BAB 4
HASIL PENELITIAN

4.1 RINGKASAN HASIL PENELITIAN

Penelitian ini merupakan pengembangan lanjutan sistem Fact Checker Directory. Pada sistem yang sebelumnya, fitur utama yang dimiliki adalah pencarian berita *hoax* dan menklasifikasikannya kedalam beberapa kategori berita *hoax*. Berita *hoax* yang ditampilkan diambil dari proses *crawling* beberapa *website* yang mengklasifikasikan dan menyediakan berita *hoax*.

Pada penelitian ini ditambahkan fitur untuk memvisualisasikan relasi antara berita *hoax* menggunakan metode *Topic Modeling* yang dipadukan dengan metode TF-IDF dan *Social Network Analisys*. Data berita *hoax* yang ada pada *database* sistem akan diolah pada proses *pre-processing* kemudian dibuat pemodelan topiknya. Setelah pemodelan topik terbentuk, tiap topik yang dihasilkan akan dijadikan node dan saling dihubungkan keterkaitannya.

4.2 LDA TOPIC MODELING

Dalam *Topic Modeling* terdapat banyak faktor yang dapat mempengaruhi keoptimalan suatu model. Salah satu cara untuk melakukan penyesuaian dan evaluasi model supaya mendapatkan model yang optimal adalah menggunakan *Coherence Measurement*. *Coherence Measurement* adalah metriks evaluasi yang umumnya digunakan untuk mengevaluasi pemodelan topik dengan mengukur tingkat kesamaan semantik antara kata-kata dengan skor tinggi dalam suatu topik (Bellaouar et al., 2021). *Coherence Measurement* dalam pemodelan topik dapat digunakan untuk mengukur seberapa dapat ditafsirkan topik tersebut bagi manusia. Dalam hal ini, topik direpresentasikan sebagai N kata teratas dengan probabilitas tertinggi untuk menjadi bagian dari topik tersebut. Secara singkat, skor koherensi mengukur seberapa mirip kata-kata ini satu sama lain.

Pada penelitian ini tipe *Coherence Measurement* yang digunakan adalah CV *Coherence*, CV Coherence membuat vektor konten kata-kata menggunakan tingkat

kemunculan suatu kata lalu menghitung nilai dengan NMPI (*Normalized Pointwise Mutual Information*) (Bellaouar et al., 2021). Kemudian, karena data yang dihasilkan pada penelitian ini bersifat *dynamic* maka dalam pengujiannya digunakan studi kasus menggunakan data berita hoax dengan *keyword* "Jokowi".

4.2.1 Pengaruh Jumlah Data dan Tingkat Keoptimalan Topic Modeling

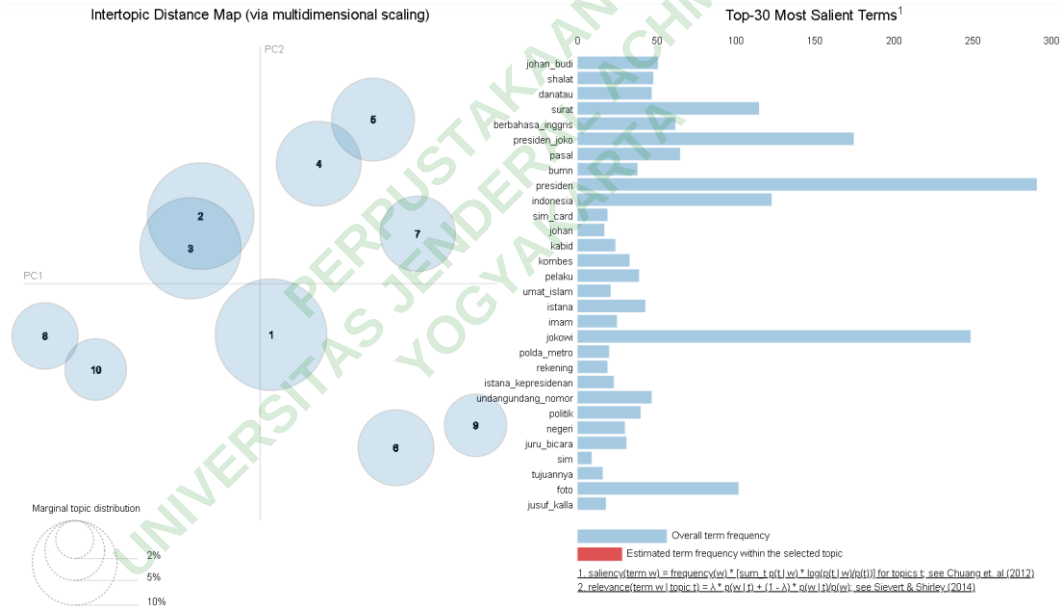
Pada platform Fact Checker Directory jumlah data pada tiap kata kunci tertentu akan berbeda-beda menyesuaikan jumlah data yang berhasil ditemukan oleh *crawler*. Untuk itu dilakukan percobaan dengan jumlah sampel data yang berbeda yaitu sampel pertama dengan kepadatan data yang tinggi sejumlah 2041 artikel, sampel kedua dengan kepadatan data sedang sejumlah 200 artikel, dan sampel ketiga dengan kepadatan data rendah sejumlah 20 artikel. Percobaan dilakukan empat kali iterasi dengan jumlah topik sebanyak sepuluh topik pada tiap sampel data untuk mengetahui rata-rata nilai evaluasi karena tiap pembentukan model akan menghasilkan sedikit perbedaan pada tiap modelnya.

Dapat dilihat pada Tabel 4.1 nilai hasil evaluasi model dengan tiga sampel berbeda menghasilkan nilai yang bervariasi pada tiap iterasinya. Pada iterasi pertama nilai evaluasi tertinggi ada pada sampel dengan kepadatan data sedang dengan nilai 0.4805, pada iterasi kedua, ketiga, dan keempat nilai tertinggi ada pada sampel data dengan kepadatan rendah dengan nilai pada iterasi kedua 0.4911, pada iterasi ketiga 0.6506, dan pada iterasi keempat 0.5067. Nilai yang bervariasi pada tiap sampel dapat dijelaskan karena berdasarkan definisi dan rumus perhitungan *Coherence Measurement*, nilainya bergantung pada data yang digunakan untuk menghitungnya. Sebagai contoh dalam kasus data dengan kepadatan tinggi dengan jumlah data 2041 baris, skor 0,48 mungkin cukup baik tetapi dalam kasus data dengan kepadatan sedang dan rendah tidak cukup baik.

Tabel 4.1 Evaluasi Coherence Measurement

Sampel	Iterasi 1	Iterasi 2	Iterasi 3	Iterasi 4
Kepadatan data tinggi (2041 baris)	0.4048	0.4293	0.4831	0.4170
Kepadatan data sedang (200 baris)	0.4805	0.4574	0.4577	0.4361
Kepadatan data rendah (20 baris)	0.4056	0.4911	0.6506	0.5067

Pada Gambar 4.1, Gambar 4.3, dan Gambar 4.5 dapat dilihat persebaran topik dengan perbedaan susunan kata yang dimunculkan dan jumlah frekuensi kemunculan tiap kata. Makin banyak data yang disediakan, maka kemungkinan susunan kata semakin bervariasi dengan jumlah frekuensi kemunculan yang tinggi.

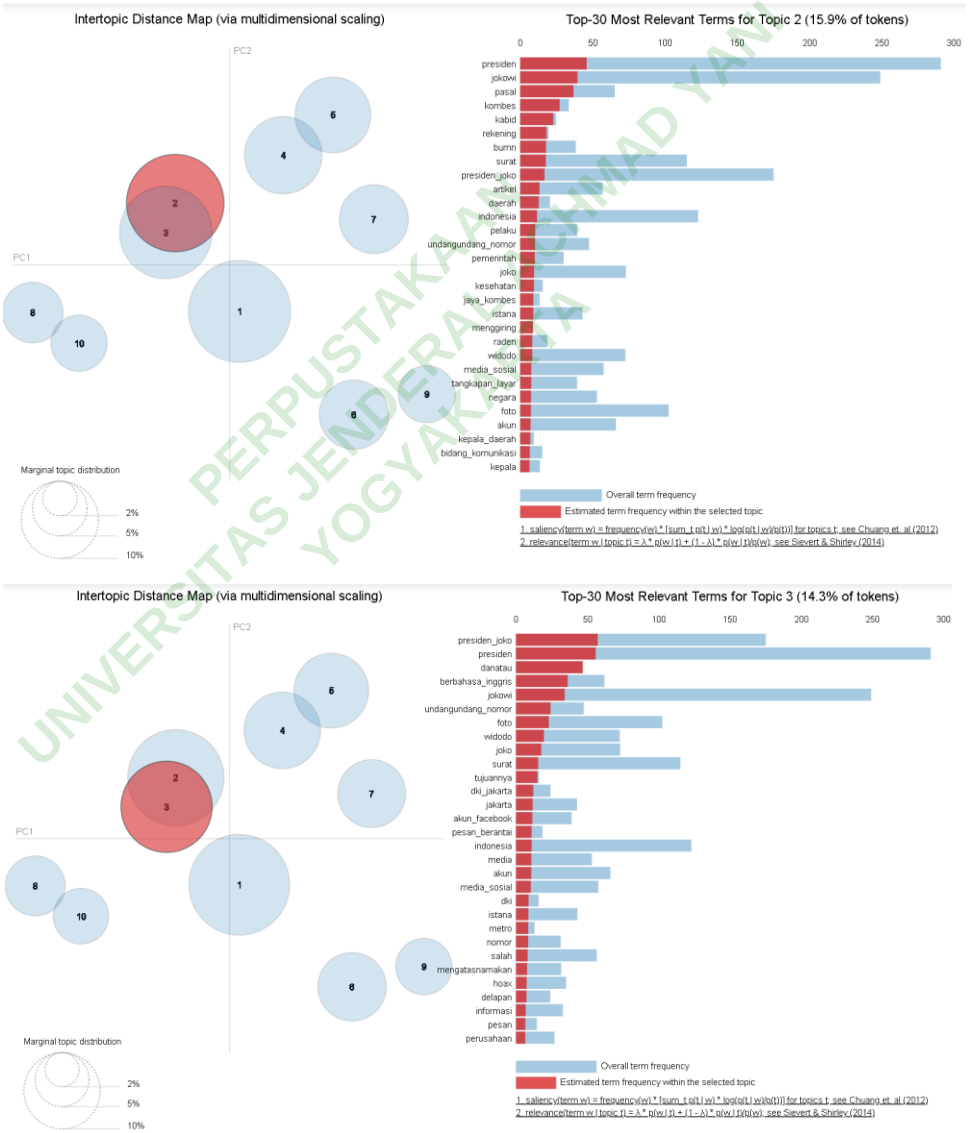


Gambar 4.1 Visualisasi Model Sampel Kepadatan Data Tinggi

Pada Gambar 4.1 Sampel dengan kepadatan data tinggi, kata dengan frekuensi kemunculan paling tinggi ada pada kata “presiden” dengan frekuensi kemunculan hampir menyentuh 300 kali kemunculan. Sedangkan untuk kata yang paling jarang muncul ada pada kata “sim” dengan frekuensi kemunculan dibawah 50 kali kemunculan. Selain itu persebaran topik yang dicapai cukup optimal dengan

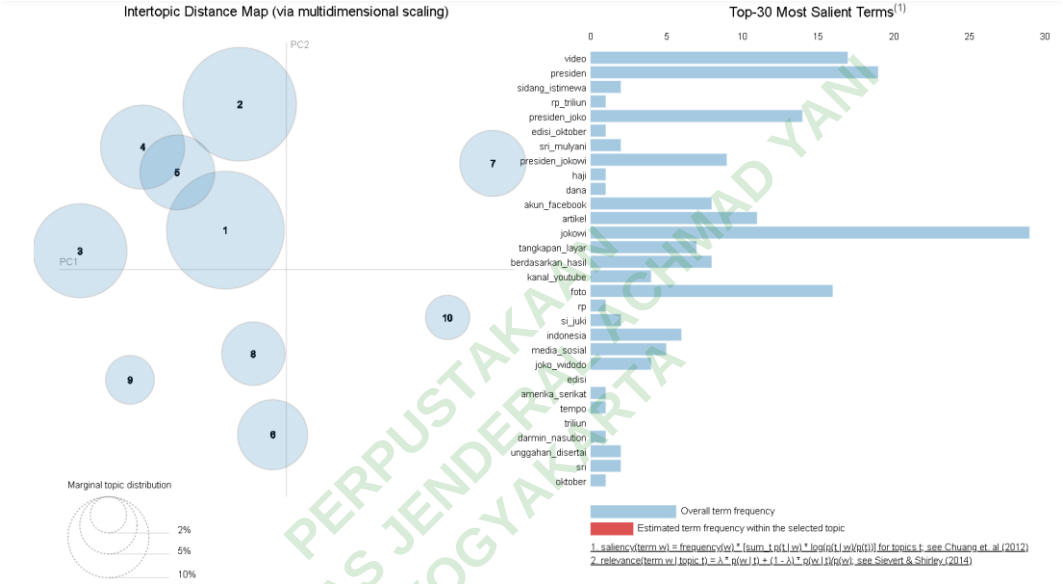
hanya ada dua topik yang hampir sepenunhnya saling tumpeng tindih yaitu topik 2 dan topik 3.

Pada Gambar 4.2 dapat dilihat perbandingan kata pada topik 2 dan topik 3 yang saling tumpang tindih. Pada topik 2, lima kata penyusun teratas adalah “presiden”, “Jokowi”, “pasal”, “kombes”, dan “kabit”. Sedangkan pada topik 3, lima kata penyusun teratas adalah “presiden_joko”, “presiden”, “danatau”, “berbahasa_inggris”, dan “Jokowi”.



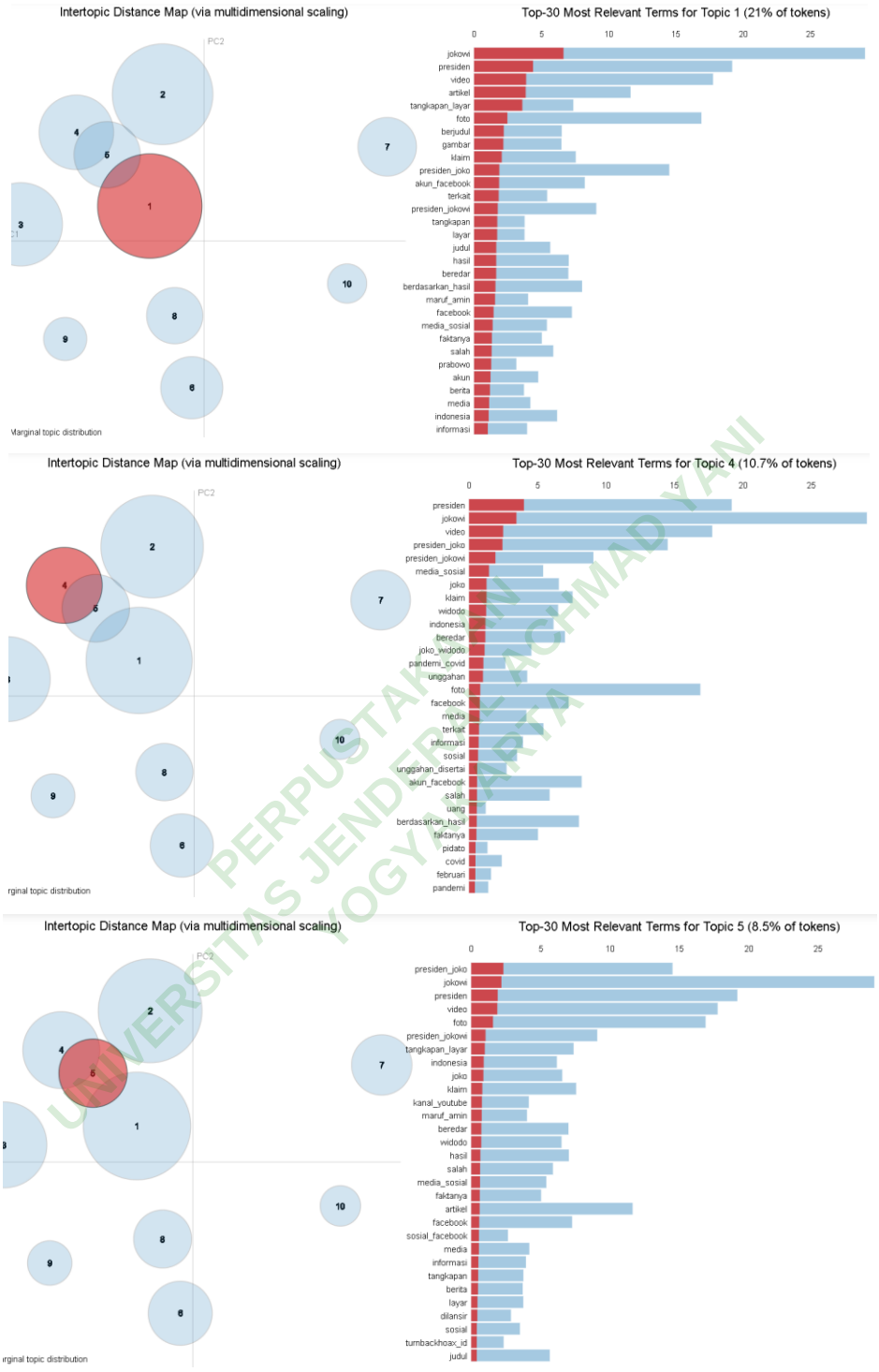
Gambar 4.2 Perbandingan Penyusun Kata Pada Topik 2 dan 3

Pada Gambar 4.3 sampel dengan kepadatan data sedang, kata dengan frekuensi kemunculan paling tinggi ada pada kata “jokowi” dengan frekuensi kemunculan hampir menyentuh 30 kali kemunculan. Sedangkan untuk kata yang paling jarang muncul ada pada kata “edisi” dam “triliun” dengan frekuensi kemunculan 1 kali kemunculan. Persebaran topik pada sampel data dengan kepadatan data sedang terdapat tiga topik yang saling tumpang tindih yaitu topik 4, topik 5, dan topik 1.



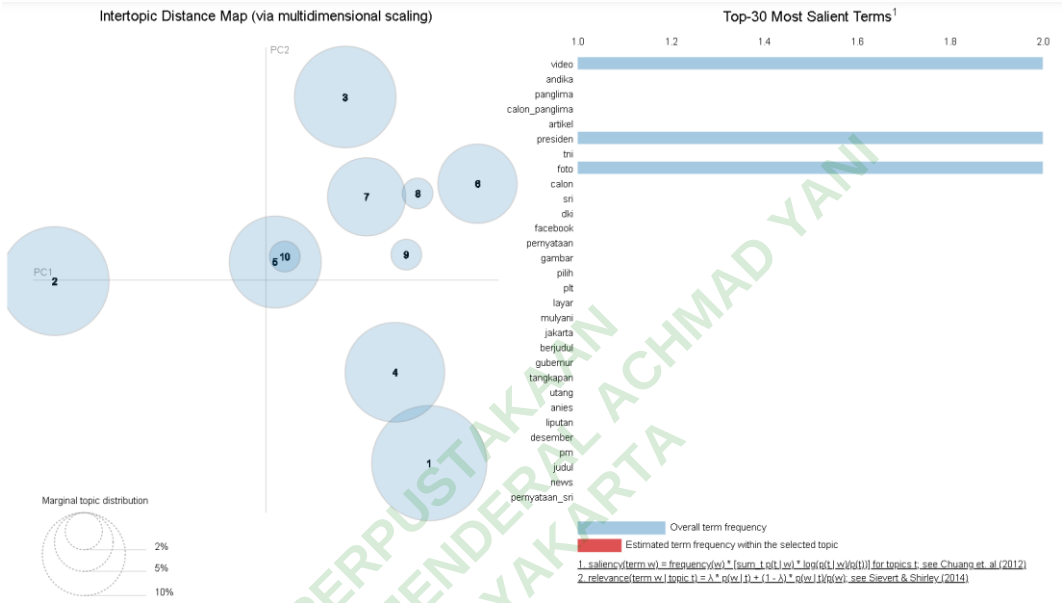
Gambar 4.3 Visualisasi Model Sampel Kepadatan Data Sedang

Pada Gambar 4.4 dapat dilihat perbandingan kata pada topik 1, topik 4 dan topik 5 yang saling tumpang tindih. Pada topik 1, lima kata penyusun teratas adalah “Jokowi”, “presiden”, “video”, “artikel”, dan “tangkapan_layar”. Pada topik 4, lima kata penyusun teratas adalah “presiden”, “Jokowi”, “video”, “presiden_joko”, dan “presiden_jokowi”. Sedangkan pada topik 5, lima kata penyusun teratas adalah “presiden_joko”, “Jokowi”, “presiden”, “video”, dan “foto”.



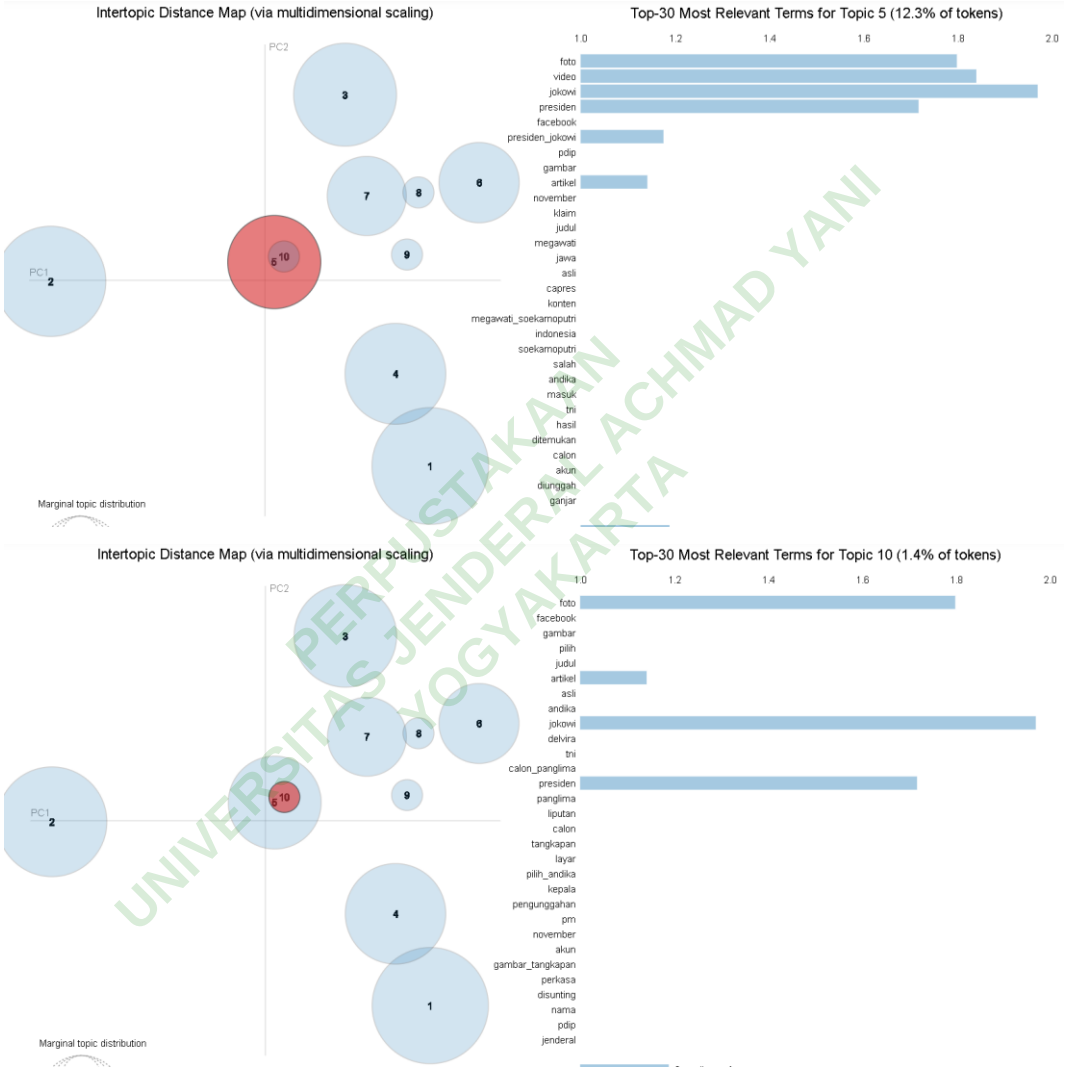
Gambar 4.4 Perbandingan Penyusun Kata Pada Topik 1, 4, dan 5

Pada Gambar 4.5 sampel dengan kepadatan data rendah frekuensi kemunculan kata terbanyak menyentuh 20 kali kemunculan pada tiga kata yaitu kata “video”, “presiden”, dan “foto”, sedangkan kata penyusun lainnya hanya muncul satu kali. Persebaran topik pada sampel data dengan kepadatan data rendah terdapat tiga topik yang saling tumpang tindih sepenuhnya yaitu topik 5 dan 10.



Gambar 4.5 Visualisasi Sampel Kepadatan Data Rendah

Pada Gambar 4.6 dapat dilihat perbandingan kata pada topik 5 dan topik 10 yang saling tumpang tindih. Pada topik 2, lima kata penyusun teratas adalah “presiden”, “Jokowi”, “pasal”, “kombes”, dan “kabit”. Sedangkan pada topik 3, lima kata penyusun teratas adalah “presiden_joko”, “presiden”, “danatau”, “berbahasa_inggris”, dan “Jokowi”.

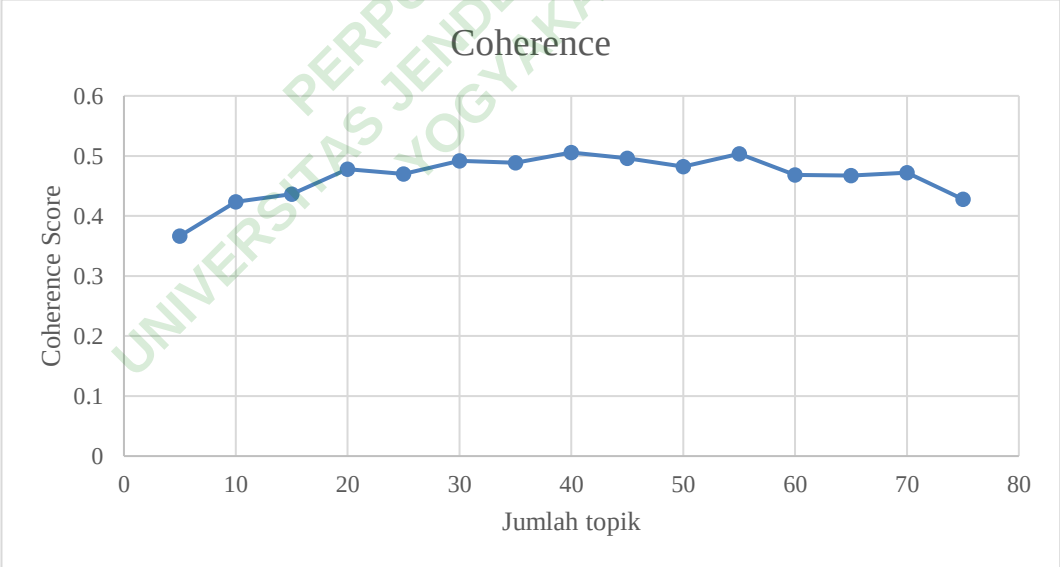


Gambar 4.6 Perbandingan Penyusun Kata Pada Topik 5 dan 10

4.2.2 Pengaruh Jumlah Topik Pada Topic Modeling

Salah satu kesulitan dalam LDA *Topic Modeling* adalah menentukan jumlah topik terbaik untuk memungkinkan model topik terbaik, sementara jumlah topik yang berbeda kemungkinan akan menghasilkan struktur korpus yang sangat berbeda. Jumlah topik yang tidak mencukupi dapat membuat model LDA terlalu kasar untuk mengidentifikasi pengklasifikasi yang akurat. Di sisi lain, jumlah topik yang berlebihan dapat menghasilkan model yang terlalu kompleks, sehingga menyulitkan interpretasi dan validasi subyektif (W. Zhao et al., 2015).

Untuk menentukan jumlah topik yang optimal penelitian ini melakukan evaluasi menggunakan nilai CV *Coherence Measurement*. Untuk sampel data yang digunakan adalah berita *hoax* dengan kata kunci "jokowi" dengan jumlah sampel 500 artikel. Pada Gambar 4.7 data dengan sampel 500 sampel data, hasil evaluasi *Coherence Measurement* menunjukkan nilai tertinggi pada jumlah topik 40 topik dengan nilai 0.50, kemudian untuk jumlah topik selanjutnya mengalami penurunan nilai secara bertahap lalu kembali menyentuh nilai 0.50 pada jumlah topik 55 topik.



Gambar 4.7 Coherence Score 500 Sampel Data

Pada Tabel 4.2 dapat dilihat semakin banyak sampel data yang digunakan maka semakin banyak jumlah topik yang dapat dihasilkan dengan nilai *Coherence Measurement* yang paling optimal. Sebaliknya, jika semakin sedikit sampel data yang digunakan maka semakin sedikit jumlah topik yang dapat dihasilkan dengan nilai *Coherence Measurement* yang paling optimal.

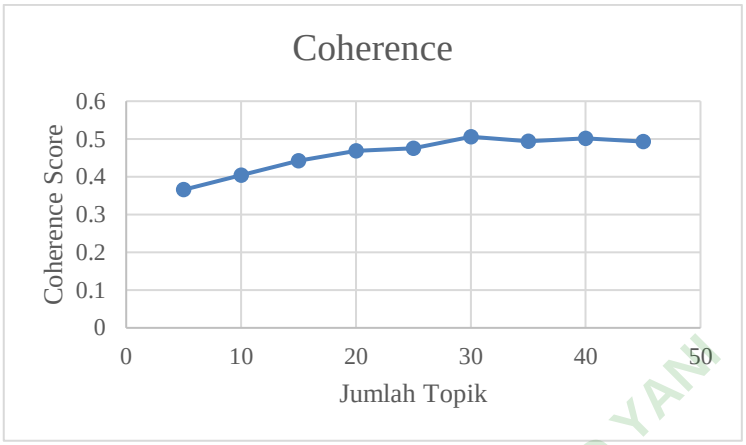
Tabel 4.2 Hasil Evaluasi Pemodelan Topik

Jumlah Topik	Sampel 10 Data	Sampel 300 Data	Sampel 500 Data
5	0.43	0.45	0.36
10	0.50	0.47	0.42
15	0.55	0.48	0.43
20	0.53	0.47	0.47
25	0.53	0.49	0.47
30	0.54	0.49	0.49
35	0.52	0.48	0.48
40	0.47	0.46	0.50
45	0.50	0.46	0.49
50		0.45	0.48
55		0.42	0.50
60		0.40	0.46
65		0.39	0.46
70		0.44	0.47
75		0.42	0.45

Penentuan jumlah topik maksimal dalam evaluasi *Coherence Measurement* dilakukan secara manual dengan melakukan beberapa kali percobaan dengan jumlah topik maksimal yang berbeda-beda. Pemilihan jumlah topik maksimal dalam evaluasi *Coherence Measurement* dipilih berdasarkan nilai fluktuasi yang dihasilkan pada tiap hasil evaluasi. Iterasi percobaan akan dihentikan ketika nilai *coherence score* mengalami penurunan terus-menerus setelah didapatkan nilai *coherence score* nilai *coherence score* tertinggi.

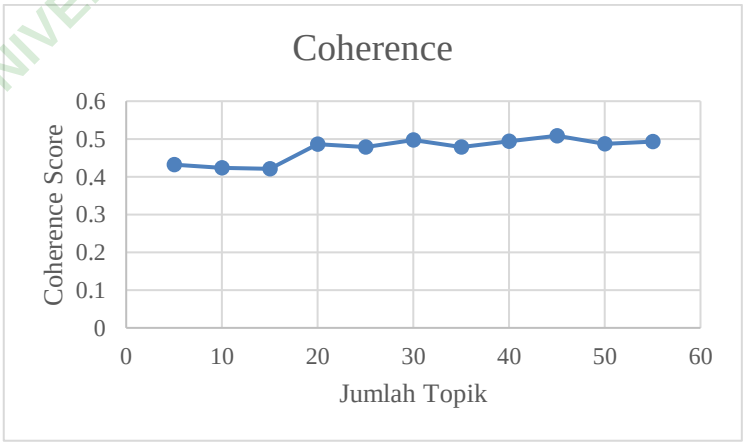
Dapat dilihat pada Gambar 4.8 hasil evaluasi dengan menggunakan 500 sampel data dan jumlah topik maksimal sebanyak 45 topik, nilai *coherence score* tertinggi ada pada 40 topik dengan nilai *coherence score* sebesar 0.5. Karena masih

belum terdapat penurunan nilai *coherence score* yang signifikan, maka dilakukan iterasi selanjutnya dengan jumlah topik maksimal yang ditambahkan.



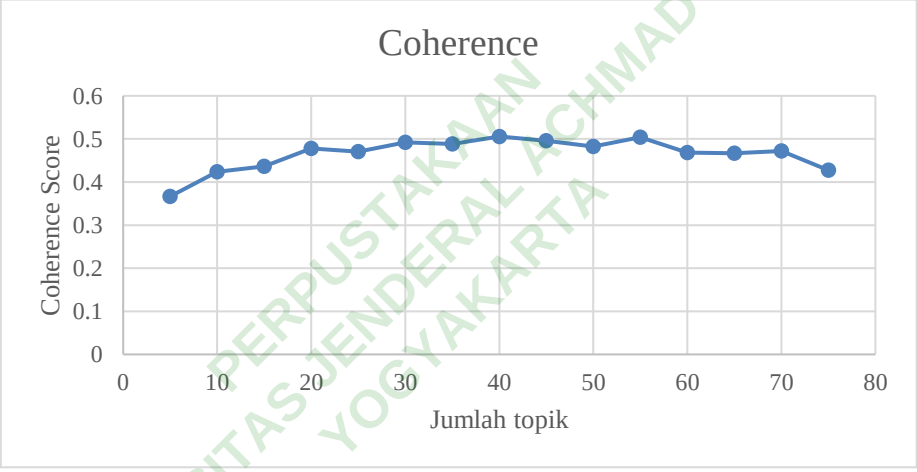
Gambar 4.8 Coherence Score 500 Sampel Data & 45 Topik

Karena pada percobaan dengan topik maksimal sebanyak 45 topik masih belum optimal, maka percobaan dilanjutkan dengan jumlah topik maksimal yang ditambahkan menjadi 55 topik. Pada Gambar 4.9 hasil evaluasi dengan menggunakan 500 sampel data dan jumlah topik maksimal sebanyak 55 topik, nilai *coherence score* tertinggi ada pada 45 topik dengan nilai *coherence score* sebesar 0.5. Karena masih belum terdapat penurunan nilai *coherence score* yang signifikan, maka dilakukan iterasi selanjutnya dengan jumlah topik maksimal yang ditambahkan.



Gambar 4.9 Coherence Score 500 Sampel Data & 55 Topik

Karena pada percobaan dengan topik maksimal sebanyak 55 topik masih belum optimal, maka percobaan dilanjutkan dengan jumlah topik maksimal yang ditambahkan menjadi 75 topik. Pada Gambar 4.10 hasil evaluasi dengan menggunakan 500 sampel data dan jumlah topik maksimal sebanyak 75 topik, nilai *coherence score* tertinggi ada pada 55 topik dengan nilai *coherence score* sebesar 0.50. Dapat dilihat setelah nilai *coherence score* tertinggi didapatkan pada 55 topik, nilai *coherence score* selanjutnya mulai mengalami penurunan secara bertahap. Karena nilai *coherence score* mulai mengalami penurunan, maka dapat disimpulkan bahwa jumlah topik paling optimal berada pada jumlah topik sebanyak 55 topik.

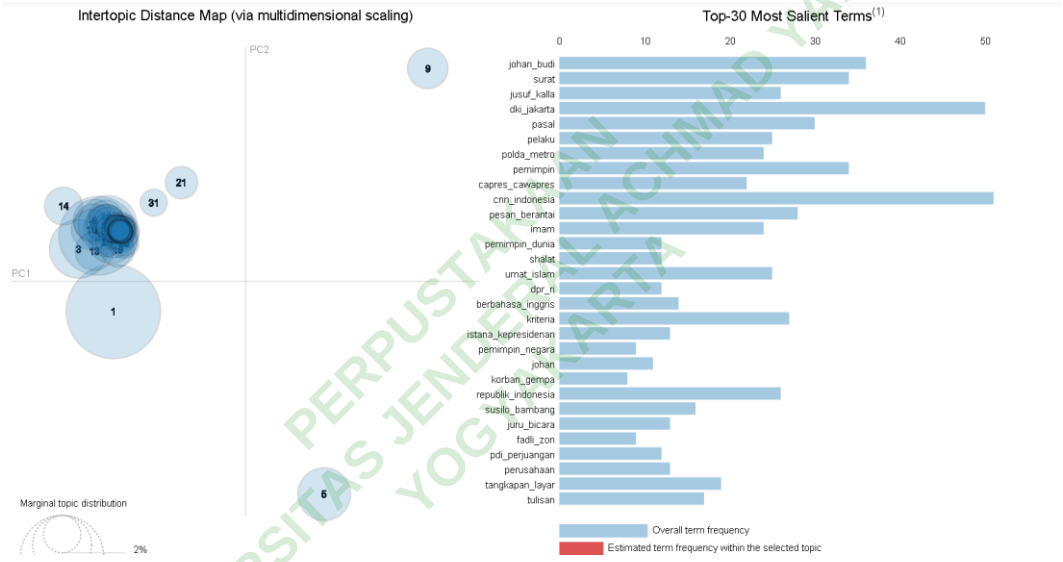


Gambar 4.10 Coherence Score 500 Sampel Data & 75 Topik

4.3 ANALISIS TOPIC MODELING

Pada penelitian ini data yang digunakan dalam pembentukan SNA merupakan data *dynamic* sehingga untuk melakukan percobaan maka pada penelitian ini akan menggunakan studi kasus dengan kata kunci "Jokowi". Untuk mempersiapkan analisis SNA, perlu dilakukan pelabelan manual terlebih dahulu pada hasil pemodelan topik. Proses pelabelan topik dilakukan dengan menganalisis susunan kata pada *term of topic* yang dihasilkan pada proses pemodelan topik.

Dalam pelabelan topik untuk lebih memahami struktur kata yang menyusun setiap topik pada model topik, dilakukan visualisasi *intertopic distance map* dari 55 topik yang didapatkan pada pemodelan topik. Visualisasi ini memberikan gambaran tentang bagaimana topik terkait satu sama lain berdasarkan persamaan kata yang muncul pada tiap topik. Gambar 4.11 menunjukkan *intertopic distance map* 55 topik yang dihasilkan oleh algoritma LDA, kita dapat melihat bahwa beberapa topik memiliki jarak satu sama lain dan beberapa topik lainnya tumpang tindih dengan topik lainnya. Topik yang tumpang tindih menunjukkan bahwa ada kata-kata serupa yang membangun topik yang berbeda.



Gambar 4.11 *Intertopic Distance Map*

Ide dasar representasi topik dalam LDA adalah topik yang terdiri dari sekumpulan kata. Oleh karena itu, topik yang tumpang tindih umumnya akan berbagi beberapa kata dalam representasi topik tersebut. Berdasarkan Gambar 4.11, kita dapat melihat bahwa topik yang memiliki jarak antar topik yang jauh berarti topik tersebut tidak memiliki hubungan. Namun, topik dengan jarak yang dekat atau bahkan tumpang tindih antar topik berarti topik tersebut memiliki hubungan atau mungkin memiliki tema topik yang serupa.

Tabel 4.3 merangkum kategori topik dari proses agregasi sebagai hasil dari pemodelan topik. Tabel 4.3 menunjukkan bahwa proses agregasi topik menghasilkan enam kategori topik. Setiap topik yang dihasilkan terdiri dari beberapa kumpulan kata yang memiliki bobot kemunculan yang berbeda. Topik yang memiliki kemunculan kata yang sama dianggap memiliki kategori topik yang sama, kita bisa melihat kata mana yang paling signifikan dalam membangun kategori topik. Berdasarkan kesamaan kemunculan kata pada 55 topik, diperoleh enam kategori topik seperti terlihat pada Tabel 4.3 kolom deskripsi topik.

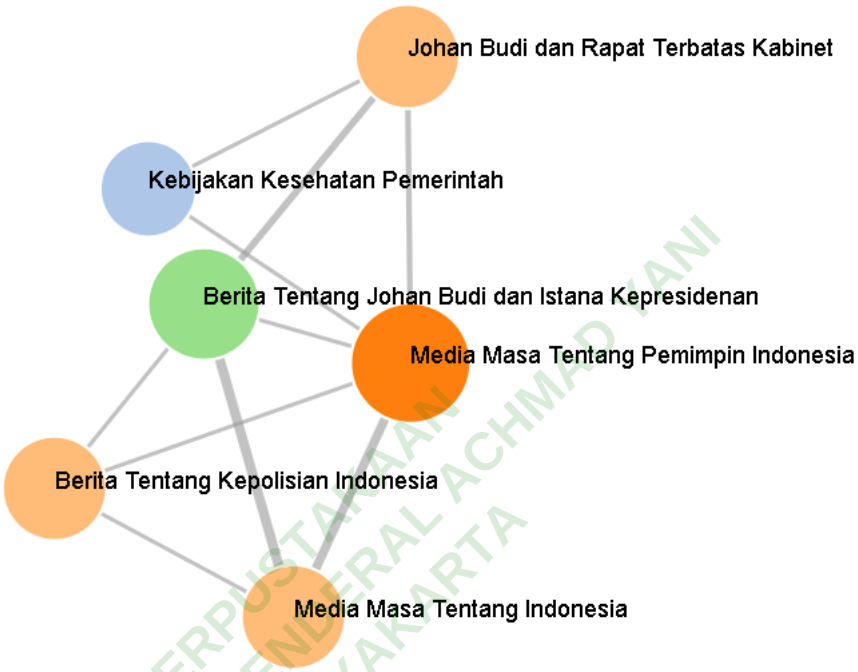
Tabel 4.3 Daftar Topik yang Diidentifikasi dari Klaster *Term* LDA

Kategori Topik	**Klaster *Term LDA***	**Deskripsi Topik**
Kategori 1	1	Media Masa Tentang Pemimpin Indonesia
Kategori 2	5	Berita Tentang Kepolisian Indonesia
Kategori 3	9	Berita Tentang Johan Budi dan Istana Kepresidenan
Kategori 4	21	Media Masa Tentang Indonesia
Kategori 5	31	Johan Budi dan Rapat Terbatas Kabinet
Kategori 6	2, 3, 4, 6, 7, 8, 10 , 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55	Kebijakan Kesehatan Pemerintah

4.4 SOCIAL NETWORK ANALYSIS

Setelah melakukan analisis terhadap *Intertopic Distance Map* dan melakukan agregasi pada topik untuk melakukan pelabelan topik. Didapatkan model SNA seperti pada Gambar 4.12. Dapat dilihat bahwa pada *Node* kategori topik 1 “Media Masa Tentang Presiden Indonesia” merupakan topik yang paling berpengaruh dengan jumlah koneksi paling banyak dengan kategori topik lainnya.

Selain itu kategori topik 1 menjadi penghubung bagi interaksi kategori topik lain di dalam jaringan.



Gambar 4.12 Model SNA

4.4.1 *Centrality Measures* dan Hasil Analisis SNA

Dalam teori graf dan network analysis, terdapat empat metode yang dapat digunakan untuk mengukur centrality, yaitu dengan metode menghitung *degree centrality*, *betweeness centrality*, *closeness centrality* dan *eigenvector centrality*. Pada penelitian ini akan digunakan tiga metode perhitungan, yaitu *degree centrality*, *betweeness centrality* dan *closeness centrality*.

Dapat dilihat pada Tabel 4.4 merupakan hasil hitung terhadap nilai degree centrality, betweeness centrality, dan closeness centrality. Diketahui bahwa kategori topik yang paling berpengaruh terhadap interaksi jaringan berita hoax Jokowi adalah kategori topik 1 tentang "Media Masa Tentang Pemimpin Indonesia". Dengan nilai degree centrality tertinggi kategori topik 1 menjadi topik paling berpengaruh dari total jumlah interaksi yang dihasilkan dengan topik lain.

Nilai betweenness centrality pada kategori topik 1 juga mendapatkan nilai yang paling tinggi, dimana nilai tersebut menjadikan kategori topik 1 sebagai penghubung bagi interaksi topik lain di dalam jaringan. Kemudian nilai closeness centrality tertinggi juga ada pada kategori topik 1, dimana nilai tersebut menjadikan topik kategori 1 sebagain kategori topik yang paling banyak memiliki kedekatan dengan kategori topik lainnya.

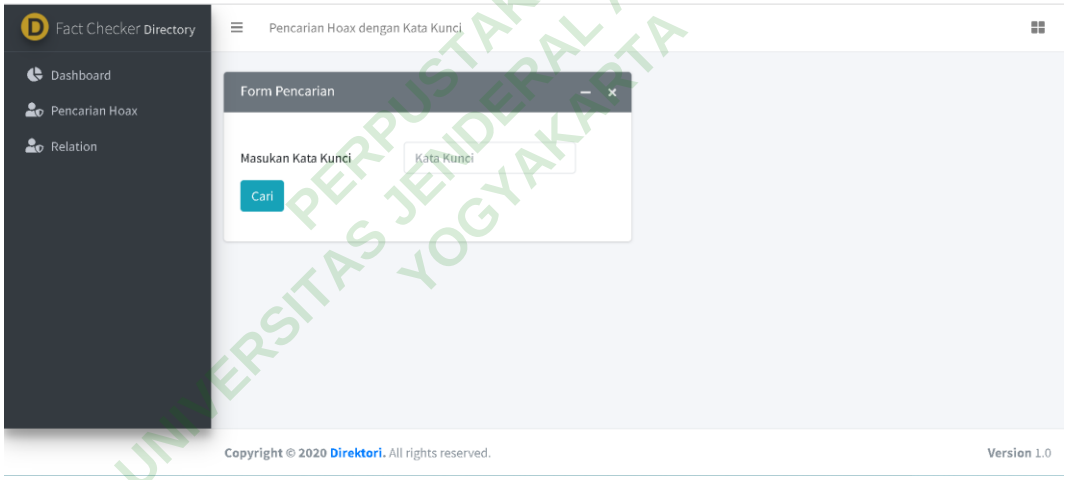
Tabel 4.4 Nilai *Centrality Node* Berpengaruh Pada SNA Jokowi

Node	**Degree Centrality**	**Betweeness Centrality**	**Closeness Centrality**
	Score / (Rank)	**Score / (Rank)**	**Score / (Rank)**
Kategori 1	5 / (1)	0.56 / (1)	1 / (1)
Kategori 2	3 / (3)	0.33 / (4)	0.71 / (3)
Kategori 3	4 / (2)	0.40 / (2)	0.83 / (2)
Kategori 4	3 / (3)	0.33 / (4)	0.71 / (3)
Kategori 5	3 / (3)	0.36 / (3)	0.71 / (3)
Kategori 6	2 / (4)	0.33 / (4)	0.62 / (4)

4.5 HASIL TAMPILAN ANTARMUKA

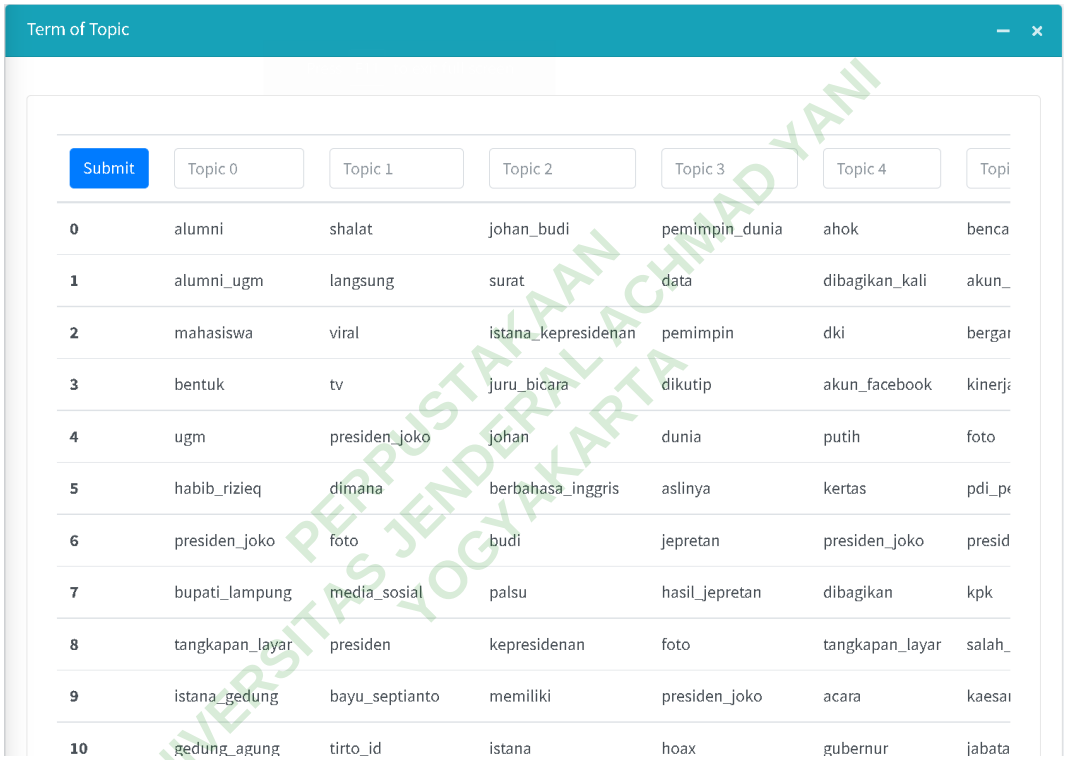
Hasil akhir dari penelitian ini berupa *website* yang menampilkan grafik relasi antar topik serta atribut dari tiap topik seperti nilai *degree centrality*, *betweenness centrality*, *closeness centrality*, dan *term of topic* pada tiap topik sehingga user dapat menganalisis apakah setiap berita *hoax* yang beredar saling terkait atau tidak serta dapat mengetahui tujuan adanya jaringan berita *hoax* yang saling terhubung.

Tampilan awal dapat dilihat pada Gambar 4.13 dimana pengguna akan disediakan *form input* pencarian menggunakan kata kunci tertentu. Untuk menampilkan grafik relasi pengguna diharuskan memasukkan kata kunci tertentu sehingga sistem akan melakukan pencarian data pada *database* yang kemudian akan ditampilkan.



Gambar 4.13 Tampilan Awal Fitur *Relation*

Setelah pengguna melakukan pencarian dengan kata kunci tertentu sistem akan menampilkan tabel *term of topic* seperti pada Gambar 4.14 untuk dilakukan pelabelan manual pada topik. Setelah user melakukan pelabelan pada topik, sistem akan membentuk grafik SNA kemudian ditampilkan. Kolom penamaan label yang disediakan berjumlah sepuluh kolom, namun user tidak diharuskan untuk mengisi semua kolom jika pada kolom *term of topic* tertentu memiliki kumpulan kata yang hampir sama.



Gambar 4.14 Tampilan Tabel Term of Topic

Berikut Gambar 4.15 merupakan tampilan grafik SNA yang telah dibentuk setelah user melakukan pelabelan manual. Terdapat tiga grafik yang ditampilkan yaitu grafik *degree centrality*, *betweenness centrality*, dan *closeness centrality*. Jika diklik, node akan menampilkan atribut berupa nama topik, nilai *degree centrality*, *betweenness centrality*, dan *closeness centrality*.

Gambar 4.15 Tampilan Grafik SNA