframework for constructing database application systems that meetvarious requirements for high availability.

Tabel 4.3. Contoh data *author paper*

id_author	**Nama *Author***	***id_paper***	**Tahun Publikasi**	**Daftar *Co- Author***
3	Wenhu Wu	1696382	2006	Jian Liu, Thomas Fang Zheng, Wenhu Wu
4	Zhiyuan Zeng	1163076	2008	Bo Li, Zhiyuan Zeng, Jianzhong Zhou, Mu Zhou
16	Nadia Dahmani	2063320	2014	Nadia Dahmani, François Clautiaux, Saoussen Krichen, El-Ghazali Talbi
18	Z. Liu	2330044	2011	Z. Liu, M. -T. Sun, C. -W. Lin, Z. Zhang, Z. Liu, H. H. Chen, Y. -P. Tan, O. C. Au
19	Ren-Hong Wang	891103	2006	Xue-Zhang Liang, Ren-Hong Wang, Li-Hong Cui, Jie-Lin Zhang, Ming Zhang

Proses keempat dalam pemilihan dataset adalah mencari sitasi dan tahun sitatasi pada masing-masing artikel ilmiah yang telah terpilih pada proses ketiga. Dalam proses ini hanya menyimpan data yang memiliki informasi *id paper*, *id citation*, dan tahun sitasi yang diterima oleh masing-masing artikel ilmiah secara lengkap. Dari proses seleksi dataset keempat meghasilkan data artikel ilmiah sebanyak **168.692 artikel** dengan **1.929.411 sitatsi**, dimana semua data tersebut akan digunakan dalam proses ekstraksi fitur produktivitas *author* berdasarkan topik dan proses ekstraksi fitur kolaborasi *author* berdasarkan topik. Berikut contoh beberapa data *paper citation* yang ditunjukkan pada Tabel 4.4.

Pemilihan dataset dengan empat proses tersebut menghasilkan sejumlah dataset yang akan digunakan dalam ujicoba dan evaluasi metode usulan, yaitu ekstraksi fitur berdasarkan topik pada artikel ilmiah untuk pengelompokan potensi penulis dalam jaringan kolaborasi dinamis. Dataset tersebut dibagi menjadi dua bagian, yaitu sebagai data latih (*training*) dan data uji (*testing*).

Tabel 4.4. Contoh data *paper citation*

id_paper	***id_citation***	**Tahun Sitasi**
134	4682	1986
134	65980	1989
134	77960	1989
134	285184	1985
134	290545	1988

Dalam penelitian ini, proses *training* akan dilakukan pada data artikel dan data sitasi dengan tahun publikasi dan tahun sitasi 2000 hingga tahun 2009, serta akan mengakumulasikan data artikel dan data sitasi dengan tahun kurang dari 2000. Berikut ini merupakan penjabaran dari distribusi data latih artikel ilmiah dan data sitasi berdasarkan tahun publikasi dan tahun sitasi, yang akan dipaparkan secara mendetail pada Tabel 4.5 dan Tabel 4.6.

Tabel 4.5. Data latih artikel berdasarkan tahun publikasi

Tahun Publikasi	**Jumlah Artikel**	**Tahun Publikasi**	**Jumlah Artikel**
< 2000	27.644	2005	18.271
2000	5.438	2006	21.104
2001	6.609	2007	20.669
2002	9.030	2008	23.515
2003	9.322	2009	31.411
2004	10.958	**Total Artikel**	**183.971**

Tabel 4.6. Data latih sitasi berdasarkan tahun sitasi

Tahun Sitasi	Jumlah Data Sitasi	Tahun Sitasi	Jumlah Data Sitasi
< 2000	99.727	2005	100.999
2000	21.662	2006	114.377
2001	28.653	2007	130.025
2002	41.955	2008	137.201
2003	42.970	2009	186.330
2004	54.011	**Total Data Sitasi**	**957.910**

Tabel 4.7. Data uji artikel berdasarkan tahun publikasi

Tahun Publikasi	Jumlah Artikel	Tahun Publikasi	Jumlah Artikel
2010	25.246	2014	15.322
2011	23.856	2015	4.490
2012	23.985	2016	191
2013	18996	**Total Artikel**	**112.086**

Tabel 4.8. Data uji sitasi berdasarkan tahun sitasi

Tahun Sitasi	Jumlah Data Sitasi	Tahun Sitasi	Jumlah Data Sitasi
2010	186.862	2014	179.965
2011	184.978	2015	38.520
2012	198.389	2016	2.792
2013	179.995	**Total Data Sitasi**	**971.501**

Dalam proses *testing* untuk menguji dan mengevaluasi metode usulan, penulis menggunakan data artikel ilmiah dan data sitasi dengan tahun publikasi dan tahun sitasi 2010 hingga tahun 2016. Berikut ini merupakan penjabaran dari distribusi data uji artikel ilmiah dan data sitasi berdasarkan tahun publikasi dan tahun sitasi, yang akan dipaparkan secara mendetail pada Tabel 4.7 dan Tabel 4.8.

4.2.2 Ekstraksi Topik dalam Artikel Ilmiah

Proses awal pada penelitian ini adalah ekstraksi topik dalam artikel ilmiah. Ekstraksi topik didapatkan dari proses *clustering* terhadap dataset yang telah diolah dengan empat proses pada tahap persiapan dataset. Adapun hasil yang didapatkan adalah informasi judul dan abstrak artikel ilmiah. Selanjutnya agar teks yang ada pada artikel ilmiah mudah untuk dibaca dan diolah oleh sistem, beberapa tahap preprosesing akan diterapkan terhadap data artikel ilmiah tersebut. Berikut ini adalah penjelasan dan contoh hasil dari preprosesing teks :

1. *Tokenizing*, *Tokenizing* memiliki peran untuk memecah teks dalam artikel ilmiah menjadi unit terkecil kata-kata atau term.
2. *Case Folding*, *Case folding* memiliki peran untuk merubah term dalam artikel ilmiah menjadi huruf kecil (*lowercase*).
3. *Stopwords Removal*, *Stopwords removal* memiliki peran untuk menghapus kata henti dan tanda baca dalam bahasa inggris.
4. *Lemmatization*, *Lemmatization* memiliki peran untuk merubah term-term dalam Bahasa inggris menjadi bentuk dasar.

Tabel 4.9 menunjukkan contoh salah satu abstrak pada artikel ilmiah berserta hasil preprosesing teks tanpa proses *lemmatization* dan dengan *lemmatization*. Terlihat bahwa dengan penerapan proses *lemmatization* akan merusak arti pada beberapa kata. Sebagai contoh kata "*reliability between a pair of nodes*" terdeteksi benar tanpa proses *lemmatization* dengan hasil "*reliability pair nodes*" sedangkan jika dengan proses *lemmatization* menghasilkan "*reliability pair nod*". Sehingga kata "*nodes*" kehilangan makna yang sebenarnya jika menerapkan proses *lemmatization* dalam tahap *preprocessing*.

Tabel 4.9. Contoh hasil preposesing artikel ilmiah

Keterangan	Hasil
Text dalam artikel ilmiah	Reliability Optimization in the Design of Distributed Systems : The reliability of a distributed system depends on the reliabilities of its communication links and computing elements, as well as on the distribution of its resources, such as programs and data files. A useful measure of reliability in distributed systems is the terminal reliability between a pair of nodes which is the probability that at least one communication path exists between these nodes. An interesting optimization problem is that of maximizing the terminal reliability between a pair of computing elements under a given budget constraint. Analytical techniques to solve this problem are applicable only to special forms of reliability expressions. In this paper, three iterative algorithms for terminal reliability maximization are presented. The first two algorithms require the computation of terminal reliability expressions, and are therefore efficient for only small networks. The third algorithm, which is developed for large distributed systems, does not require the computation of terminal reliability expressions; this algorithm maximizes approximate objective functions and gives accurate results. Several examples are presented to illustrate the approximate optimization algorithm and an estimation of the error involved is also given.
Hasil Preprosesing tanpa proses lemmatization	reliability optimization design distributed systems reliability distributed system depends reliabilities communication links computing elements well distribution resources programs data files useful measure reliability distributed systems terminal reliability pair nodes probability least one communication path exists nodes interesting optimization problem maximizing terminal reliability pair computing elements given budget constraint analytical techniques solve problem applicable special forms reliability expressions paper three iterative algorithms terminal reliability maximization presented first two algorithms require computation terminal reliability expressions therefore efficient small networks third algorithm developed large distributed systems require

	computation terminal reliability expressions algorithm maximizes approximate objective functions gives accurate results several examples presented illustrate approximate optimization algorithm estimation error involved also given
Hasil Preprosesing dengan proses lemmatization	reliability optimization design distribute systems reliability distribute system depend reliabilities communication link compute elements well distribution resources program data file useful measure reliability distribute systems terminal reliability pair nod probability least one communication path exist nod interest optimization problem maximize terminal reliability pair compute elements give budget constraint analytical techniques solve problem applicable special form reliability expressions paper three iterative algorithms terminal reliability maximization present first two algorithms require computation terminal reliability expressions therefore efficient small network third algorithm develop large distribute systems require computation terminal reliability expressions algorithm maximize approximate objective function give accurate result several examples present illustrate approximate optimization algorithm estimation error involve also give

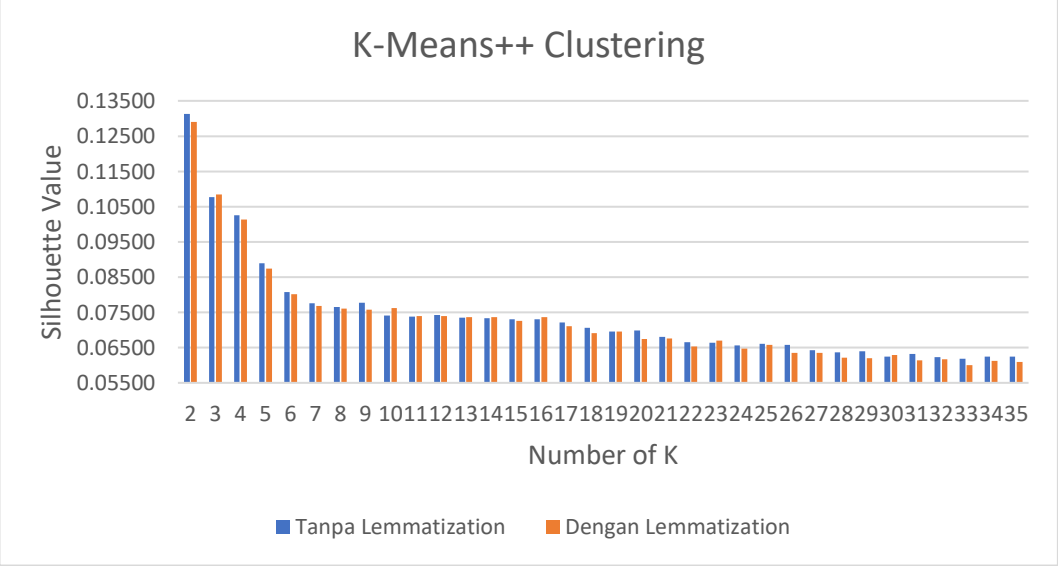
Setelah dilakukan tahap *preprosesing* teks dalam artikel, langkah selanjutnya yaitu melakukan proses representasi kata kedalam bentuk vector. Penelitian ini menggunakan Word2Vec sebagai pembentukan *vector space model*, yang merupakan salah satu proses dalam *word embedding*. Proses detail tentang Word2Vec telah dijelaskan dalam sub-bab 2.4. Setelah didapatkan vektor yang berisi pembobotan dari setiap term pada artikel, vektor tersebut akan dijadikan fitur yang mewakili setiap artikel ilmiah dalam proses *clustering*.

Pada penelitian ini penulis melakukan beberapa sekenario dalam proses *clustering* artikel ilmiah. Penulis membandingkan *clustering* artikel ilmiah dengan metode k-means++ *clustering*, metode birch *clustering*, metode averagelink agglomerative *clustering*, metode completelink agglomerative *clustering*, metode singlelink agglomerative *clustering*, dengan melakukan percobaan pada jumlah *cluster* yang berbeda pada masing-masing percobaan.

Adapun jumlah klaster yang diuji coba adalah 2 sampai dengan 35 klaster. *Cosine similarity* juga digunakan untuk pengukuran jarak antar artikel. Kemudian hasil dari *cluster* tersebut dilakukan analisa dengan *Silhouette* untuk menentukan jumlah *cluster* yang optimal terhadap kumpulan artikel tersebut. Berikut penjelasan secara mendetail mengenai masing-masing metode *clustering* dan hasil percobaan.

A. K-Means++ Clustering

K-Means++ merupakan metode pengembangan dari metode K-Means untuk mengelompokkan dokumen [13]. Pada umumnya metode K-Means memilih data titik pusat klaster secara acak. Hal tersebut menyebabkan metode K-Means dapat menghasilkan solusi yang sub-optimal. Oleh karena itu, K-Means++ mengatasi permasalahan tersebut dengan mengoptimasi pemilihan titik pusat klaster awal sebelum metode K-Means dijalankan. Pada umumnya K-Means dan K-Means++ menggunakan pendekatan VSM (*Vector Space Model*), dimana dokumen dimodelkan dalam vektor yang memiliki kata sebagai fitur. Setelah kumpulan dokumen telah melewati tahap pra-pemrosesan teks, seluruh kata pada kumpulan dokumen diekstrak untuk dijadikan fitur dokumen, pendekatan ini disebut juga *"bag-of-words model"*. Dalam percobaan ini penulis melakukan percobaan pada metode K-Means++ dengan proses *lemmatization* dan *non-lemmatization*.

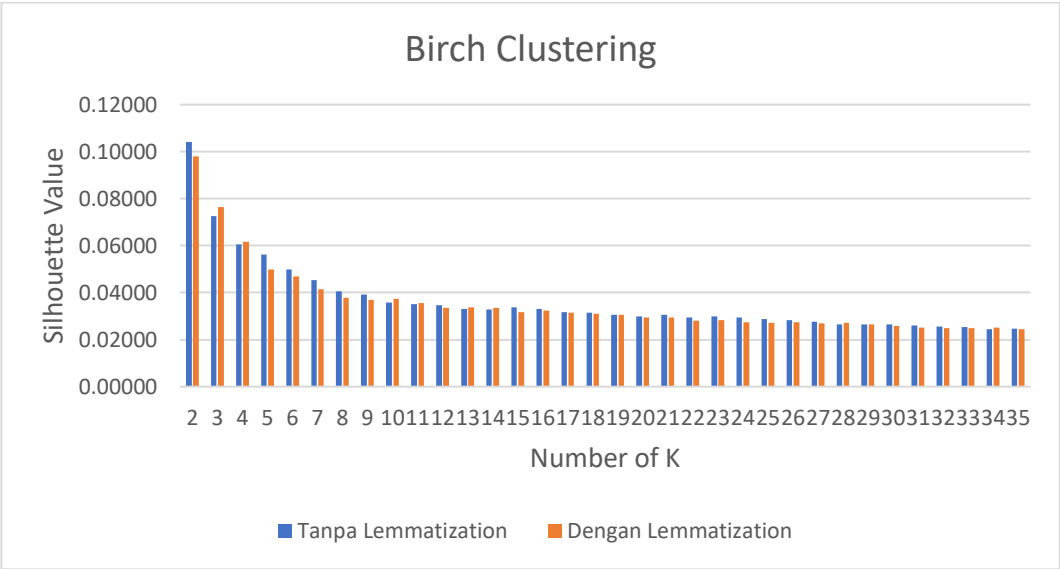


Gambar 4.1. Hasil *clustering* artikel ilmiah dengan K-Means++

Gambar 4.1 menunjukan perbandingan hasil *clustering* artikel ilmiah dengan menggunakan lemmatization proses dan tanpa menggunakan lemmatization proses pada tahap preposesing teks. Pada skenario percobaan ini, penulis menggunakan algoritma *clustering* K-Means++ pada jumlah k = 2 hingga k = 35. Hasil uji coba menunjukkan bahwa hasil *clustering* pada k = 2 sampai k = 8 mengalami penurunana seiring bertambahnya jumlah *cluster*, namun hasil *cluster* menunjukkan kenaikan yang signifikan pada jumlah k = 9 dengan nilai silhouette sebesar 0,07766. Secara keseluruhan hasil *clustering* tanpa menggunakan lemmatization proses lebih unggul daripada dengan menggunakan lemmatization pada tahap preposesing teks.

B. Birch Clustering

BIRCH (*balanced iterative reducing and clustering using hierarchies*) adalah salah satu metode unsupervised yang digunakan untuk melakukan pengelompokan hierarkikal dalam kumpulan data yang sangat besar. Keuntungan BIRCH adalah kemampuannya dalam mengelompokkan data metrik secara bertahap dan dinamis. BIRCH *clustering* juga mampu untuk multi-dimensi dalam upaya menghasilkan hasil pengelompokan yang efisien dalam batasan memori dan waktu [14].



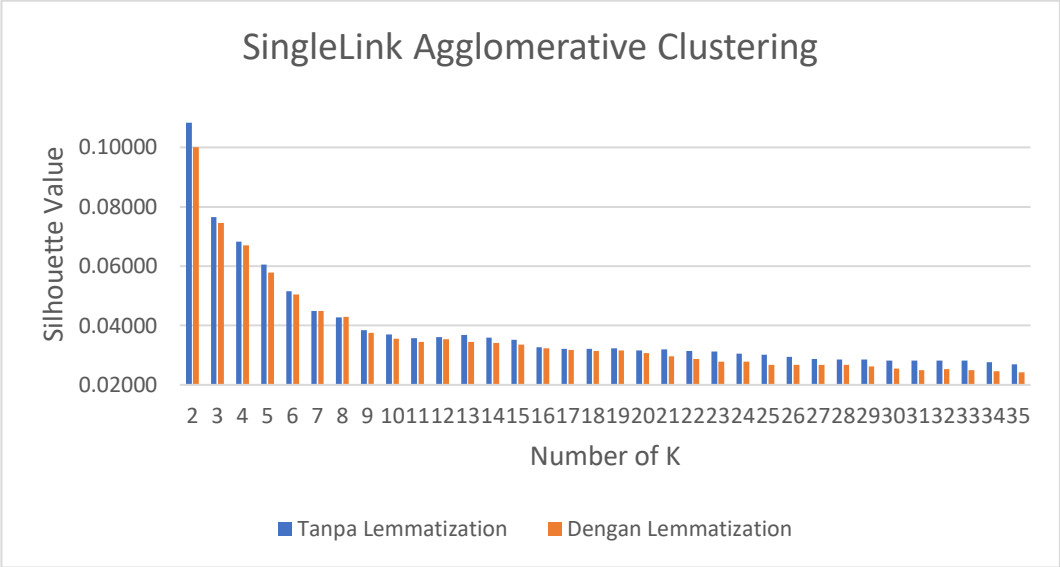
Gambar 4.2. Hasil *clustering* artikel ilmiah dengan Birch

Gambar 4.2 merupakan perbandingan hasil *clustering* artikel ilmiah dengan menggunakan lemmatization proses dan tanpa menggunakan lemmatization proses di tahap preposesing teks. Pada skenario percobaan ini, penulis menggunakan algoritma *clustering* Birch pada jumlah k = 2 hingga k = 35. Hasil uji coba menunjukkan bahwa hasil *clustering* mengalami penurunana seiring bertambahnya jumlah *cluster*, dan beberapa kenaikan seperti pada k = 10 dengan lemmatization dan k = 15 tanpa lemmatization, namun kenaikan tersebut tidak cukup signifikan sehingga tidak dapat menentukan jumlah *cluster* dengan metode *clustering* Birch. Secara keseluruhan hasil *clustering* tanpa menggunakan lemmatization proses sama unggulnya dengan menggunakan lemmatization pada tahap preposesing teks.

C. SingleLink Agglomerative Clustering

Klasterisasi hirarkikal dokumen teks merupakan metode pengelompokan dokumen yang bekerja dengan membangun sebuah hirarki kelompok dokumen atau klaster. Hirarkikal *clustering* dapat disebut juga sebagai dendogram. Klasterisasi hirarkikal telah digunakan untuk pembentukan taksonomi konsep pada suatu teks [15]. Kelebihan klasterisasi hirarkikal dibandingkan dengan metode klasterisasi lain adalah tingkat hirarki yang dapat ditentukan sesuai kebutuhan [16]. Metode single linkage *clustering* didasarkan pada jarak minimum. Jarak antara satu *cluster* dan *cluster* lain diukur berdasarkan obyek yang mempunyai jarak terdekat.

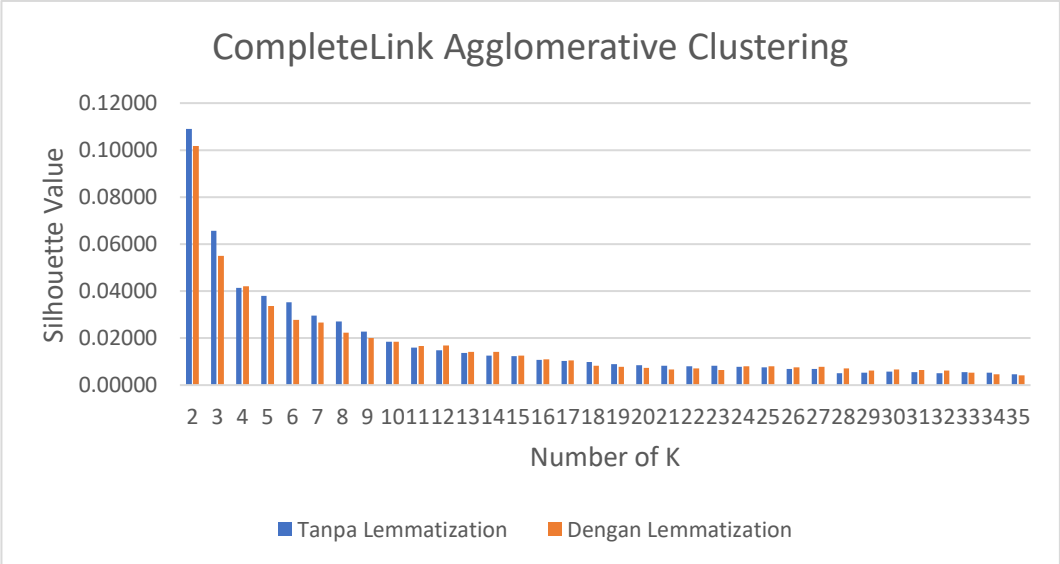
Gambar 4.3 merupakan perbandingan hasil *clustering* artikel ilmiah dengan menggunakan lemmatization proses dan tanpa menggunakan lemmatization proses di tahap preposesing teks. Pada skenario percobaan ini, penulis menggunakan algoritma SingleLink Agglomerative Clustering pada jumlah k = 2 hingga k = 35. Hasil uji coba menunjukkan bahwa hasil *clustering* pada k = 2 sampai k = 12 mengalami penurunana seiring bertambahnya jumlah *cluster*, namun hasil *cluster* menunjukkan kenaikan yang signifikan pada jumlah k = 13 dengan nilai silhouette sebesar 0,03676. Secara keseluruhan hasil *clustering* tanpa menggunakan lemmatization proses lebih unggul daripada dengan menggunakan lemmatization pada tahap preposesing teks.



Gambar 4.3. Hasil *clustering* artikel ilmiah dengan SingleLink

D. CompleteLink Agglomerative Clustering

Metode complete linkage *clustering* didasarkan pada jarak maksimum. Jarak antara satu *cluster* dan *cluster* lain diukur berdasarkan obyek yang mempunyai jarak terjauh. Metode ini sangat ampuh untuk memperkecil *variance* di dalam *cluster*, karena melibatkan centroid pada saat penggabungan antar *cluster*. Metode ini juga baik untuk data yang mengandung outlier.

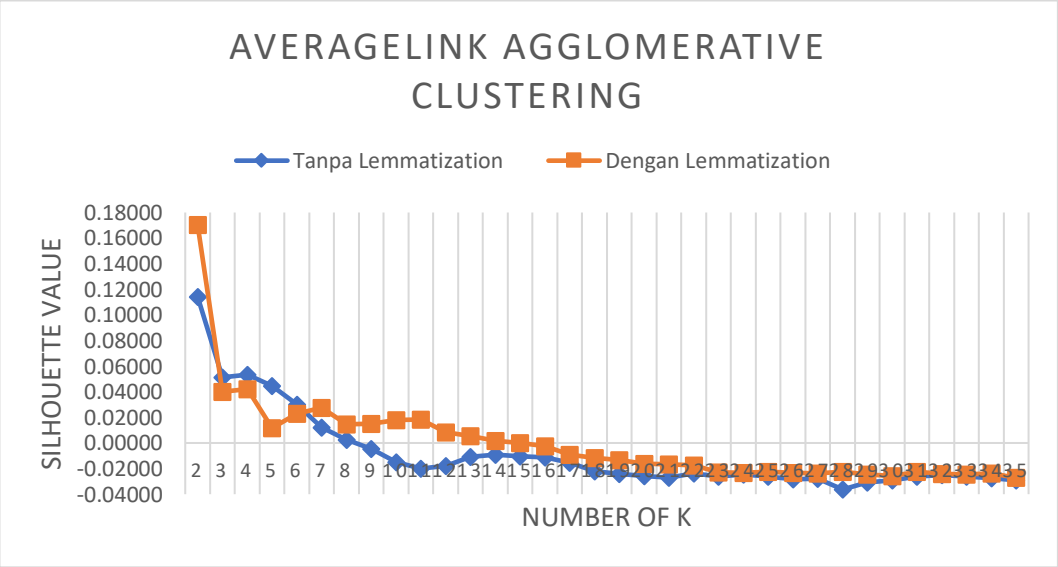


Gambar 4.4. Hasil *clustering* artikel ilmiah dengan CompleteLink

Gambar 4.4 merupakan perbandingan hasil *clustering* artikel ilmiah dengan menggunakan lemmatization proses dan tanpa menggunakan lemmatization proses di tahap preposesing teks. Pada skenario percobaan ini, penulis menggunakan algoritma CompleteLink Agglomerative Clustering dengan k = 2 hingga k = 35. Hasil uji coba menunjukkan bahwa hasil *clustering* mengalami penurunana seiring bertambahnya jumlah *cluster*, dan beberapa kenaikan seperti pada k = 12 dengan lemmatization, namun kenaikan tersebut tidak cukup signifikan sehingga tidak dapat menentukan jumlah *cluster* dengan metode *clustering* CompleteLink Agglomerative. Secara keseluruhan hasil *clustering* tanpa menggunakan lemmatization proses sama unggulnya dengan menggunakan lemmatization dalam tahap preposesing teks.

E. AverageLink Agglomerative Clustering

Metode average linkage *clustering* menghitung jarak antara dua *cluster* yang disebut sebagai jarak rata-rata dimana jarak tersebut dihitung pada masing-masing *cluster*. Metode ini relative yang terbaik dari metode-metode hierarkikal. Namun membutuhkan waktu komputasi yang paling tinggi dibandingkan dengan metode hierarkikal yang lain.



Gambar 4.5. Hasil *clustering* artikel ilmiah dengan AverageLink

Gambar 4.5 merupakan perbandingan hasil *clustering* artikel ilmiah dengan menggunakan lemmatization proses dan tanpa menggunakan lemmatization proses di tahap preposesing teks. Pada skenario percobaan ini, penulis menggunakan algoritma AverageLink Agglomerative Clustering pada jumlah $k = 2$ hingga $k = 35$. Hasil uji coba menunjukkan bahwa hasil *clustering* pada $k = 2$ sampai $k = 5$ mengalami penurunana seiring bertambahnya jumlah *cluster*, namun hasil *cluster* menunjukkan kenaikan yang signifikan pada jumlah $k = 6$ dengan nilai silhouette sebesar 0,02736. Secara keseluruhan hasil *clustering* dengan menggunakan lemmatization proses lebih unggul daripada dengan menggunakan lemmatization pada tahap preposesing teks.

4.2.3 Ekstraksi Fitur Produktivitas Berdasarkan Artikel Ilmiah

Setiap *author* memiliki beberapa fitur yang akan menggambarkan kinerjanya secara individual. Di dalam proses ekstraksi fitur produktifitas, dataset yang berisi informasi artikel setiap *author* beserta tahun publikasinya dan sitasi yang diterima masing-masing artikel per tahun akan diproses bersama dengan topik artikel ilmiah yang telah didapatkan pada proses sebelumnya. Dari proses ini, akan dihasilkan 6 macam fitur produktifitas *author*. Gambar 4.6 menunjukkan cuplikan dari hasil fitur produktivitas pada masing-masing artikel.

id_author	tahun_pengujian	T1f1	T1f2	T1f3	T1f4	T1f5	T1f6
39	2000	0	0	0	0	0	0
39	2001	0	0	0	0	0	0
39	2002	0	0	0	0	0	0
39	2003	0	0	0	0	0	0
39	2004	1	1	1	0	0	0
39	2005	5	6	5.5	0	0	0
39	2006	5	11	7.833	2	2	2
39	2007	2	13	6.417	0	2	1
39	2008	0	13	4.117	0	2	0.667
39	2009	0	13	3.084	1	3	1.5
57	2000	0	0	0	0	0	0
57	2001	0	0	0	0	0	0
57	2002	0	0	0	0	0	0
57	2003	0	0	0	0	0	0
57	2004	0	0	0	0	0	0
57	2005	0	0	0	0	0	0
57	2006	0	0	0	0	0	0
57	2007	0	0	0	0	0	0
57	2008	0	0	0	0	0	0
57	2009	2	2	2	0	0	0

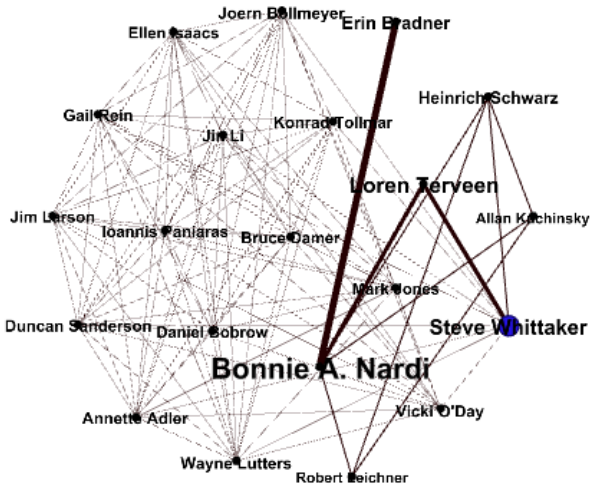
Gambar 4.6. Hasil fitur produktivitas berdasarkan topik

Dari hasil fitur produktifitas berdasarkan topik di atas dapat dilihat bahwa tidak semua *author* memiliki fitur produktifitas (fitur 1, fitur 2, fitur 3, fitur 4, fitur 5, dan fitur 6) pada topik 1. Hanya beberapa *author* seperti *author* dengan *id* 39 dan 57 yang memiliki fitur produktifitas dalam topik 1. Hal tersebut dikarenakan oleh tidak semua *author* dalam dataset memiliki artikel ilmiah dalam bidang keahlian di topik 1, yaitu topik *System Performance*.

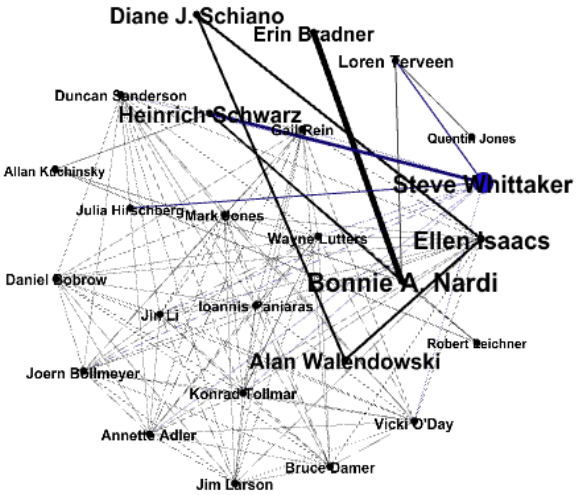
4.2.4 Ekstraksi Fitur Kolaborasi Berdasarkan Artikel Ilmiah

Selain fitur produktifitas, sangatlah penting dalam proses pengelompokan potensi penulis berdasarkan bidang penelitian untuk mengamati kemampuan kolaborasi *author* dengan *author* yang lain. Ekstraksi fitur kolaborasi *author* diproses berdasarkan topik penelitian *author* yang telah dididapat dari tahap sebelumnya, yaitu ekstraksi topik artikel ilmiah.

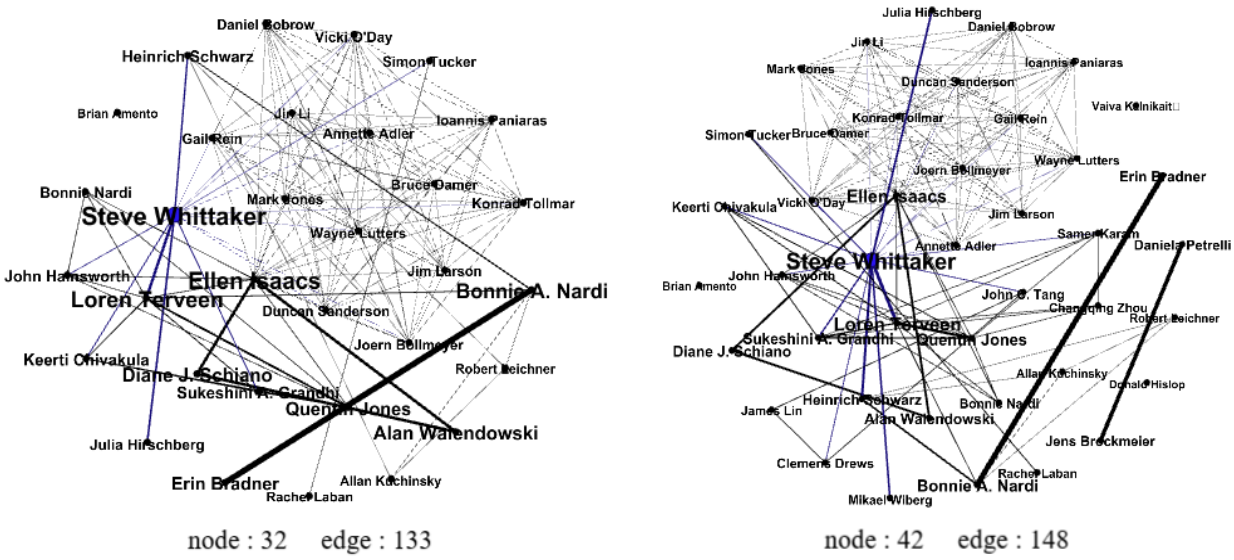
Dalam penelitian ini menggunakan dua macam pembobotan graph untuk memodelkan kemampuan kolaborasi *author,* yaitu pembobotan edge berdasarkan quantity atau frekuensi publikasi artikel ilmiah *author* dan pembobotan edge berdasarkan impact atau sitasi yang diterima oleh setiap artikel.



node : 21 edge : 109



node : 25 edge : 114



Gambar 4.7. Hasil graph pada tahun 2000, 2003, 2006, dan 2009

Gambar 4.7 menujukkan hasil graph yang terbentuk pada tahun 2000, 2003, 2006, dan 2009 guna untuk mengekstraksi fitur kolaborasi yang menyoroti hubungan yang dibentuk oleh dua *author* terkemuka, yaitu Bonnie A. Nardi dan Steve Whittaker, dapat diamati bahwa semakin bertambahnya tahun penelitian, maka semakin padat dan banyak relasi atau *co-authorship* yang terbentuk antar *author* oleh kedua *author* tersebut.

Dari masing-masing graph yaitu *quantitative graph* dan *impact graph* digunakan untuk proses ekstraksi fitur kolaborasi berdasarkan topik pada artikel ilmiah, yaitu fitur 7, fitur 8, fitur 9, fitur 10, fitur 11, fitur 12, dan fitur 13. Berikut pada Gambar 4.8 menunjukkan cuplikan dari hasil fitur kolaborasi berdasarkan topik artikel ilmiah dari metode yang diusulkan pada penelitian ini.

Dari hasil fitur kolaborasi berdasarkan topik pada Gambar 4.8 dapat dilihat bahwa tidak semua *author* memiliki fitur kolaborasi (fitur 7, fitur 8, fitur 9, fitur 10, fitur 11, fitur 12, dan fitur 13) pada topik 1. Hanya beberapa *author* seperti *author* dengan *id* 39 yang memiliki fitur kolaborasi dalam topik 1. Hal tersebut dikarenakan tidak semua *author* dalam dataset memiliki artikel ilmiah dalam bidang keahlian di topik 1, yaitu topik *System Performance.*

id_author	tahun_pengujian	T1f7	T1f8	T1f9	T1f10	T1f11	T1f12	T1f13
39	2000	0	0	0	0	0	0	0
39	2001	0	0	0	0	0	0	0
39	2002	0	0	0	0	0	0	0
39	2003	0	0	0	0	0	0	0
39	2004	2	2	0.092	4	0	0	0
39	2005	8	13	0.473	104	0	0	0
39	2006	18	20.667	0.807	372.006	1	0.075	4
39	2007	18	14.5	0.72	261	0.352	0.027	1.408
39	2008	18	9.567	0.479	172.206	0.2	0.017	0.8
39	2009	18	7.233	0.258	130.194	0.207	0.017	0.828
57	2000	0	0	0	0	0	0	0
57	2001	0	0	0	0	0	0	0
57	2002	0	0	0	0	0	0	0
57	2003	0	0	0	0	0	0	0
57	2004	0	0	0	0	0	0	0
57	2005	0	0	0	0	0	0	0
57	2006	0	0	0	0	0	0	0
57	2007	0	0	0	0	0	0	0
57	2008	0	0	0	0	0	0	0
57	2009	12	16	0.571	192	0	0	0

Gambar 4.8. Hasil fitur kolaborasi berdasarkan topik

4.2.5 Ekstraksi Fitur Dinamika *Author* Berdasarkan Artikel Ilmiah

Untuk menangkap kedinamisan dalam kinerja penelitian *author*, penelitian ini mendefinisikan beberapa indikator perubahan yang akan diterapkan pada masing-masing fitur yang telah dijelaskan sebelumnya. Terdapat lima macam fitur dinamika *author*, diantaranya adalah perubahan minimum, perumahan maksimum, perubahan terakhir, total perubahan, dan nilai representative. Berikut pada Gambar 4.9 menunjukkan cuplikan dari hasil fitur dinamika *author* berdasarkan topik penelitian artikel ilmiah.

id	id_author	T1f1_min	T1f1_max	T1f1_end	T1f1_sum	T1f1_rep	T1f2_min	T1f2_max	T1f2_end	T1f2_sum	T1f2_rep
1	1001469	0	0	0	0	0	0	0	0	0	0
2	100187	-1	1	0	0	0.1	0	1	0	1	0.1
3	100224	0	0	0	0	0	0	0	0	0	0
4	1002526	-1	5	0	4	2.6	0	5	5	25	2.8
5	1002587	0	0	0	0	0	0	0	0	0	0
6	100261	-1	1	0	0	0.1	0	1	0	1	0.2
7	100296	0	0	0	0	0	0	0	0	0	0
8	100445	-1	1	0	0	0.1	0	1	0	1	0.1
9	100513	-3	2	1	-3	3	1	5	3	24	5.9
10	1006676	0	0	0	0	0	0	0	0	0	0
11	1007004	0	0	0	0	0	0	0	0	0	0
12	1007123	-3	2	1	-1	0.8	0	3	1	6	0.9
13	100772	-3	6	-3	2	2	0	6	3	19	2.5
14	1008	-1	2	1	3	0.9	0	3	3	9	0.9
15	10098	-2	2	-2	0	0.2	0	2	0	2	0.2
16	1010000	0	0	0	0	0	0	0	0	0	0
17	101030	-5	6	-5	7	11.1	8	16	11	107	11.5

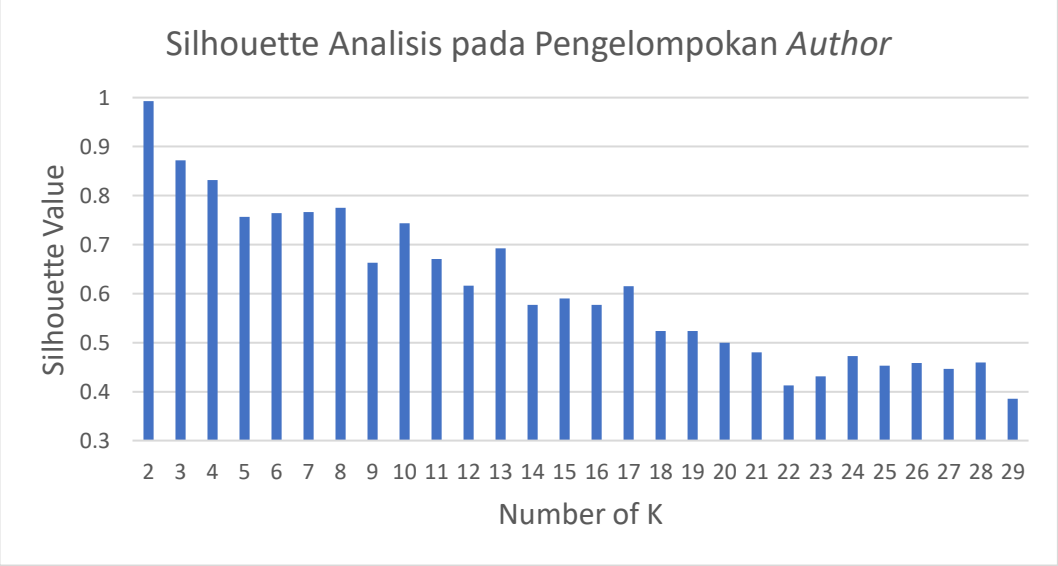
Gambar 4.9. Hasil fitur dinamika *author* berdasarkan topik

Semua fitur yang dihasilkan dari proses ekstraksi fitur produktifitas berdasarkan topik dan ekstraksi fitur kolaborasi berdasarkan topik menghasilkan 10 nilai yang didapatkan disetiap tahun pengujian. Pada penelitian ini menggunakan tahun pengujian 2000 hingga 2009, sehingga masing-masing *author* memiliki 10 baris fitur produktifitas dan kolaborasi berdasarkan topik.

Dari hasil fitur dinamika *author* berdasarkan topik di atas dapat dilihat bahwa tidak semua *author* memiliki fitur dinamika *author* pada topik 1. Hanya beberapa *author* seperti *author* dengan *id* 100513, 1002526, 1007123 dan seterusnya. Hal tersebut dikarenakan oleh tidak semua *author* dalam dataset memiliki artikel ilmiah dalam bidang keahlian di topik 1, yaitu topik *System Performance*.

4.2.6 Pengelompokan Potensi *Author* Berdasarkan Topik Artikel Ilmiah

Setelah didapatkan fitur-fitur dari proses ekstraksi topik pada artikel ilmiah, ekstraksi fitur produktifitas, ekstraksi fitur kolaborasi, dan ekstraksi fitur dinamika *author*, semua fitur tersebut digunakan untuk mengelompokkan potensi penulis dengan cara unsupervised yaitu metode *clustering*. Algoritma *clustering* yang digunakan adalah K-Means++ dengan jumlah *cluster* yang berbeda. Gambar 4.10 menunjukkan hasil *clustering author* pada k = 2 hingga k = 29. Dari hasil uji coba menunjukkan nilai silhouette terbaik ada pada *cluster* sejumlah 10.



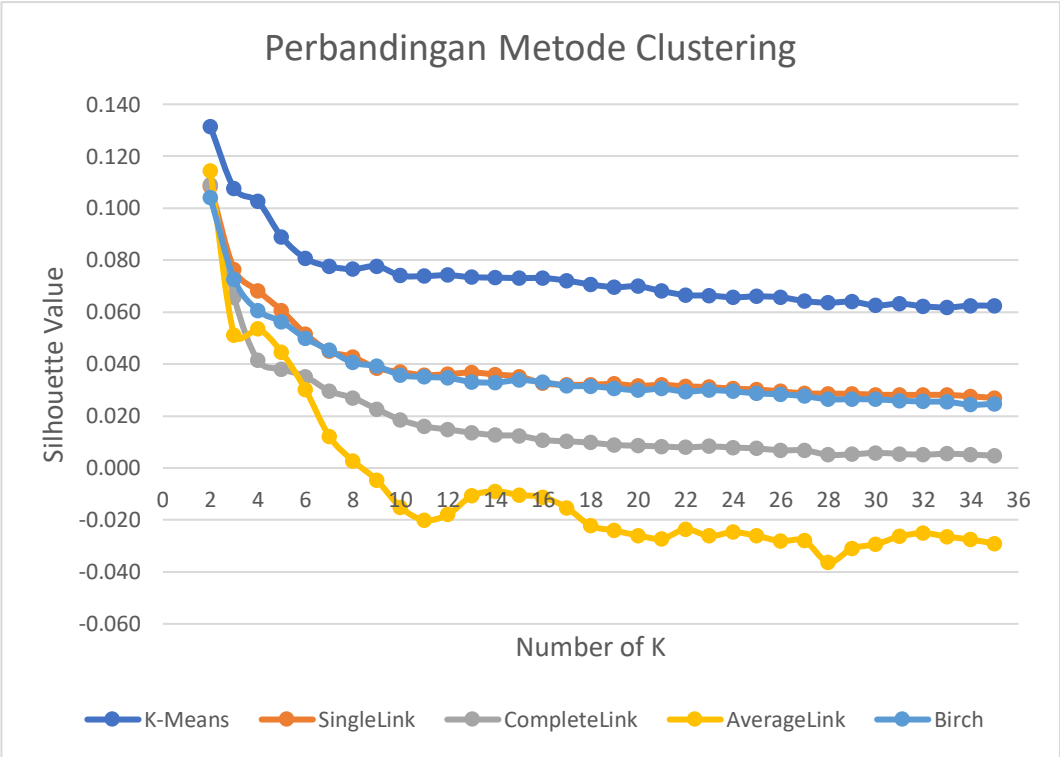
Gambar 4.10. Hasil eksperimen *clustering author*

4.3 Hasil Pengujian dan Analisis

Setelah memperoleh semua hasil matriks fitur pada setiap *author*, yaitu fitur produktifitas berdasarkan topik, fitur kolaborasi berdasarkan topik, dan fitur dinamika *author* berdasarkan topik. Fitur-fitur tersebut akan diolah dengan metode clustering K-Means ++ untuk mengelompokkan *author* berdasarkan potensi dan bidang keahliannya melalui artikel ilmiah yang telah dipublikasi.

4.3.1 Hasil Pengujian pada Ekstraksi Topik

Dari sepuluh percobaan yang telah dipaparkan pada bagian 4.2.2, mengenai clustering artikel ilmiah dengan beberapa metode dan percobaan terhadap tahap lemmatization pada preprosesing, dapat disimpulkan bahwa proses *clustering* tanpa melakukan tahap lemmatization menghasilkan nilai silhouette yang relative unggul dibandingkan hasil *clustering* yang menerapkan tahap lemmatization pada semua metode *clustering* yang diuji, yaitu metode *clustering* k-means++ *clustering*, birch *clustering*, averagelink agglomerative *clustering*, completelink agglomerative *clustering*, singlelink agglomerative *clustering*.



Gambar 4.11. Perbandingan hasil *clustering* tanpa lemmatization

Pada Gambar 4.11 menujukkan perbandingan pada kelima metode *clustering* yang telah dilakukan uji coba tanpa menggunakan lemmatization pada tahap preprosesing teks. Terlihat bahwa metode K-Means++ paling unggul dalam nilai silhouettenya dengan jumlah *cluster* optimal $k = 9$ dengan nilai *silhouette* sebesar 0,07766. Oleh karena itu penelitian ini memutuskan menggunakan jumlah *cluster* sebanyak sembilan sebagai kelompok topik artikel ilmiah.

Setelah didapatkan daftar artikel pada masing-masing topik, kami membuat 2 tabel yang berbeda pada setiap topiknya. Tabel pertama yaitu berisi tentang informasi data artikel ilmiah, tahun publikasi, *author* dan co-*authorship* pada masing-masing topik. Sedangkan pada tabel kedua berisi tentang informasi data sitasi yang diperoleh oleh masing-masing arikel ilmiah disetiap topik. Berikut pada Tabel 4.10 menjelaskan secara detail tentang jumlah data *author* dan jumlah data sitasi pada masing-masing topik yang digunakan dalam penelitian.

Tabel 4.10. Data uji sitasi berdasarkan tahun sitasi

Topik	**Jumlah Artikel**	**Jumlah Data *Author***	**Jumlah Data Sitasi**
Topik 1	31.003	106.551	219.250
Topik 2	30.557	86.776	322.245
Topik 3	32.472	96.378	181.384
Topik 4	27.675	72.579	193.152
Topik 5	31.214	101.215	198.686
Topik 6	35.455	121.777	207.567
Topik 7	33.257	105.847	98.898
Topik 8	34.972	114.747	201.779
Topik 9	39.736	134.865	306.450
Total	**296.341**	**940.735**	**1.929.411**

4.3.2 Hasil Pengujian pada Pengelompokan Potensi *Author*

Hasil uji coba pada bagian 4.2.6 menunjukkan bahwa semua *author* pada dataset dikelompokkan dengan metode clustering K-Means++ menjadi 10 kelompok berdasarkan potensi dan bidang penelitian.

Tabel 4.11. Hasil Pengujian pada Pengelompokan Potensi Author

Kelompok *Author*	**Jumlah *Author***	**Jumlah *Author* Menurut Range Publikasi**
Cluster 1	3078	521 *author* = 25-50, 519 *author* = 51-75, 516 *author* = 76-100, 495 *author* = 101-125, 486 *author* = 126-150, 276 *author* = 151-175, 265 *author* > 175
Cluster 2	2	265 *author* > 175
Cluster 3	56	1 *author* = 126-150, 1 *author* = 151-175, 54 *author* > 175
Cluster 4	8	8 *author* > 175
Cluster 5	39	4 *author* = 101-125, 8 *author* = 126-150, 4 *author* = 151-175, 23 *author* > 175
Cluster 6	39	2 *author* = 101-125, 1 *author* = 126-150, 6 *author* = 151-175, 30 *author* > 175
Cluster 7	13	13 *author* > 175
Cluster 8	5	5 *author* > 175
Cluster 9	18	1 *author* = 126-150, 4 *author* = 151-175, 13 *author* > 175
Cluster 10	207	2 *author* = 51-75, 9 *author* = 76-100, 22 *author* = 101-125, 25 *author* = 126-150, 39 *author* = 151-175, 110 *author* > 175

Pada proses sebelumnya, penelitian ini memilih *author* berdasarkan jumlah publikasi artikel ilmiah yang dibagi menjadi 7 range, yaitu *author* yang memiliki publikasi 25-50 artikel, 51-75 artikel, 76-100 artikel, 101-125 artikel, 126-150 artikel, 151-175 artikel, dan *author* yang memiliki publikasi lebih dari 175 artikel disemua tahun. Pada kesempatan kali ini penelitian ini akan membandingkan setiap *author* yang berada pada setiap *cluster* dengan jumlah publikasi artikel ilmiah yang dimiliki oleh masing-masing *author*.

Tabel 4.11 menjelaskan hasil pengujian pada pengelompokan potensi author, beserta jumlah *author* per *cluster* dan jumlah *author* pada masing-masing *range* jumlah publikasi artikel ilmiah secara detail. Dapat disimpulkan bahwa kelompok potensi *author* tidak dapat dilihat hanya dari segi jumlah publikasinya saja. Hal tersebut dapat dilihat dari hasil uji coba pada pengelompokan potensi *author* berdasarkan topik artikel ilmiah, bahwa kelompok *author* pada *cluster* 1 memiliki 7 *range* jumlah publikasi yang berbeda antar anggota *author*nya.

Hal tersebut disebabkan oleh beberapa faktor, diantaranya *author* yang memiliki jumlah publikasi artikel ilmiah yang sedikit memiliki satu bidang penelitian saja, sedangan *author* yang memiliki jumlah publikasi artikel ilmiah yang tinggi memiliki lebih dari satu macam topik penelitian, sehingga mereka terkelompokkan menjadi satu *cluster* karena memiliki jumlah publikasi dan sitasi yang seimbang pada topik penelitan yang sama.

Adapun analisa yang dilakukan pada masing-masing kelompok *author* dalam setiap topik artikel ilmiah. Pada Topik pertama, pada jumlah publikasi tidak terjadi peningkatan. Sedangkan pada jumlah sitasi terjadi peningkatan sejak tahun 2014 hingga tahun 2016 pada data *cluster* 1, *cluster* 2, dan *cluster* 4. Topik kedua, pada jumlah publikasi terjadi peningkatan sejak tahun 2013 hingga tahun 2016 pada *cluster* 1, *cluster* 2, dan *cluster* 4. Sedangkan pada jumlah sitasi tidak terjadi peningkatan bahkan terjadi penurunan. Topik ketiga, pada jumlah publikasi terjadi peningkatan sejak tahun 2012 hingga tahun 2016 pada *cluster* 1, *cluster* 2, dan *cluster* 4 meskipun dari tahun 2015 *cluster* 1 dan *cluster* 2 mengalami penurunan. Sedangkan pada jumlah sitasi terjadi peningkatan sejak tahun 2014 hingga tahun 2016 pada *cluster* 3 dan *cluster* 4.

Topik keempat, pada jumlah publikasi terjadi peningkatan sejak tahun 2013 pada *cluster* 2. Sedangkan pada jumlah sitasi tidak terjadi peningkatan bahkan terjadi penurunan. Topik kelima, pada jumlah publikasi tidak terjadi peningkatan, begitu pula pada jumlah sitasi. Topik keenam, pada jumlah publikasi terjadi peningkatan sejak tahun 2015 *cluster* 2 dan *cluster* 4. Sedangkan pada jumlah sitasi terjadi peningkatan sejak tahun 2015 hingga tahun 2016 pada *cluster* 1, *cluster* 2, dan *cluster* 4. Pada topik ketujuh, pada jumlah publikasi tidak terjadi peningkatan, sedangkan pada jumlah sitasi terjadi peningkatan sejak tahun 2015 hingga tahun 2016 pada *cluster* 1, *cluster* 2, dan *cluster* 4.

Topik kedelapan, pada jumlah publikasi terjadi peningkatan sejak tahun 2014 pada *cluster* 1, *cluster* 2, dan *cluster* 4. Sedangkan pada jumlah sitasi terjadi peningkatan sejak tahun 2014 pada semua *cluster* kecuali kelas *cluster* 10. Pada topik kesembilan, jumlah publikasi terjadi peningkatan sejak tahun 2014 pada *cluster* 2, *cluster* 4, dan *cluster* 6. Sedangkan pada tahun sitasi terjadi peningkatan sejak tahun 2014 pada *cluster* 5, *cluster* 4, dan *cluster* 3. Namun beberapa kelas mengalami penurunan pada tahun 2015 seperti pada *cluster* 1 dan *cluster* 2, tetapi *cluster* 8 dan *cluster* 9 mengalami peningkatan pada tahun 2016.

4.3.3. Analisa Hasil Uji Coba

Pada analisa hasil uji coba akan merangkum kinerja dari metode yang diusulkan pada penelitian ini, dan juga akan menganalisis kelebihan dan kekurangan dari setiap proses yang dilakukan sehingga menghasilkan pengelompokan potensi penulis berdasarkan topik pada artikel ilmiah.

Pada Tabel 4.10 menjelaskan tentang hasil analisis 3 potensi teratas dari setiap kelompok *author* pada masing-masing topik artikel ilmiah. Dari hasil uji coba menjelaskan bahwa kelompok *author* pada *cluster* 1 sangat berpotensi di topik 9 yaitu bidang penelitian *Data Manipulation*, kemudian diikuti oleh bidang penelitian *Information System* dan *System Performance*. Selanjutnya pada *cluster* 2 sangat berpotensi pada bidang penelitian *Information System*, kemudian diikuti oleh bidang penelitian *Data Manipulation* dan *System Optimization*.

Tabel 4.12. Analisis potensi kelompok *author* pada masing-masing topik

Cluster	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**
C1	3	-	2	-	-	-	-	-	1
C2	-	-	1	-	-	-	3	-	2
C3	-	-	-	-	-	-	1	2	3
C4	-	-	3	-	-	-	-	2	1
C5	-	-	-	-	2	-	1	3	-
C6	-	-	-	3	-	-	-	2	1
C7	-	-	-	-	3	-	1	2	-
C8	-	-	-	-	-	-	3	1	2
C9	2	-	-	-	3	1	-	-	-
C10	-	-	-	-	-	-	3	1	2

Pada *cluster* 3, kelompok *author* tersebut sangat berpotensi pada bidang penelitian di topik 7 yaitu dibidang *System Optimization,* kemudian diikuti oleh bidang penelitian *Image Processing* dan *Data Manipulation*. Sedangkan pada *cluster* 4 sangat berpotensi pada bidang penelitian di topik 9, yaitu *Data Manipulation*, kemudian diikuti oleh bidang penelitian *Image Processing*, *Information System*. Selanjutnya pada *cluster* 5 sangat berpotensi pada bidang penelitian di topik 9, yaitu *Data Manipulation*, kemudian diikuti oleh bidang penelitian *Image Processing, Information System*.

Kemudian pada *cluster* 6 sangat berpotensi pada bidang penelitian *Information System*, kemudian diikuti oleh bidang penelitian *Data Manipulation* dan *System Optimization*. Sedangkan *cluster* 7 sangat berpotensi pada bidang penelitian *Information System,* kemudian diikuti oleh bidang penelitian *Data Manipulation* dan *System Optimization*. Begitu juga dengan *cluster* 8, kelompok *author* tersebut sangat berpotensi pada bidang penelitian di topik 9, yaitu *Data*

Manipulation, kemudian diikuti oleh bidang penelitian *Image Processing*, *Information System*. Dan yang terakhir adalah kelompok *author* pada *cluster* 10, kelompok *author* ini sangat berpotensi pada bidang *Image Processing*, kemudian diikuti oleh bidang penelitian *Data Manipulation* dan *System Optimization*.

Hasil analisis pada metodologi usulan secara keseluruhan adalah terdapat kekurangan pada metode *clustering* untuk mengelompokkan artikel ilmiah berdasarkan topik. Penggunaan algoritma K-Means++ memperoleh hasil yang paling unggul dibandingkan keempat metode lain yang telah disebutkan sebelumnya. Meskipun dengan menggunakan algoritma K-Means++ mampu memperoleh hasil silhouette yang paling unggul, namun hasil tersebut kurang efisien untuk penentuan *cluster* topik yang sesuai dengan topik yang terkandung pada artikel ilmiah. Hal tersebut dapat menyebabkan kerusakan pada serangkaian proses ekstraksi fitur di tahap berikutnya.Oleh karena itu, perlu adanya metode pembobotan kata pada artikel ilmiah yang memperhatikan unsur *semantic*, *sentatic*, dan *co-ocurrance* pada masing-masing term dalam dokumen agar lebih merepresentasikan fitur artikel ilmiah jika diterapkan metode *clustering* untuk proses pengelompokan artikel ilmiah berdasarkan topik.

BAB 5
KESIMPULAN DAN SARAN

Bab ini menjelaskan mengenai kesimpulan dan saran yang didapatkan dari tahapan implementasi dan pengujian.

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut :

1. Ekstraksi fitur berdasarkan topik pada artikel ilmiah secara baik dapat dilakukan dengan menggunakan metode pembobotan Word2Vec dan algoritma clustering K-Means++, dengan nilai silhouette rata-rata 0,07766.
2. Dari hasil uji coba menyatakan bahwa pengelompokan potensi penulis dengan mempertimbangkan kombinasi dari aspek topik, produktifitas, kolaborasi dan dinamika penulis dapat dilakukan dengan baik memisahkan potensi *author* dari tahun ke tahun menurut bidang keahliannya.
3. Seperti yang ditunjukan pada skenario pengujian ekstraksi topik pada artikel ilmiah dan pengelompokan *author*, performa hasil terbaik ditunjukan oleh metode K-means++ yang mampu melakukan cluster dokumen berupa artikel ilmiah dan proses clustering author berdasarkan potensi penulis dan topik penelitian dengan rerata nilai silhouette sebesar 0,07433.

5.2 Saran

1. Perlu dilakukan penelitian terhadap *clustering* dokumen yang dapat membedakan topik artikel ilmiah lebih baik daripada algoritma K-Means++.

DAFTAR PUSTAKA

[1] N. M. Glazunov, *Foundations of scientific research (Foundations of Research Activities)*. National Aviation University, 2012.

[2] Brody, S. & Elhadad, N., 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. pp. 804–812

[3] G. Basten, *Introduction to Scientific Research Projects*. Ventus Publishing ApS, ISBN 978-87-7681-674-2, 2011.

[4] G. Tsatsaronis, I. Varlamis, S. Torge, and M. Reimann, "How to Become a Group Leader ? or Modeling Author Types based on Graph Mining," in *International Conference on Theory and Practice of Digital Libraries*, 2011, pp. 15–26.

[5] V. Durieux and P. A. Gevenois, "Bibliometric Indicators : Quality Measurements of," *Bibliometr. Indic. Qual. Meas. Sci. Publ.*, vol. 255, no. 2, pp. 342–351, 2010.

[6] X. Li, C. S. Foo, K. L. Tew, and S. Ng, "Searching for Rising Stars in Bibliography Networks," in *International Conference on Database Systems for Advanced Applications*, 2009, pp. 288–292.

[7] G. Panagopoulos, G. Tsatsaronis, and I. Varlamis, "Detecting rising stars in dynamic collaborative networks," *J. Informetr.*, vol. 11, no. 1, pp. 198–222, 2017.

[8] A. Daud, R. Abbasi, and F. Muhammad, "Finding Rising Stars in Social Networks," in *International Conference on Database Systems for Advanced Applications*, 2013, pp. 13–14.

[9] G. Tsatsaronis, B. V Elsevier, M. Reimann, G. Tsatsaronis, and M. Reimann, "Efficient community detection using power graph analysis," in *Proceedings of the 9th workshop on Large-scale and distributed informational retrieval*, 2011, no. October.

[10] N. T. Hoang, P. Do, and H. N. Le, "A Fast Algorithm for Predicting Topics of Scientific Papers Based on Co-authorship Graph Model," *Adv. Methods*

Comput. Collect. Intell., pp. 83–91, 2013.

[11] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, vol. 85.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet Web Inf. Syst.*, vol. 54, no. 1999–66, pp. 1–17, 1998.

[13] D. Arthur and S. Vassilvitskii, "K-Means++: the Advantages of Careful Seeding," *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, vol. 8, pp. 1027–1025, 2007.

[14] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proc. 1996 ACM SIGMOD Int. Conf. Manag. Data*, vol. 1, pp. 103–114, 1996.

[15] J. De Knijff, F. Frasincar, and F. Hogenboom, "Data & Knowledge Engineering Domain taxonomy learning from text : The subsumption method versus hierarchical clustering," *Data Knowl. Eng.*, vol. 83, no. 0, pp. 54–69, 2013.

[16] S. K. Popat and M. Emmanuel, "Review and Comparative Study of Clustering Techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 805–812, 2014.

[17] Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, pp.2493–2537.

[18] C. Aalla and V. Pudi, "Mining Research Problems from Scientific Literature," *2016 IEEE Int. Conf. Data Sci. Adv. Anal.*, pp. 351–360, 2016.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *World Wide Web Internet Web Inf. Syst.*, vol. 54, no. 1999–66, pp. 1–17, 1998.

[20] N. Y. Saiyad, H. B. Prajapati, and V. K. Dabhi, "A Survey of Document Clustering using Semantic Approach," *Int. Conf. Electr. Electron. Optim. Tech.*, vol. 6, no. 4, pp. 2555–2562, 2016.

[21] L. Xiong, "Survey on text clustering algorithm," *2011 IEEE 2nd Int. Conf. Softw. Eng. Serv. Sci.*, no. 4, pp. 901–904, 2011.

[22] F. De Morsier, D. Tuia, M. Borgeaud, V. Gass, and J. P. Thiran, “Cluster validity measure and merging system for hierarchical clustering considering outliers,” *Pattern Recognit.*, vol. 48, no. 4, pp. 1474–1485, 2015.

[23] C. Jin and Q. Bai, “Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co- occurrence,” *2016 Int. Conf. Inf. Syst. Artif. Intell.*, 2016.

[24] S. K. Popat and M. Emmanuel, “Review and Comparative Study of Clustering Techniques,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 805–812, 2014.

[25] Zhai, Z. et al., 2011. Clustering Product Features for Reviewon Mining. In *Proceeding WSDM ’11 Proceedings of the fourth ACM international conference*. pp. 347–354.

LAMPIRAN

LAMPIRAN 1. Hasil *clustering* artikel ilmiah dengan K-Means++

1. Preposesing tanpa menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1300	0.1306	0.1324	0.1295	0.1309	0.1321	0.1336	0.13130
3	0.1081	0.1069	0.1077	0.1071	0.1075	0.1082	0.1080	0.10764
4	0.1026	0.1014	0.1019	0.1025	0.1032	0.1029	0.1035	0.10257
5	0.0885	0.0873	0.0884	0.0909	0.0913	0.0881	0.0876	0.08887
6	0.0807	0.0804	0.0810	0.0824	0.0811	0.0789	0.0803	0.08069
7	0.0786	0.0779	0.0761	0.0765	0.0788	0.0768	0.0784	0.07759
8	0.0760	0.0758	0.0762	0.0765	0.0775	0.0777	0.0760	0.07653
9	0.0774	0.0767	0.0774	0.0778	0.0784	0.0788	0.0771	0.07766
10	0.0741	0.0706	0.0753	0.0756	0.0772	0.0743	0.0712	0.07404
11	0.0719	0.0727	0.0738	0.0748	0.0759	0.0739	0.0740	0.07386
12	0.0726	0.0737	0.0744	0.0751	0.0751	0.0732	0.0754	0.07421
13	0.0729	0.0725	0.0714	0.0741	0.0750	0.0739	0.0742	0.07343
14	0.0753	0.0754	0.0703	0.0719	0.0733	0.0735	0.0733	0.07329
15	0.0705	0.0742	0.0730	0.0738	0.0719	0.0755	0.0727	0.07309
16	0.0724	0.0709	0.0717	0.0753	0.0747	0.0730	0.0737	0.07310
17	0.0732	0.0710	0.0672	0.0704	0.0761	0.0705	0.0766	0.07214
18	0.0686	0.0708	0.0679	0.0704	0.0708	0.0721	0.0736	0.07060
19	0.0710	0.0716	0.0664	0.0711	0.0719	0.0682	0.0671	0.06961
20	0.0712	0.0680	0.0708	0.0683	0.0710	0.0691	0.0710	0.06991
21	0.0698	0.0658	0.0688	0.0671	0.0683	0.0690	0.0677	0.06807
22	0.0640	0.0636	0.0669	0.0691	0.0612	0.0691	0.0719	0.06654
23	0.0658	0.0645	0.0687	0.0650	0.0642	0.0666	0.0697	0.06636
24	0.0624	0.0658	0.0671	0.0641	0.0659	0.0702	0.0643	0.06569

25	0.0677	0.0665	0.0638	0.0667	0.0662	0.0658	0.0657	0.06606
26	0.0670	0.0630	0.0662	0.0615	0.0676	0.0651	0.0696	0.06571
27	0.0627	0.0704	0.0639	0.0621	0.0625	0.0644	0.0637	0.06424
28	0.0635	0.0610	0.0624	0.0655	0.0642	0.0649	0.0637	0.06360
29	0.0650	0.0644	0.0661	0.0611	0.0609	0.0653	0.0653	0.06401
30	0.0638	0.0612	0.0626	0.0623	0.0606	0.0645	0.0626	0.06251
31	0.0638	0.0625	0.0632	0.0647	0.0608	0.0644	0.0634	0.06326
32	0.0622	0.0608	0.0639	0.0613	0.0610	0.0626	0.0638	0.06223
33	0.0615	0.0622	0.0612	0.0589	0.0626	0.0643	0.0618	0.06179
34	0.0624	0.0642	0.0639	0.0633	0.0596	0.0635	0.0606	0.06250
35	0.0623	0.0622	0.0612	0.0620	0.0621	0.0642	0.0630	0.06243

2. Preposesing dengan menggunakan tahap lemmatization

Preprocessing Dengan Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1291	0.1298	0.1296	0.1291	0.1288	0.1281	0.1283	0.12897
3	0.1087	0.1087	0.1084	0.1085	0.1082	0.1086	0.1080	0.10844
4	0.1021	0.1001	0.1009	0.1024	0.1017	0.1011	0.1015	0.10140
5	0.0879	0.0870	0.0870	0.0854	0.0913	0.0862	0.0872	0.08743
6	0.0790	0.0797	0.0803	0.0798	0.0805	0.0831	0.0788	0.08017
7	0.0771	0.0776	0.0764	0.0772	0.0761	0.0774	0.0760	0.07683
8	0.0775	0.0779	0.0758	0.0753	0.0754	0.0765	0.0744	0.07611
9	0.0717	0.0773	0.0766	0.0770	0.0748	0.0760	0.0764	0.07569
10	0.0748	0.0769	0.0778	0.0762	0.0739	0.0757	0.0781	0.07620
11	0.0752	0.0734	0.0746	0.0754	0.0730	0.0724	0.0736	0.07394
12	0.0773	0.0737	0.0752	0.0714	0.0728	0.0740	0.0737	0.07401
13	0.0738	0.0746	0.0760	0.0726	0.0734	0.0732	0.0715	0.07359
14	0.0755	0.0726	0.0712	0.0767	0.0728	0.0740	0.0723	0.07359
15	0.0725	0.0715	0.0757	0.0706	0.0726	0.0750	0.0706	0.07264
16	0.0746	0.0721	0.0752	0.0733	0.0715	0.0754	0.0731	0.07360

17	0.0710	0.0693	0.0722	0.0723	0.0690	0.0728	0.0712	0.07111
18	0.0684	0.0688	0.0745	0.0702	0.0666	0.0673	0.0679	0.06910
19	0.0703	0.0720	0.0689	0.0712	0.0691	0.0679	0.0677	0.06959
20	0.0701	0.0668	0.0692	0.0697	0.0638	0.0659	0.0661	0.06737
21	0.0725	0.0660	0.0675	0.0682	0.0666	0.0671	0.0654	0.06761
22	0.0656	0.0666	0.0658	0.0642	0.0659	0.0655	0.0638	0.06534
23	0.0648	0.0688	0.0684	0.0671	0.0686	0.0694	0.0621	0.06703
24	0.0637	0.0631	0.0671	0.0641	0.0682	0.0633	0.0639	0.06477
25	0.0682	0.0675	0.0680	0.0595	0.0655	0.0654	0.0660	0.06573
26	0.0645	0.0674	0.0608	0.0617	0.0656	0.0638	0.0608	0.06351
27	0.0613	0.0650	0.0618	0.0634	0.0662	0.0629	0.0638	0.06349
28	0.0611	0.0627	0.0630	0.0629	0.0656	0.0622	0.0574	0.06213
29	0.0612	0.0657	0.0602	0.0603	0.0624	0.0609	0.0633	0.06200
30	0.0600	0.0630	0.0640	0.0636	0.0618	0.0633	0.0641	0.06283
31	0.0602	0.0645	0.0599	0.0596	0.0620	0.0635	0.0598	0.06136
32	0.0592	0.0614	0.0620	0.0623	0.0625	0.0635	0.0611	0.06171
33	0.0600	0.0606	0.0596	0.0593	0.0603	0.0592	0.0612	0.06003
34	0.0596	0.0634	0.0613	0.0600	0.0617	0.0605	0.0618	0.06119
35	0.0603	0.0620	0.0619	0.0608	0.0603	0.0603	0.0607	0.06090

LAMPIRAN 2. Hasil *clustering* artikel dengan SingleLink Agglomerative

1. Preposesing tanpa menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1037	0.1108	0.1094	0.117	0.0968	0.1085	0.1126	0.10840
3	0.0737	0.0741	0.0827	0.0711	0.0779	0.0825	0.073	0.07643
4	0.0686	0.0654	0.0808	0.0623	0.0637	0.0675	0.0693	0.06823
5	0.055	0.0647	0.0642	0.0637	0.0638	0.0554	0.0569	0.06053
6	0.0453	0.0489	0.0511	0.0537	0.0579	0.0558	0.0485	0.05160

7	0.0453	0.0389	0.043	0.0506	0.0469	0.0488	0.0412	0.04496
8	0.0434	0.0373	0.042	0.0485	0.0434	0.0466	0.0375	0.04267
9	0.0402	0.0302	0.044	0.0464	0.0378	0.0386	0.0319	0.03844
10	0.0382	0.0328	0.0419	0.0374	0.0372	0.0394	0.0316	0.03693
11	0.0385	0.0359	0.0401	0.0306	0.0324	0.0407	0.032	0.03574
12	0.0405	0.0331	0.0393	0.0299	0.0337	0.0424	0.0339	0.03611
13	0.0421	0.0357	0.0416	0.0287	0.0345	0.0414	0.0333	0.03676
14	0.0388	0.0365	0.0427	0.0302	0.0318	0.0396	0.0323	0.03599
15	0.0343	0.0379	0.0422	0.0296	0.0329	0.0343	0.035	0.03517
16	0.0308	0.0388	0.0374	0.0264	0.0316	0.034	0.0291	0.03259
17	0.0329	0.0346	0.0348	0.0288	0.0292	0.0341	0.0299	0.03204
18	0.034	0.0343	0.0343	0.0298	0.031	0.0324	0.0288	0.03209
19	0.034	0.0357	0.0332	0.0306	0.0322	0.0323	0.0283	0.03233
20	0.0355	0.0322	0.0354	0.0259	0.0306	0.0321	0.0295	0.03160
21	0.0355	0.0327	0.0347	0.0276	0.0296	0.0322	0.0316	0.03199
22	0.035	0.0326	0.033	0.0264	0.0303	0.0332	0.0294	0.03141
23	0.0334	0.0339	0.0329	0.0273	0.0314	0.0306	0.0285	0.03114
24	0.0344	0.0337	0.0328	0.0292	0.0304	0.027	0.0261	0.03051
25	0.0329	0.0324	0.0337	0.0282	0.0295	0.0285	0.0263	0.03021
26	0.0328	0.033	0.0315	0.0262	0.0294	0.0281	0.0253	0.02947
27	0.0332	0.0317	0.0292	0.0266	0.0289	0.0271	0.0246	0.02876
28	0.0338	0.0307	0.0285	0.0264	0.0282	0.0276	0.0251	0.02861
29	0.0315	0.0305	0.0298	0.027	0.0263	0.0284	0.0262	0.02853
30	0.0319	0.0305	0.0287	0.0264	0.0267	0.028	0.0253	0.02821
31	0.0304	0.03	0.0288	0.026	0.027	0.029	0.0257	0.02813
32	0.03	0.031	0.0273	0.0266	0.0282	0.0278	0.026	0.02813
33	0.0285	0.0307	0.0267	0.0272	0.0287	0.0282	0.0268	0.02811
34	0.029	0.0273	0.026	0.0275	0.0288	0.0279	0.0266	0.02759
35	0.0287	0.0252	0.0257	0.0259	0.0296	0.027	0.0264	0.02693

2. Preposesing dengan menggunakan tahap lemmatization

Preprocessing Dengan Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.0974	0.1099	0.1082	0.0897	0.1039	0.0864	0.1054	0.10013
3	0.0694	0.072	0.0731	0.0706	0.0812	0.0749	0.0804	0.07451
4	0.0675	0.0628	0.0683	0.0642	0.0629	0.0709	0.072	0.06694
5	0.0518	0.0614	0.0592	0.054	0.0635	0.0536	0.0618	0.05790
6	0.0436	0.051	0.0474	0.0492	0.053	0.0493	0.0603	0.05054
7	0.04	0.0535	0.0423	0.0411	0.0457	0.0463	0.0452	0.04487
8	0.0377	0.0462	0.043	0.0417	0.0454	0.0441	0.0423	0.04291
9	0.0386	0.0354	0.0352	0.04	0.0399	0.0338	0.0402	0.03759
10	0.0394	0.0395	0.0337	0.0349	0.0308	0.0316	0.039	0.03556
11	0.0354	0.0363	0.0314	0.0368	0.0318	0.0325	0.0373	0.03450
12	0.0357	0.0393	0.0303	0.0385	0.0345	0.0362	0.033	0.03536
13	0.0326	0.0355	0.0306	0.0424	0.0358	0.033	0.0309	0.03440
14	0.0345	0.031	0.0317	0.0373	0.0384	0.0349	0.0315	0.03419
15	0.0347	0.0324	0.032	0.0365	0.0384	0.0354	0.0255	0.03356
16	0.0335	0.0292	0.0313	0.0345	0.0378	0.0328	0.0269	0.03229
17	0.0348	0.0294	0.0322	0.0308	0.0379	0.0317	0.0251	0.03170
18	0.0329	0.0252	0.034	0.0324	0.038	0.0314	0.0264	0.03147
19	0.0339	0.0221	0.0349	0.0326	0.0384	0.032	0.0266	0.03150
20	0.0326	0.024	0.0302	0.0339	0.0357	0.032	0.0263	0.03067
21	0.033	0.0246	0.0317	0.0324	0.0313	0.0288	0.0256	0.02963
22	0.0332	0.0227	0.0288	0.0318	0.0305	0.0277	0.0264	0.02873
23	0.0315	0.0238	0.0279	0.0312	0.0259	0.0265	0.0277	0.02779
24	0.0302	0.0249	0.0291	0.0299	0.0273	0.0255	0.0277	0.02780
25	0.0275	0.0266	0.0295	0.0292	0.0256	0.0248	0.0243	0.02679
26	0.029	0.0253	0.0292	0.0304	0.0247	0.0247	0.0244	0.02681
27	0.0285	0.026	0.0284	0.0303	0.0251	0.0258	0.0234	0.02679

28	0.0291	0.0247	0.0271	0.0305	0.0255	0.0257	0.0245	0.02673
29	0.0268	0.0248	0.0261	0.0312	0.0259	0.0262	0.0227	0.02624
30	0.0258	0.0257	0.0242	0.0308	0.0248	0.026	0.0214	0.02553
31	0.024	0.0262	0.0222	0.0287	0.0249	0.0263	0.0227	0.02500
32	0.0249	0.0267	0.022	0.028	0.0254	0.0274	0.0227	0.02530
33	0.0239	0.0275	0.0212	0.0246	0.0263	0.0282	0.0228	0.02493
34	0.0229	0.0266	0.0223	0.0255	0.0238	0.0276	0.0228	0.02450
35	0.0238	0.0258	0.0221	0.0244	0.0238	0.0273	0.0225	0.02424

LAMPIRAN 3. Hasil *clustering* artikel dengan CompleteLink Agglomerative

1. Preposesing tanpa menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1054	0.1114	0.1075	0.1062	0.1102	0.1115	0.1113	0.10907
3	0.0772	0.0655	0.0759	0.0543	0.0644	0.0554	0.0673	0.06571
4	0.0327	0.0575	0.0403	0.0217	0.0403	0.0359	0.0614	0.04140
5	0.0269	0.0463	0.0506	0.03	0.0319	0.0242	0.0564	0.03804
6	0.0245	0.0443	0.0396	0.0389	0.02	0.0337	0.0453	0.03519
7	0.014	0.0362	0.0296	0.0308	0.0199	0.0309	0.0461	0.02964
8	0.0072	0.0344	0.0305	0.0276	0.0139	0.0348	0.0404	0.02697
9	0.0079	0.0243	0.0211	0.0238	0.0125	0.03	0.039	0.02266
10	-0.0031	0.0232	0.0118	0.0213	0.0136	0.025	0.0379	0.01853
11	0.001	0.018	0.0103	0.0203	0.0106	0.0234	0.0284	0.01600
12	0.0126	0.016	0.0083	0.0096	0.0081	0.0214	0.0277	0.01481
13	0.011	0.013	0.0077	0.0094	0.0039	0.0232	0.0268	0.01357
14	0.0097	0.0111	0.0083	0.0095	0.0028	0.0208	0.0262	0.01263
15	0.0083	0.0091	0.0105	0.0109	0.0043	0.0205	0.0229	0.01236
16	0.011	0.0048	0.0107	0.0088	0.0028	0.0148	0.0224	0.01076
17	0.0085	0.0052	0.0096	0.0076	0.0033	0.0153	0.0224	0.01027

18	0.0075	0.0057	0.0053	0.0066	0.0085	0.0137	0.021	0.00976
19	0.0065	0.0054	0.0069	0.0054	0.0067	0.016	0.0152	0.00887
20	0.0065	0.0054	0.01	0.004	0.0069	0.0119	0.0153	0.00857
21	0.0075	0.0062	0.0097	0.0024	0.006	0.0114	0.0149	0.00830
22	0.0081	0.0058	0.008	0.0048	0.0043	0.0114	0.0131	0.00793
23	0.0094	0.0055	0.0095	0.004	0.0046	0.0112	0.014	0.00831
24	0.0096	0.0059	0.0088	0.0013	0.0042	0.0118	0.0135	0.00787
25	0.009	0.0044	0.0096	0.0012	0.0026	0.0133	0.0127	0.00754
26	0.0079	0.0042	0.0105	-0.0003	0.0037	0.0129	0.0087	0.00680
27	0.0064	0.0045	0.0113	0.0005	0.0035	0.0123	0.0091	0.00680
28	0.0067	0.0038	0.0088	-0.001	0.0001	0.0115	0.0056	0.00507
29	0.0066	0.0048	0.0084	-0.0011	0.002	0.0105	0.0055	0.00524
30	0.0076	0.005	0.0117	-0.0006	0.0021	0.0095	0.0054	0.00581
31	0.0063	0.0051	0.0117	-0.0016	0.0029	0.0088	0.0048	0.00543
32	0.0045	0.007	0.0088	-0.0014	0.002	0.0097	0.0052	0.00511
33	0.004	0.0062	0.0088	-0.0015	0.0014	0.0137	0.0057	0.00547
34	0.0042	0.0071	0.0072	-0.0035	0.0012	0.0136	0.0064	0.00517
35	0.0038	0.0069	0.0072	-0.0023	0	0.0119	0.0055	0.00471

2. Preposesing dengan menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1066	0.0978	0.0861	0.1072	0.1046	0.1056	0.104	0.10170
3	0.0472	0.0503	0.044	0.0699	0.0579	0.0495	0.0661	0.05499
4	0.0261	0.051	0.0326	0.0467	0.0409	0.0377	0.0596	0.04209
5	0.0314	0.0418	0.0225	0.0419	0.0276	0.0379	0.0326	0.03367
6	0.0285	0.0388	0.0182	0.0442	0.0215	0.0205	0.022	0.02767
7	0.0227	0.0348	0.0254	0.0396	0.0241	0.0205	0.0187	0.02654
8	0.019	0.0278	0.0198	0.0394	0.015	0.0248	0.0105	0.02233
9	0.0159	0.0251	0.0201	0.0329	0.0161	0.0184	0.011	0.01993

10	0.0143	0.0271	0.0177	0.0295	0.0169	0.0158	0.0086	0.01856
11	0.0087	0.0228	0.0174	0.0264	0.0146	0.0123	0.0142	0.01663
12	0.0139	0.0234	0.0145	0.0236	0.0158	0.0138	0.0135	0.01693
13	0.0106	0.02	0.0147	0.0209	0.0146	0.0069	0.0114	0.01416
14	0.0066	0.0231	0.0193	0.0182	0.0141	0.0074	0.0105	0.01417
15	0.0077	0.0207	0.0171	0.0145	0.0089	0.0083	0.0105	0.01253
16	0.0061	0.0191	0.0131	0.0156	0.0062	0.0081	0.0092	0.01106
17	0.006	0.0204	0.0131	0.0125	0.0063	0.0091	0.0054	0.01040
18	0.0039	0.0131	0.0089	0.0113	0.0056	0.0116	0.0037	0.00830
19	0.0045	0.011	0.0096	0.0118	0.0064	0.0106	0.0008	0.00781
20	0.0048	0.0101	0.01	0.0119	0.0067	0.0082	-0.0005	0.00731
21	0.002	0.0086	0.0068	0.0157	0.0021	0.0075	0.0032	0.00656
22	0.0066	0.0066	0.0097	0.0158	0.0012	0.0069	0.0024	0.00703
23	0.0069	0.0039	0.0101	0.0155	0.001	0.0068	0.0006	0.00640
24	0.0088	0.0051	0.0109	0.0178	0.0022	0.0075	0.0038	0.00801
25	0.0078	0.0054	0.0105	0.0185	0.0017	0.0078	0.0039	0.00794
26	0.0086	0.0035	0.0104	0.0178	0.0019	0.0075	0.0039	0.00766
27	0.0084	0.0037	0.011	0.0177	0.0016	0.0072	0.0044	0.00771
28	0.0111	0.0033	0.0074	0.0166	0.0008	0.0076	0.0032	0.00714
29	0.0092	0.0026	0.0072	0.017	0.0001	0.0049	0.0017	0.00610
30	0.01	0.0036	0.0073	0.0159	0.0025	0.0062	0.001	0.00664
31	0.0101	0.0027	0.006	0.0158	0.004	0.0067	-0.0011	0.00631
32	0.0097	0.0032	0.0062	0.0164	0.0031	0.0034	0.001	0.00614
33	0.011	0.0035	0.0049	0.0153	0.0035	-0.0008	0.0002	0.00537
34	0.0068	0.0021	0.0038	0.0144	0.0041	-0.0005	0.001	0.00453
35	0.0044	0.0019	0.0029	0.0139	0.0038	0.0011	0.0006	0.00409

LAMPIRAN 4. Hasil *clustering* artikel dengan AverageLink Agglomerative

1. Preposesing tanpa menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1604	0.1203	0.1108	0.1605	0.0734	0.1096	0.0647	0.11424
3	0.0431	0.0798	0.0914	0.1084	0.0305	-0.0253	0.0301	0.05114
4	-0.0434	0.0914	0.0858	0.0899	0.0683	0.0256	0.0569	0.05350
5	-0.0013	0.0381	0.0538	0.0683	0.0684	0.0267	0.0573	0.04447
6	-0.0015	0.03	0.0321	0.0335	0.0364	0.0245	0.056	0.03014
7	-0.044	0.0312	-0.0024	0.0284	0.0196	0.015	0.0374	0.01217
8	-0.0439	0.0254	-0.0061	0.0156	0.0161	0.0142	-0.0023	0.00271
9	-0.0424	0.0178	-0.0092	0.0043	-0.0098	0.0137	-0.0071	-0.00467
10	-0.0432	-0.0186	0.0133	0.0051	-0.0153	-0.0398	-0.0073	-0.01511
11	-0.044	-0.0194	0.0129	-0.0084	-0.0267	-0.0429	-0.0118	-0.02004
12	-0.0451	-0.0235	0.0049	0.0208	-0.0289	-0.0413	-0.012	-0.01787
13	-0.0104	-0.0459	-0.0146	0.016	-0.0309	-0.0043	0.015	-0.01073
14	-0.0133	-0.0457	-0.0101	0.0071	-0.0054	-0.0049	0.0092	-0.00901
15	-0.0176	-0.0214	-0.0128	-0.0103	-0.0142	-0.0054	0.0088	-0.01041
16	-0.0176	-0.0221	-0.0128	-0.0117	-0.0146	-0.0056	0.0058	-0.01123
17	-0.0187	-0.0226	-0.0166	-0.0229	-0.0096	-0.0061	-0.0109	-0.01534
18	-0.0219	-0.0354	-0.0166	-0.0335	-0.0267	-0.0063	-0.015	-0.02220
19	-0.0249	-0.0354	-0.0151	-0.0338	-0.0354	-0.006	-0.0179	-0.02407
20	-0.0248	-0.0358	-0.0206	-0.0355	-0.0248	-0.0093	-0.0311	-0.02599
21	-0.0245	-0.0353	-0.0215	-0.0361	-0.0269	-0.0147	-0.0316	-0.02723
22	-0.01	-0.0353	-0.0012	-0.0359	-0.0269	-0.0173	-0.0385	-0.02359
23	-0.016	-0.0362	-0.0018	-0.0358	-0.0296	-0.0249	-0.039	-0.02619
24	-0.0188	-0.0383	-0.0001	-0.0172	-0.0302	-0.0248	-0.0436	-0.02471
25	-0.0212	-0.0397	-0.0014	-0.0191	-0.0311	-0.0247	-0.0458	-0.02614
26	-0.0242	-0.0414	-0.0028	-0.0206	-0.0311	-0.0306	-0.0465	-0.02817

27	-0.0225	-0.0467	-0.0022	-0.0226	-0.0314	-0.0323	-0.0385	-0.02803
28	-0.0551	-0.0605	-0.0049	-0.0233	-0.0337	-0.0362	-0.0403	-0.03629
29	-0.0507	-0.0408	-0.0063	-0.021	-0.0337	-0.0356	-0.0289	-0.03100
30	-0.0515	-0.0369	-0.0142	-0.0091	-0.0284	-0.0365	-0.0288	-0.02934
31	-0.0447	-0.0383	-0.0187	-0.0069	-0.0105	-0.0369	-0.0285	-0.02636
32	-0.0467	-0.0393	-0.0207	-0.0071	-0.0123	-0.0379	-0.0116	-0.02509
33	-0.0467	-0.0419	-0.0207	-0.0097	-0.013	-0.0406	-0.0124	-0.02643
34	-0.0467	-0.043	-0.0219	-0.0096	-0.0155	-0.0438	-0.0125	-0.02757
35	-0.0482	-0.0431	-0.0221	-0.0125	-0.0164	-0.0477	-0.0135	-0.02907

2. Preposesing dengan menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.4035	0.1369	0.2198	0.0662	0.1943	0.2047	-0.0334	0.17029
3	0.0437	0.1367	0.0385	0.0568	-0.0174	0.0607	-0.0395	0.03993
4	0.02	0.0916	0.065	0.0676	-0.0173	0.0577	0.0107	0.04219
5	-0.0361	0.0746	0.058	0.0333	-0.03	-0.0271	0.0104	0.01187
6	-0.0361	0.08	0.0537	0.0178	0.014	0.0205	0.0116	0.02307
7	0.0094	0.0693	0.0261	0.0165	0.01	0.0157	0.0445	0.02736
8	0.0106	0.0499	-0.0285	0.006	0.01	0.0093	0.0447	0.01457
9	-0.0162	0.0578	-0.0271	0.0287	-0.0113	0.0292	0.0426	0.01481
10	-0.0216	0.0569	-0.0039	0.0212	0.0169	0.0202	0.0352	0.01784
11	0.0105	0.0408	-0.015	0.022	0.0185	0.0215	0.0301	0.01834
12	0.0109	0.0154	-0.0121	0.0153	0.0184	0.0118	-0.002	0.00824
13	0.0103	0.0125	-0.0123	0.0064	0.0191	0.0039	-0.0021	0.00540
14	0.0104	0.0059	-0.013	-0.0052	0.019	-0.0028	-0.0027	0.00166
15	0.0091	0.006	-0.018	0.0129	0.019	-0.0028	-0.0256	0.00009
16	0.001	0.0042	-0.0182	0.0119	0.0153	-0.0063	-0.026	-0.00259
17	-0.0014	0.0029	-0.0194	0.0054	-0.013	-0.011	-0.0283	-0.00926
18	-0.0042	0.0027	-0.0205	0.0013	-0.0176	-0.014	-0.0285	-0.01154

19	-0.0067	0.0008	-0.0211	-0.0001	-0.0177	-0.017	-0.0317	-0.01336
20	-0.0113	-0.0094	-0.0231	-0.0007	-0.0188	-0.0171	-0.0336	-0.01629
21	-0.0119	-0.0094	-0.0282	-0.0007	-0.0198	-0.0125	-0.0341	-0.01666
22	-0.012	-0.0126	-0.0283	-0.0014	-0.0242	-0.0125	-0.0325	-0.01764
23	-0.0278	-0.0181	-0.0324	-0.0059	-0.0254	-0.0174	-0.033	-0.02286
24	-0.0293	-0.0185	-0.0324	-0.0097	-0.0253	-0.0151	-0.0323	-0.02323
25	-0.0312	-0.0196	-0.0332	0.005	-0.0271	-0.0155	-0.0357	-0.02247
26	-0.0237	-0.0241	-0.039	0.0072	-0.0303	-0.017	-0.0359	-0.02326
27	-0.0262	-0.0313	-0.0247	0.0059	-0.037	-0.0175	-0.0359	-0.02381
28	-0.0314	-0.0315	-0.0277	0.0057	-0.0335	-0.0017	-0.0364	-0.02236
29	-0.0394	-0.0295	-0.0333	0.001	-0.0364	-0.0028	-0.0322	-0.02466
30	-0.0412	-0.0318	-0.0338	-0.0005	-0.0366	-0.0044	-0.0325	-0.02583
31	-0.028	-0.0318	-0.0307	-0.0014	-0.0307	-0.0031	-0.0325	-0.02260
32	-0.0278	-0.0324	-0.0305	-0.0042	-0.0354	-0.0033	-0.034	-0.02394
33	-0.0279	-0.0328	-0.0314	-0.0027	-0.0361	-0.0063	-0.0342	-0.02449
34	-0.0282	-0.021	-0.0314	-0.0063	-0.0371	-0.009	-0.0345	-0.02393
35	-0.0322	-0.0295	-0.0397	-0.0061	-0.0374	-0.0109	-0.0345	-0.02719

LAMPIRAN 5. Hasil *clustering* artikel ilmiah dengan Birch

1. Preposesing tanpa menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1115	0.1064	0.1127	0.1046	0.0959	0.1002	0.0971	0.10406
3	0.0792	0.0776	0.0585	0.0644	0.0783	0.0757	0.0743	0.07257
4	0.0628	0.0735	0.0557	0.0562	0.0601	0.0531	0.0631	0.06064
5	0.065	0.0539	0.0558	0.0464	0.0551	0.0555	0.0624	0.05630
6	0.0547	0.0466	0.0483	0.0462	0.0471	0.0572	0.0497	0.04997
7	0.0537	0.0391	0.0425	0.0413	0.0454	0.0498	0.0456	0.04534
8	0.0467	0.0386	0.037	0.0388	0.041	0.0387	0.043	0.04054

9	0.0416	0.0404	0.037	0.0413	0.039	0.0373	0.0377	0.03919
10	0.035	0.0371	0.0403	0.0326	0.0374	0.0283	0.0395	0.03574
11	0.0353	0.0388	0.0382	0.0289	0.0388	0.028	0.0373	0.03504
12	0.0356	0.0339	0.0355	0.0267	0.0404	0.0309	0.0395	0.03464
13	0.0371	0.029	0.0341	0.0285	0.0394	0.0314	0.0324	0.03313
14	0.0354	0.0307	0.0329	0.0297	0.0351	0.032	0.0346	0.03291
15	0.038	0.0317	0.0333	0.0307	0.0352	0.0333	0.0347	0.03384
16	0.0353	0.0288	0.0348	0.0271	0.0359	0.0351	0.0343	0.03304
17	0.0333	0.0261	0.0349	0.0257	0.0351	0.0352	0.031	0.03161
18	0.0323	0.0261	0.0349	0.0277	0.0349	0.032	0.0323	0.03146
19	0.0308	0.0274	0.0322	0.0275	0.0357	0.0286	0.0324	0.03066
20	0.0322	0.0264	0.0307	0.0262	0.0339	0.028	0.0319	0.02990
21	0.0339	0.0282	0.0307	0.0254	0.0344	0.0291	0.0317	0.03049
22	0.0357	0.027	0.0261	0.0269	0.0298	0.0299	0.0305	0.02941
23	0.0362	0.0275	0.0272	0.0275	0.0304	0.0301	0.0308	0.02996
24	0.0354	0.0259	0.0277	0.0279	0.03	0.0316	0.0281	0.02951
25	0.0314	0.0253	0.0267	0.0285	0.0306	0.0304	0.0279	0.02869
26	0.032	0.0262	0.0258	0.0268	0.0292	0.0315	0.0271	0.02837
27	0.0324	0.0252	0.0265	0.0256	0.0264	0.0316	0.0256	0.02761
28	0.0326	0.0235	0.0252	0.023	0.0246	0.0297	0.0266	0.02646
29	0.0305	0.0221	0.0257	0.0237	0.0256	0.0309	0.0267	0.02646
30	0.0285	0.0222	0.0266	0.0233	0.0259	0.0308	0.0275	0.02640
31	0.0281	0.0226	0.0245	0.022	0.0256	0.0304	0.0283	0.02593
32	0.028	0.0232	0.0229	0.0233	0.0235	0.0291	0.029	0.02557
33	0.0264	0.0233	0.0218	0.0238	0.0242	0.0292	0.0294	0.02544
34	0.0256	0.0225	0.0221	0.0223	0.0215	0.0282	0.0282	0.02434
35	0.0258	0.0229	0.0212	0.0234	0.0221	0.0283	0.0292	0.02470

2. Preposesing dengan menggunakan tahap lemmatization

Preprocessing Tanpa Lemmatization								
Jumlah K	P 1	P 2	P 3	P 4	P 5	P 6	P 7	Rata-Rata
2	0.1046	0.1035	0.0954	0.1039	0.1019	0.0912	0.0857	0.09803
3	0.0725	0.0771	0.0842	0.0732	0.0795	0.0717	0.0772	0.07649
4	0.0551	0.0734	0.0663	0.0731	0.0509	0.0653	0.0483	0.06177
5	0.0426	0.0569	0.0542	0.048	0.0488	0.0601	0.0386	0.04989
6	0.0461	0.0495	0.0508	0.0424	0.0438	0.0589	0.0363	0.04683
7	0.0396	0.0458	0.048	0.0403	0.0369	0.0507	0.0292	0.04150
8	0.0405	0.0373	0.0439	0.0429	0.0367	0.0368	0.0273	0.03791
9	0.0383	0.0412	0.0396	0.0432	0.0371	0.0359	0.0235	0.03697
10	0.0406	0.0391	0.0388	0.0429	0.0384	0.0319	0.0296	0.03733
11	0.0379	0.0348	0.0365	0.0383	0.0371	0.0334	0.0309	0.03556
12	0.0323	0.0319	0.0295	0.0397	0.0369	0.0329	0.0314	0.03351
13	0.0336	0.0333	0.0311	0.0402	0.0315	0.0346	0.0321	0.03377
14	0.0315	0.0361	0.0299	0.0395	0.0298	0.036	0.0325	0.03361
15	0.034	0.0333	0.0288	0.0363	0.0273	0.0275	0.0348	0.03171
16	0.0356	0.0338	0.0316	0.0353	0.0295	0.0266	0.0345	0.03241
17	0.036	0.0328	0.0284	0.0326	0.0307	0.0241	0.036	0.03151
18	0.0342	0.0326	0.0297	0.0292	0.031	0.0232	0.0366	0.03093
19	0.0349	0.0313	0.0304	0.0287	0.0281	0.0246	0.0355	0.03050
20	0.0356	0.0305	0.0275	0.028	0.0278	0.0225	0.0337	0.02937
21	0.0355	0.031	0.0281	0.0278	0.0291	0.024	0.0306	0.02944
22	0.0335	0.0292	0.0252	0.0269	0.0263	0.0235	0.0323	0.02813
23	0.0339	0.0302	0.0247	0.0273	0.0271	0.024	0.0313	0.02836
24	0.0316	0.0274	0.0226	0.0272	0.0285	0.0237	0.0307	0.02739
25	0.0322	0.0278	0.0233	0.0267	0.0261	0.0244	0.0304	0.02727
26	0.0311	0.028	0.0246	0.0273	0.0272	0.0245	0.0283	0.02729
27	0.0316	0.0289	0.0232	0.0289	0.0241	0.0257	0.0266	0.02700

28	0.0323	0.0291	0.0223	0.0292	0.0257	0.0246	0.0269	0.02716
29	0.0324	0.0297	0.0228	0.0279	0.0252	0.0256	0.0224	0.02657
30	0.0313	0.0297	0.0225	0.0269	0.0237	0.026	0.0206	0.02581
31	0.029	0.0291	0.0237	0.0263	0.0221	0.0264	0.0198	0.02520
32	0.0296	0.0282	0.0225	0.0257	0.0223	0.0266	0.0197	0.02494
33	0.0259	0.0281	0.023	0.0263	0.0231	0.0278	0.0193	0.02479
34	0.0261	0.0291	0.0237	0.0267	0.0228	0.0274	0.0198	0.02509
35	0.0257	0.0289	0.0242	0.0238	0.0219	0.0285	0.0182	0.02446

BIOGRAFI PENULIS

Amelia Sahira Rahma merupakan wanita sholeha yang berada ditengah keluarga bahagia dari perpaduan kota Surabaya (Ibu Failun Indrawati) dan Sidoarjo (Bapak Rindi Sudjono). Lahir sebagai anak ketiga dari empat bersaudara pada tanggal 17 Agustus 1993 di kota Surabaya, Jawa Timur. Amelia mengenyam pendidikan sekolah dasar di SDN Baratajaya Surabaya pada tahun 1999 sampai 2005. Dilanjutkan dengan sekolah menengah pertama di SMP Muhammadiyah 5 Surabaya pada tahun 2005 sampai 2008, dan sekolah menengah atas di SMA Khadijah Surabaya pada tahun 2008 sampai 2011. Mengingat banyaknya cita-cita dan rasa penasaran akan teknologi, Amelia memutuskan untuk melanjutkan pendidikan di Universitas Muhammadiyah Sidoarjo guna mendalami ilmu pengetahuan dibidang Teknik Informatika dan mendapatkan gelar S.Kom pada tahun 2011 sampai 2015. Pada tahun 2016 sampai 2018, dengan penuh keberkahan dari ALLAH SWT, Amelia mendapatkan kesempatan untuk melanjutkan pendidikan S2 di Program Pascasarjana, Departemen Informatika, ITS Surabaya guna mengeksplor dan lebih mendalami bidang keahlian Komputasi Cerdas dan Visualisasi, serta mendapatkan gelar M.Kom atas pendidikan tersebut.

Email : **ameliasahirarahma170893@gmail.com**