Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 1, No. 1, Mei 2023, hlm 47-55

Topic Modeling dan Social Network Analysis digunakan untuk mencari keterkaitan topik pada Tweet Pembahasan Saham

Muhammad Ilham Zulqarnain [a,1,] , Puji Winar Cahyo [a,2,*]

[a] Program Studi Informatika Universitas Jenderal Achmad Yani Yogyakarta, Jl. Siliwangi Ringroad Barat, Sleman and 55293, Indonesia
[1] ilham.mancity13@gmail.com; [2] pwcahyo@gmail.com*

* corresponding author

ABSTRACT

Pada tahun 2020, jumlah orang yang melakukan *trading* di Indonesia mengalami peningkatan meskipun terjadi pandemic covid19. Dalam hitungan jumlah investor pada tahun 2020 mencapai 3.5 juta investor sedangkan pada tahun 2021 meningkat menjadi 7.5 investor. Melalui adanya peningkatan ini, maka jumlah posting tentang saham dan tutorial mengenai *trading* saham di media sosial meningkat cukup drastis. Maka penelitian ini mencoba untuk melakukan analisis keterkaitan topik pembicaraan saham pada sosial media Twitter dengan menggunakan integrasi *topic modelling* dan Social Network Analysis (SNA). Proses pembagian topik ideal menggunakan *coherence measurement* menentukan sebanyak 5 topik ideal. Melalui lima topik yang dihasilkan dari *topic modelling* tersebut kemudian dilakukan analisis menggunakan SNA sehingga menghasilkan nilai *degree centrality, betweeness centrality*, dan *closeness centrality* yang sama pada setiap topik. Nilai tersebut diantaranya: 4 untuk nilai *degree centrality*, 0.4 untuk *betweeness centrality* dan 1 *untuk closeness centrality*. Melalui hasil tersebut maka perlunya evaluasi dalam pembentukan SNA dengan menggunakan *topic modeling*. Evaluasi tersebut salah satunya bisa dilakukan melalui identifikasi pada *tweet* yang memiliki kesamaan pembahasan meskipun dengan penulisan redaksi yang berbeda, atau dapat dilakukan dengan cara menambah variasi data dengan cara memperlama waktu pengambilan.

ARTICLE INFO

Article history
Received: 9 Maret 2023
Revised: 5 April 2023
Accepted: 10 Mei 2023

Keywords
topic modeling
lda
sna
twitter
saham

1. Introduction

Pada tahun 2020, jumlah orang yang melakukan *trading* di Indonesia mengalami peningkatan meskipun terjadi pandemic covid19. Peningkatan berada pada angka 2,8 persen dari tahun 2019 menuju tahun 2020, sedangkan pada tahun 2020 menuju tahun 2021 mengalami peningkatan cukup tinggi yaitu pada 36,78 persen. Dalam hitungan jumlah investor pada tahun 2020 mencapai 3.5 juta investor sedangkan pada tahun 2021 meningkat menjadi 7.5 investor [1]. Melalui adanya peningkatan ini, maka aplikasi *trading* untuk perdagangan saham meningkat cukup drastis dibersamai dengan jumlah posting tentang

saham dan tutorial mengenai *trading* saham di media sosial. Sementara itu *Trading* adalah proses negosiasi harga antara pembeli dan penjual sampai tercapai kesepakatan akhir antara pembeli dan penjual, atau *trading* dapat dipahami sebagai bentuk usaha dalam bentuk kegiatan usaha jual sama seperti pembeli dan penjual di pasar buah atau supermarket. Jika di pasar buah yang diperjualbelikan berupa buah- buahan, di dalam *trading* yang diperjualbelikan adalah saham, valuta asing, komoditi, dan lain-lain [2].

Seiring bertambahnya jumlah posting pada media sosial terutama Twitter mengenai saham maka dapat dilakukan analisis topik terkait hasil diskusi pada media sosial tersebut. Untuk mendapatkan topik pada suatu diskusi dapat menggunakan metode pemodelan topik salah satunya menggunakan algoritma *Latent Dirichlet Allocation* (LDA), melalui LDA tersebut topik dapat dilakukan ekstraksi berdasar tingkat kedekatan dan saling keterkaitan pada topik yang dihasilkan [3]. Sedangkan untuk keterkaitan lebih detail pengaruh topik dengan masing-masing aktor dapat menggunakan metode *Social Network Analysis* (SNA) [4]. Penggunaan SNA dapat diintegrasikan dengan LDA sehingga akan menghasilkan keterkaitan antar topik dan keterkaitan antar aktor melalui pencarian aktor yang berperan sebagai sentralitas suatu topik [5]. Dari beberapa metode analisis topik diskusi yang telah dipelajari tersebut maka penelitian ini mencoba untuk melakukan analisis diskusi pembicaraan saham pada platform Twitter dengan menggunakan integrasi *topic modeling* dan SNA untuk mengetahui keterkaitan topik dengan membentuk grafik analisis berbentuk *Social Network* yang dihasilkan dari *Topic Modeling* menggunakan LDA.

2. Method

Penelitian ini bertujuan untuk mencari keterkaitan topik melalui hubungan antar kata *(term)* yang membentuk topik tersebut. Proses untuk mendapatkan hubungan *term* pada setiap topik melalui beberapa tahapan diantaranya adalah penggunaan metode *Topic Modeling* dan *Social Network Analysis* secara tahapan dapat dilihat pada Fig. 1.

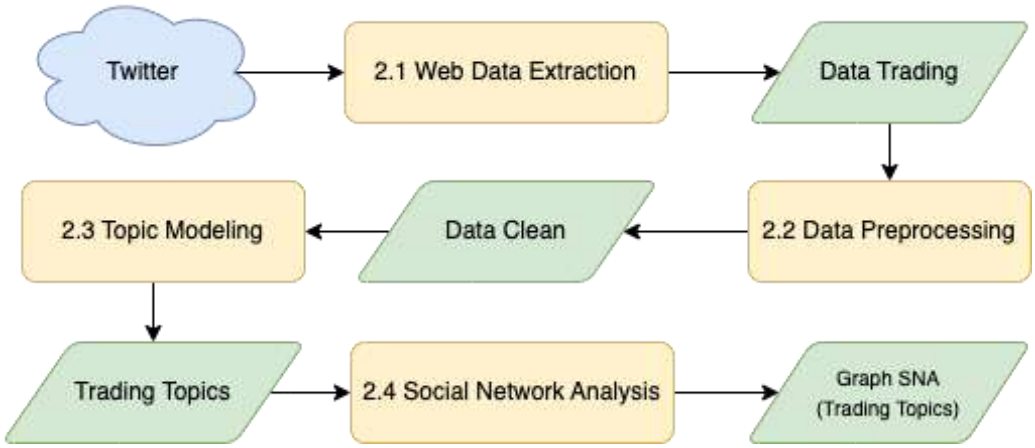

Fig. 1. Arsitektur Penelitian

Melalui Fig. 1 dapat dijelaskan bahwa data diambil dari Platform Twitter yang kemudian dilakukan ekstraksi data sehingga menghasilkan data pembahasan trading. Data yang telah berhasil diekstraksi kemudian dibersihkan untuk dapat diproses kedalam *topic modelling.* Hasil dari *topic modeling* adalah data *term* yang membentuk suatu topik sehingga setiap topik dapat saling memiliki keterkaitan atau bahkan tidak memiliki keterkaitan sama sekali. Proses identifikasi keterkaitan tersebut dilakukan pada tahapan *social network analysis* menggunakan penghitungan kedekatan pada term masing-masing kata pembentuk topik tersebut. Secara detail dapat dibahas sebagai berikut:

2.1. Web Data Extraction

Proses *web data extraction* merupakan cara untuk mendapatkan nilai informasi melalui pengambilan informasi yang diperlukan pada suatu halaman web [6]. Upaya pengambilan data pada penelitian ini menggunakan *library* snscraper dengan *timeline* waktu sejak Januari 2022 sampai pada Mei 2022. Data *tweet* yang diambil menggunakan kata kunci "saham".

2.2. Data Preprocessing

Hasil data *tweet* dari pengambilan menggunakan *library* snscraper yang sudah terkumpul diteruskan menuju tahap *text cleaning*, *casefolding*, *tokenizing*, *stopwords removal* dan *frase filtering*. Secara tahapan dapat dilihat sebagai berikut:

- Text Cleaning: bertujuan untuk menghilangkan karakter yang tidak dibutuhkan pada proses pembuatan model. Contohnya karakter seperti tanda baca.
- Casefolding: bertujuan untuk mengubah format penulisan huruf menjadi cetak huruf kecil semuanya.
- Tokenizing: bertujuan untuk memotong kalimat menjadi beberapa bagian kata yang terpisah dan pada saat yang sama menghilangkan karakter tertentu seperti tanda baca, angka, dan karakter non-abjad. Karakter tersebut dianggap sebagai pembatas kata dan tidak memengaruhi pemrosesan kata [7].
- Stopword Removal: bertujuan untuk menghilangkan kata umum yang tidak banyak memengaruhi pemrosesan kata. Sebagai contoh *stopword* pada bahasa Indonesia diantaranya : "yang", "ini", "dari, "ke" [7].
- Frase Filtering: Pada penelitian ini dilakukan identifikasi frasa yang mengandung 2 kata misalnya "media sosial" dan frasa yang mengandung 3 kata misalnya "sekolah menengah atas", oleh karena itu diperlukan penyaringan bigram dan trigram

2.3. Topic Modeling

Setelah data selesai dilakukan *preprocessing*, kemudian dilanjutkan menuju pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA). Dua masukan utama untuk pemodelan topik menggunakan LDA adalah *dictionary* dan *corpus*. Preprocessing akan membentuk *dictionary* berupa pemetaan kata kedalam id secara unik. Kemudian dilanjutkan pada pembentukan *corpus* yang berisi *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) sesuai dengan (1).

$$W_{t,d} = tf_{t,d} \times log \frac{N}{df_t} \quad (1)$$

Persamaan (1) digunakan untuk menghitung TF-IDF dengan $W_{t,d}$ merupakan bobot dokumen *ke-i* terhadap kata *ke-j*. Sedangkan $tf_{t,d}$ merupakan frekuensi kata ke *t* pada dokumen *d*, sementara itu df_t merupakan banyaknya dokumen yang mengandung kata ke-*t* dan *N* merupakan jumlah dokumen.

Setelah *dictionary* dan *corpus* dihasilkan kemudian dibentuk model LDA dan dilakukan evaluasi dengan metode *coherence measurement* untuk memperoleh model yang optimal. Model visualisasi pemodelan topik menggunakan LDA dapat dilihat pada Fig. 2.

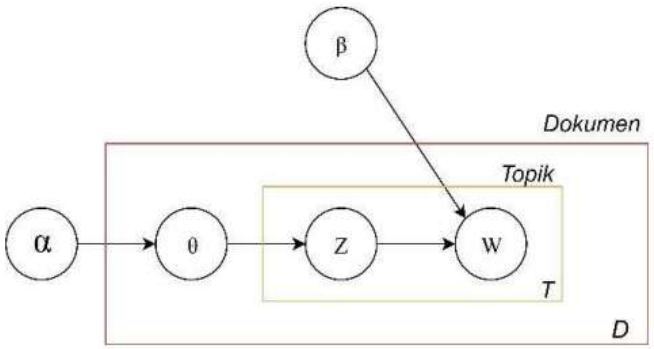

Fig. 2. Visualisasi Pemodelan Topik dengan menggunakan LDA

Terdapat tiga tingkatan pada LDA Modeling seperti yang ditunjukan pada Fig. 2 Pada *D = [d1, d2, d3, ..., dq]* dokumen terdapat parameter *α* dan *β* yang merupakan parameter distribusi topik *(T = [t1, t2, t3, ..., t j])* pada tingkatan *corpus*. Paramater α digunakan untuk menentukan distribusi topik dalam dokumen, semakin banyak topik yang dibahas di suatu dokumen, maka semakin tinggi nilai *alpha* dalam suatu dokumen. Parameter *β* berfungsi untuk menentukan distribusi di dalam topik, semakin sedikit kata-kata yang terdapat dalam topik atau mengandung kata-kata yang lebih spesifik, maka nilai beta semakin kecil, begitu sebaliknya. Variabel *θD* adalah variabel yang berada di tingkat dokumen *(D)*. Variabel *θ* merepresentasikan distribusi topik untuk dokumen tertentu. Semakin tinggi nilai *θ*, maka semakin banyak topik yang ada di dalam dokumen, sedangkan semakin kecil nilai *θ* , maka dapat dikatakan dokumen tersebut semakin spesifik pada topik tertentu. Variabel *Z* dan *Wm* adalah variabel tingkat kata. Variabel *Z* merepresentasikan topik dari kata tertentu pada sebuah dokumen sedangkan variabel *W* merepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen [8].

2.4. Social Network Analysis

Topik yang telah berhasil diekstraksi pada tahap *topic modeling* dijadikan sebagai komponen utama pembentuk relasi pada model *Social Network Analysis* (SNA). Relasi pada SNA diatur oleh *node* yang membentuk kedekatan antar topik. Kedekatan antar topik yang digunakan berdasar pada *term* yang membentuk topik. Hasil *term* dari topik LDA umumnya akan saling tumpang tindih dengan term pada topik ang lain. Hal ini kemudian dijadikan dasar untuk membentuk *social network*. *Term* tumpang tindih tersebut dapat dicontohkan sesuai pada Fig. 1. yaitu terdapat 2 topik: topik 1 dan topik 2 kemudian masing-masing topik memiliki keterkaitan *term* yang sama diantaranya: saham , beli, gue, dan banget. Maka dari hubungan keterkaitan antara *term* tersebut dijadikan sebagai dasar terbentuknya SNA.

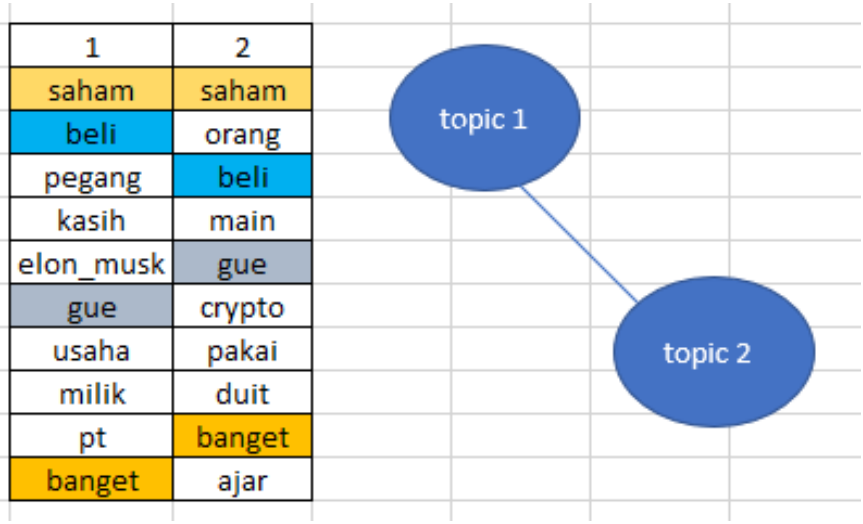

Fig. 3. Pembentukan Relasi SNA

3. Results and Discussion

3.1. Topic Modeling

Hasil dari pemodelan topik menggunakan LDA menghasilkan nilai *coherence score* paling tinggi yaitu dengan penentuan 5 topik yang ideal kemudian divisualisasikan pada grafik intertopic: Distance Map. Contoh visualisasi dapat dilihat pada Fig. 4 dan Fig. 5 merupakan visualisasi keterkaitan antara term pada topik 1 dan topik 2.

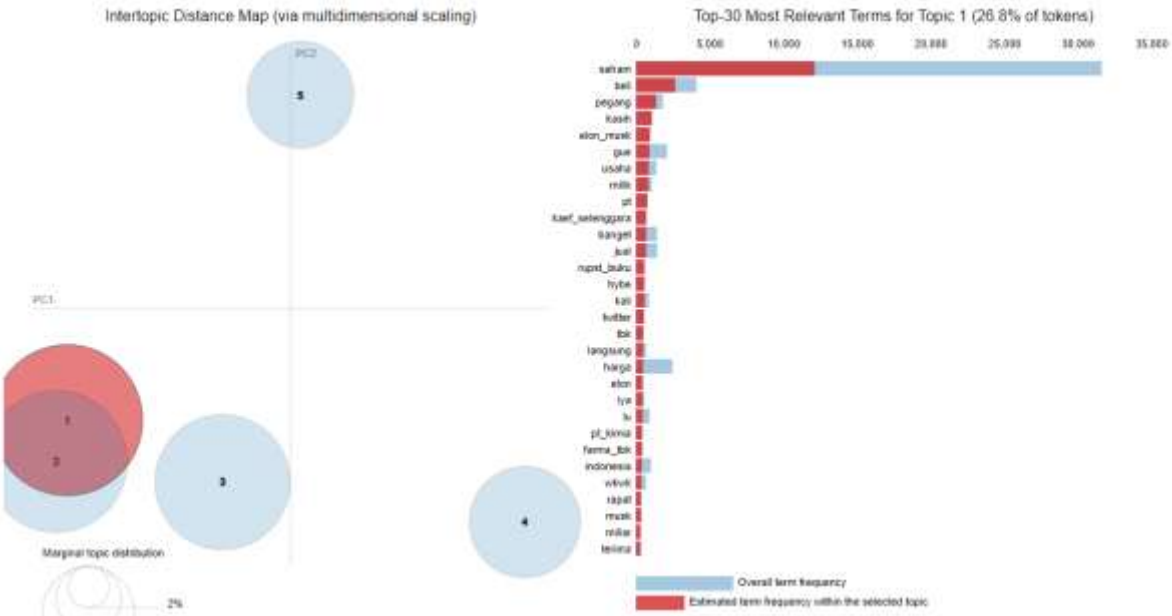

Fig. 4. Intertopic distance map topik 1

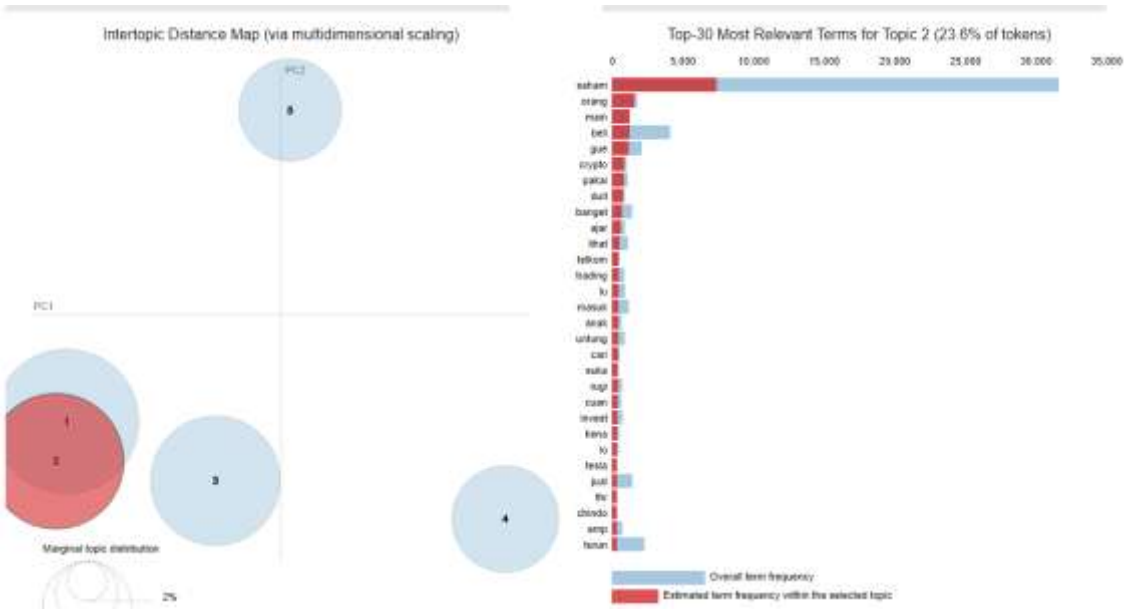

Fig. 5. Intertopic distance map topik 2

Dapat dilihat pada Fig. 4 dan Fig. 5 merupakan grafik *term* yang membentuk suatu topik. Apabila dilihat dari keseluruhan topik maka dapat diketahui masing-masing *term* yang membentuk setiap topik dapat dilihat pada Table 1.

Table 1. Term penyusun topik

Topik 0	Topik 1	Topik 2	Topik 3	Topik 4
ihsg	saham	saham	saham	saham
saham	beli	orang	goto	investasi
kuat	pegang	beli	rp	pasar
indeks	kasih	main	mei	turun
tutup	elon_musk	gue	dividen	harga
buka	gue	crypto	triliun	uang
persen	usaha	pakai	pegang	reksadana
dagang	milik	duit	bank	suku_bunga
bursa	pt	banget	selenggara	wall_street
merah	banget	ajar	gojek_tokopedia	pilih

Seperti yang ditunjukan pada Table 1, setiap topik mempunyai kata (*term*) yang menyusun. Hasil dari *term* tersebut kemudian dianalisis menurut isi pembicaraan dan keterkaitan pada keseluruhan *term* pada topik tersebut sehingga menghasilkan inti pembicaraan yang dapat disimpulkan sesuai Table 2.

Table 2. Inti Pembicaraan

Topik	Kata Penyusun Topik (term)	Inti Pembicaraan
Topik 0	"ihsg", "saham", "kuat", "indeks", "tutup", "buka", "persen", "dagang", "bursa", "merah"	Menguatnya saham IHSG
Topik 1	"saham", "beli", "pegang", "kasih", "elon_musk", "gue", "usaha", "milik", "pt", "banget"	Saham yang dimiliki oleh Elon Musk
Topik 2	"saham", "main", "pakai", "ajar" "orang", "gue", "duit", "beli", "crypto", "banget",	Tentang cryptocurrency
Topik 3	"saham", "goto", "rp", "mei", "dividen", "triliun", "pegang", "bank", "selenggara", "gojek_tokopedia"	Saham Gojek Tokopedia(GoTo).
Topik 4	"saham", "investasi", "pasar", "turun", "harga", "uang", "reksadana", "suku_bunga", "wall_street", "pilih"	Pasar saham yang turun.

3.2. Social Network Analysis

Dari hasil penyusunan topik pada Table 1 kemudian dilanjutkan pencarian relasi topik antara suatu topik dengan topik yang lainnya. Relasi antar topik berdasarkan *term* dapat dilihat pada Table 3.

Table 3. Relasi topik berdasarkan term

	relation	from	to	term	weigth
0	topik 0 - topik 1	topik 0	topik 1	Saham	1
1	topik 0 - topik 2	topik 0	topik 2	Saham	1
2	topik 0 - topik 3	topik 0	topik 3	Saham	1
3	topik 0 - topik 4	topik 0	topik 4	Saham	1
4	topik 1 - topik 2	topik 1	topik 2	Saham	4
5	topik 1 - topik 2	topik 1	topik 2	Beli	
6	topik 1 - topik 2	topik 1	topik 2	Gue	
7	topik 1 - topik 2	topik 1	topik 2	banget	
8	topik 1 - topik 3	topik 1	topik 3	Saham	2
9	topik 1 - topik 3	topik 1	topik 3	Pegang	
10	topik 1 - topik 4	topik 1	topik 4	Saham	1
11	topik 2 - topik 3	topik 2	topik 3	Saham	1
12	topik 2 - topik 4	topik 2	topik 4	Saham	1
13	topik 3 - topik 4	topik 3	topik 4	Saham	1

Table 3 menjelaskan bahwa penghitungan *weight* menggunakan banyaknya relasi antar *term* pada topik yang sejenis, seperti contoh topik 1 memiliki 4 relasi *term* pada topik 2, sehingga bobot untuk relasi kedua topik tersebut adalah 4. Contoh lain adalah topik *1* mempunyai relasi topik dengan topik *3* sebanyak 2 kali, sehingga dinyatakan bobot topik *1* memiliki relasi dengan topik *3* adalah 2. Hasil bobot tersebut kemudian digunakan sebagai pembentukan relasi *node* dengan menggunakan *library* NetworkX. Hasil pembentukan SNA melalui NetworkX kemudian diambil nilai *degree*, *between* dan *closeness centrality* sehingga membentuk hasil seperti pada Table 4.

Table 4. Nilai centrality node

Node	Score		
	Degree Centrality	Betweeness Centrality	Closeness Centrality
Topik 0	4	0,4	1

Topik 1	4	0,4	1
Topik 2	4	0,4	1
Topik 3	4	0,4	1
Topik 4	4	0,4	1

3.3. Analisis Hasil Relasi Topik

Sesuai hasil SNA yang dapat dilihat pada Table 4 maka ke-5 topik memiliki *degree centrality*, *betweeness centrality* dan *closeness centrality* yang sama. Contoh kasus topik menguatnya Saham IHSG, memiliki nilai *degree centrality* 4 yang diartikan bahwa topik menguatnya Saham IHSG memiliki relasi dengan ke 4 topik lainnya, nilai dari *betweeness centrality* 0,4 dapat diartikan setiap topik memiliki jarak kedekatan yang sama satu sama lain. Nilai dari clossness centrality 1 yang diartikan bahwa menguatnya Saham IHSG dekat dengan topik yang lainnya salah satunya dengan Saham Gojek Tokopedia (GoTo) dikarenakan Saham IHSG adalah salah satu *index* harga saham gabungan Indonesia.

Menguatnya saham IHSG salah satunya dipengaruhi karena naiknya harga saham yang ada di dalam negeri. Seperti yang dimuat pada media. mediaindonesia.com pada artikel yang berjudul "IHSG Menguat karena Investor Prediksi Inflasi AS" didalam artikel tersebut dimuatnya beberapa Saham yang naik di indeks LQ45 seperti saham GOTO sehingga membuat harga saham IHSG menguat pada tanggal 14 Juli 2022 [9].

4. Conlusion

Proses penentuan topik yang ideal pada *topic modelling* menggunakan *coherence measurement* menghasilkan pembagian topik ideal sebanyak 5 topik. Melalui lima topik yang dihasilkan dari *topic modelling* tersebut kemudian dilakukan analisis menggunakan SNA sehingga menghasilkan nilai *degree centrality*, *betweeness centrality*, dan *closeness centrality* yang sama pada setiap topik. Nilai tersebut diantaranya: 4 untuk nilai *degree centrality*, 0.4 untuk *betweeness centrality* dan 1 untuk *closeness centrality*. Melalui hasil tersebut maka perlunya evaluasi dalam pembentukan SNA dengan menggunakan LDA. Evaluasi tersebut salah satunya bisa dilakukan melalui identifikasi pada *tweet* yang memiliki kesamaan pembahasan meskipun dengan penulisan redaksi yang berbeda, atau dapat dilakukan dengan cara menambah variasi data dengan memperlama waktu pengambilan.

References

[1] M. H. R. Lubis, "Analisis Pertumbuhan Investor Ritel pada Masa Pandemi dan Implikasi Pajak Penghasilan Final Atas Penjualan Saham di Bursa," J. Pajak Indones., vol. 6, no. 2, pp. 245–264, 2022.

[2] I. J. Tjendra, A. A. S, J. Cahyadi, and J. Siwalankerto, "Perancangan Buku Panduan Dasar Trading Untuk Pemula," J. DKV Adiwarna, Univ. Kristen Petra, vol. 1, no. 8, pp. 1–9, 2016.

[3] P. W. Cahyo, M. Habibi, A. Priadana, and A. B. Saputra, "Analysis of Popular Hashtags on Instagram Account The Ministry of Health," in Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020), 2021, vol. 34, no. Ahms 2020, pp. 270–273.

[4] C. Jovanica, D. D. Rahmintanigrum, and H. A. Nuradni, "Analisis Pengaruh Aktor pada Tagar # roketchina di Media Sosial Twitter Menggunakan Social Network Analysis (SNA)," vol. 10, no. 1, pp. 43–56, 2022.

[5] M. Iqbal and S. Pramana, "Penerapan Social Network Analysis dan Latent Dirichlet Allocation untuk Pemetaan Publikasi Penelitian Dosen Politeknik Statistika STIS," J. Apl. Stat. Komputasi Stat., vol. 13, no. 2, pp. 1–14, 2021.

[6] P. W. Cahyo, K. Kusumaningtyas, and U. S. Aesyi, "A User Recommendation Model for Answering Questions on Brainly Platform," J. INFOTEL, vol. 13, no. 1, pp. 7–12, 2021.

[7] P. W. Cahyo and M. Habibi, "Entity Profiling to Identify Actor Involvement in Topics of Social Media Content," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 14, no. 4, pp. 417–428, 2020.

[8] I. M. Kusnanta, B. Putra, and P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)," J. Tek. ITS, vol. 6, no. 2, pp. 4–9, 2017.

[9] A. N. Gumay, "IHSG Menguat karena Investor Prediksi Inflasi AS," mediaindonesia.com, 2022.