HASIL CEK_IMPLEMENTASI ALGORITMA SPECTRAL CLUSTERING UNTUK ANALISIS SENTIMEN

by Sugiyarto Cek_implementasi Algoritma Spectral Clustering Unt

Submission date: 16-Nov-2020 08:21AM (UTC+0700)
Submission ID: 1447038460
File name: 1229-2961-1-RV_1.pdf (352.52K)
Word count: 2373
Character count: 14339

IMPLEMENTASI *ALGORITMA SPECTRAL CLUSTERING* UNTUK ANALISIS SENTIMEN

Qonitat rohmah [1)], Sugiyarto Surono[2)]

[1)] *Universitas Ahmad Dahlan, Jl. Kapas 9, Semaki, Umbulharjo, Yogyakarta; qonitat1600015013@webmail.uad.ac.id*

[2)] *Universitas Ahmad Dahlan, Jl. Kapas 9, Semaki, Umbulharjo, Yogyakarta; sugiyarto@math.uad.ac.id*

Abstract

Data mining is a study that collects, cleans, processes, analyzes and benefits from data. One of the techniques known in data mining is the Spectral Clustering technique. Spectral clustering is a technique that follows the Connectivity approach, where this method classifies points that are connected or directly adjacent. The purpose of this study is to determine the level of public sentiment towards the 2017 Jakarta Pilkada using the Spectral Clustering method. The test data was obtained from the scraping process on Twitter from October 1, 2016 to April 20, 2017. In this study, input data consisting of tweet data and output data were used in the form of sentiments that have been clustered into 3, namely positive, negative and neutral. Obtained 4571 negative data, 1899 neutral data and 1588 positive data. with the highest possible win rate in the first round on Ahok.

***Keyword**:* Sentiment Analysis, Pilkada Jakarta 2017, Spectral Clustering, Clustering, Data mining

1. Pendahuluan

Pada saat ini Data semakin berkembang dalam segala macam bidang ilmu, Data yang semakin banyak mengakibatkan terjadinya banjir data, sehingga di butuhkan Konsep Data mining. Data mining adalah studi yang mengumpulkan, membersihkan, mengolah, menganalisis, dan memperoleh manfaat dari data (Angarwal, Charu c, 2015). Data mining adalah proses untuk menggali informasi yang diperlukan secara otomatis dalam repositori data yang sangat besar. Teknik pengambilan data memungkinkan untuk menjelajahi database besar guna menemukan pola baru yang mungkin tidak diketahui (Van Dongen, 2000)

Ada beberapa Teknik yang digunakan pada Data Mining, *Clustering* adalah salah satu metode Data Mining *unsupervised* yang mengelompokan data menurut karakteristik yang sama pada masing masing data, dalam jurnal Retno Tri Wulandari metode *cluster* merupakan suatu metode yang mencari dan

mengelompokan data sesuai dengan kemiripan karakteristik antar data satu dengan data lainnya.

Terdapat banyak *Algoritma Clustering* yang sudah dikenal secara umum, namun *Algoritma Clustering* yang sering digunakan adalah *K-Means*. Alasannya karena metode *K-mean* mudah dalam implementasinya serta memiliki waktu komputasi yang cepat. Metode ini memiliki beberapa kelemahan salah satunya karena hanya mempertimbangkan jarak data ke masing masing *centroid* pada setiap *cluster* maka ada beberapa masalah pada analisis persebarannya serta bergantung pada inisial *centroidnya*. Metode *Spectrla Clustering* merupakan metode yang dapat diusulkan guna memperbaiki kelemahan *K-mean* (Trivedi, S., A. Pardos, Z., & N. Sar, G. 2008). *Spectral Clustering*, pengelompokkan didasarkan atas kesamaan antara setiap data. Kesamaan tersebut dilihat dari keterkaitan antara setiap data. Pada *Spectral Clustering* akan dibentuk sebuah graf dari data yang ada. Di mana simpul dari graf tersebut merupakan titik pada data. sisi berupa hubungan antar data yang biasanya bernilai jarak dari dua *record* yang berhubungan (Trivedi, S., A. Pardos, Z., & N. Sar, G. 2008).

Pelaksanaan Pilkada merupakan wujud dari demokrasi tidak langsung (*indirect democracy*), yang bertujuan agar Kepala Daerah mengeluarkan kebijaksanaan daerah atas nama rakyat sehingga kepala daerah harus dipilih secara langsung oleh rakyatnya sendiri melalui Pilkada (Marijan, 2010). Dengan kata lain Pilkada merupakan suatu sarana pemberian mandat dari rakyat kepada Kepala Daerah dengan Harapan agar Kepala Daerah yang terpilih dapat bertindak sesuai kepentingan rakyat.

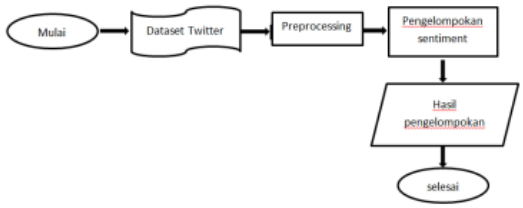
Pada sebelum maupun pada saat pilkada sedang berlangsung banyak opini opini masyarakat terhadap Kandidat pilkada, Twitter merupakan salah satu media sosial yang banyak digunakan masyarakat untuk menyampaikan opini opini mereka terhadap pelaksanaan pilkada. Opini pada twitter belum dapat diidentifikasikan apakah opini tersebut bersifat positif, negative atau netral. Informasi yang diterima pada media sosial twitter apabila tidak dipelajari lebih dalam tentang sumber data dan kebenaran informasi cenderung akan menjadi berita palsu atau "kampanye Hitam'. Agar opini tersebut dapat bermanfaat maka diperlukan berbagai proses sehingga diperoleh informasi yang penting melalui analisis sentimen untuk menentukan strategi kampanye yang tepat untuk Pilkada.

Analisis sentiment atau dikenal juga *opinion mining* (penambangan data) adalah proses mencari informasi dari suatu opini atau pendapat dari suatau data untuk topic tertenti (H. Kaur, V. Mangat dan N, 20017). Analisis sentimen dilakukan dengan tujuan agar mengetahui pendapat seseorang terhadap suatu peristiwa atau masalah, cenderung positif, negative atau netral. Teknik yang dapat digunakan untuk analisis sentiment adalah *text mining*.

Data yang akan ditambang untuk Proses *clustering* adalah pesan dan komentar-komentar yang sudah diposting pada media sosial Twitter. Postingan yang pernah ditulis pada twitter kemudian di *per-processing* sehingga menjadi data yang dapat dikelompokan menurut jenis pesan. Variabel-variabel yang dibutuhkan pada penelitian ini adalah *tweet,* tanggal dan kandidat, yang akan diproses menggunakan *text mining* yang akan di *cluster* menjadi 3 kelompok yaitu postingan bersifat positif, netral, atau negatif. Dalam penelitian ini saya tertarik untuk menganalisis pilkada 2017 menggunakan Twiter untuk mengelompokan data menggunakan *Spectral Clustering*.

2. Metodologi Penelitian

Penelitian ini melalui beberapa tahap seperti pada gambar 1 yaitu pengumpulan data, pra-pemrosesan data, pengembangan sistem menggunakan metode *Spectral Clustering*, serta evaluasi sistem.



Gambar 1. *Flowchart* penelitian

2.1 *Text Data*

Text mining adalah salah satu teknologi yang digunakan untuk data yang semakin bertambah banyak sehingga teks yang tidak sistematis dapat dianalisis (Francis dan Flynn, 2010). *Text Mining* dapat menjadi solusi dari Data Mining yaitu cara untuk mengetahui informasi dari sebuah data (Feldman dan Sanger, 2007). Konsen data mining hampir mirip dengan Konsep Data mining hanya saja *Text mining* dapat bekerja pada teks yang tidak terstuktur atau semi terstruktur seperti *E-mail*.

Kurniawan, et al.(2012) menjelaskan terdapat beberapa langkah yang dapat dilakukan dalam *text mining* :

1. *Text Pre-processing*

Pada tahap ini data kotor dibersihkan, data yang tidak diperlukan dibuang dan dirapikan sehingga dapat diolah. Tindakan yang dilakukan pada tahap ini :

- *To lower case*, proses ini mengubah huruf besar menjadi huruf kecil
- *Tokenizing*, proses ini menguraikan yang semula kalimat menjadi kata kata.

2. *Feature Selection*

Pada tahap ini tindakan yang dilakukan adalah:

- *Stopword (stopword removal)* , pada proses ini setiap kata akan diperiksa, jika didalamnya terdapat kata kata yang tidak sesuai dengan kamus maka akn dibuang, kamus yang digunakan pada penilitian ini adalag *stopwordID.txt*.
- *stemming* adalah Proses yang mengubah *text* menjadi kata dasar.

2.2 Proses *Clustering* Menggunakan *Spectral Clustering*

Metode *cluster* adalah metode untuk mengkelompokan data yang memiliki kemiripan karakteristik antar data pada dokumen. . *Cluster* adalah metode penambangan data tanpa pengawasan (H. Kaur, V. Mangat dan N. 2017). Analisis Clustering megelompokan berdasarkan informasi yang tersedia pada data, analisis ini bersifat *unsupervised* sehingga semakin besar (homogenitas) dalam satu kelompok dan semakin besar perbedaan pada setiap kelompok maka semakin jelas pengelompokannya.

Pengelompokan *spektral* adalah pengelompokan multi-arah teknik yang menggunakan vektor *eigen* dari sebuah matriks afinitas diinduksi dari data untuk melakukan pengelompokan. Pengelompokan *Spektral* adalah teknik yang populer dikarenakan kesederhanaan, intuisi dan kemampuan untuk pengelompokan titik data yang tidak dapat dipisahkan secara linear. Selain itu juga dapat memberikan hasil perhitungan yang sebanding atau lebih baik dibandingkan metode metode lainnya (Luxburg,2007).

Pengelompokan Spektral adalah teknik yang populer karena kesederhanaan, intuisi, dan kemampuannya untuk mengelompokkan titik data yang tidak dapat diakses secara linier. Disamping itu juga dapat memberikan hasil perhitungan yang sebanding atau lebih baik dari metode lainnya. (Luxburg, U. V. 2007). Teknik Spectral Clustering menggunakan spektrum

(eigenvalues) dari matriks Kesamaan untuk melakukan reduksi dimensional sebelum pengelompokan dalam dimensi yang lebih sedikit. Matriks Kesamaan dapat didefinisikan sebagai matriks simetris A dimana $A_{ij} \geq 0$ menunjukkan atau afinitas antara titik x_i dan x_j. Pendekatan umum Pengelompokan Khusus adalah dengan menggunakan metode pengelompokan standar (seperti k-mean) pada vektor eigen yang relevan dari matriks Laplacian A. Untuk menghitung, vektor eigen ini sering dihitung sebagai vektor yang dihitung sebagai nilai eigen vektor yang sesuai dengan beberapa nilai eigen terbesar dari fungsi Laplacian. Matriks Laplacian yang didefinisikan sebagai:

$$L = D - A$$

Dimana D adalah matrik diagonal :

$$D_{ii} = \sum_{j=1}^{n} A_{ij}$$

Teknik *Spectral Clustering* yang populer adalah algoritma pemotongan yang dinormalisasi atau algoritma Shi-Malik yang diperkenalkan oleh Jianbo Shi dan Jitendra Malik.Teknik ini membagi titik menjadi dua himpunan (B_1, B_2) berdasarkan vektor *eigen v* yang sesuai dengan nilai *eigen* terkecil dari matrik *Laplacian* yang didefinisikan pada persamaan :

$$L^{norm} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

Algoritma mengambil vektor *eigen* yang sesuai dengan nilai *eigen* terbesar dari matriks ketetanggaan yang dinormalisasi dengan berjalan acak $P = D - A$.

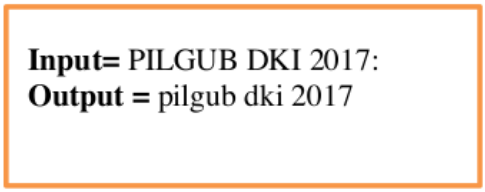
3. Hasil dan Pembahasan

Data yang digunakan diperoleh menggunakan tweetscraper menggunakan katakunci " PILKADA JAKARTA 2017" yang diambil dari tanggal 1 Oktober 2016 sampai 20 April 2017. Data yang diperoleh sebanyak 8058 dengan 3 kandidat yaitu Ahok, Anies Baswedan dan Agus Harimurty yudhoyono.

3.1 *Pre-processing* data

- *Case Folding*

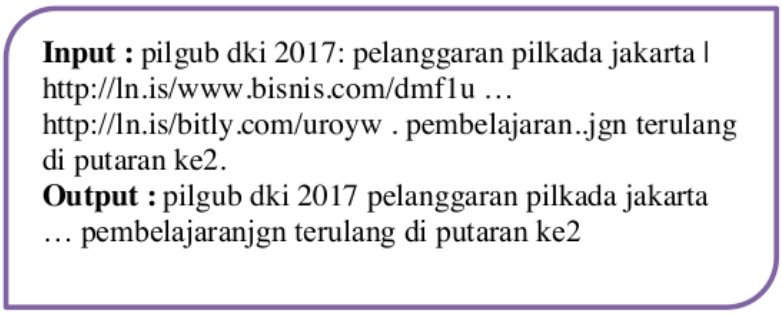
  Proses ini mengubah huruf besar menjadi huruf kecil. Hal ini bertujuan agar mempermudah dalam proses penganalisisan.



Gambar 2. Contoh *case folding*

- *Normalized*

  Pada data *tweet* terdapat karakter yang tidak berpengaruh pada proses selanjutnya, maka karakter tersebuh akan di hilangkan untuk mempermudah pada proses selanjtnya seperti, alamat *link* dan *username*



Gambar 3. Contoh *Normalized*

- *Stopword removal*

  Proses selanjutnya adalah *Stopword removal*, pada proses ini kata perkata akan diperiksa, jika di dalamnya terdapat kata yang tidak terdapat pada *stopword* maka akan di hapuskan. Di sini kami menggunakan file stopwordsID.txt sebagai acuan *stopword*.

Input =
pilgub dki 2017 pelanggaran pilkada jakarta URL … URL pembelajaranjgn terulang di putaran ke2
Output =
['pilgub', 'dki', 'pelanggaran', 'pilkada', 'jakarta', 'pembelajaranjgn', 'terulang', 'putaran', 'ke2']

Gambar 4. Contoh *Stopword Removal*

3.2 Mebuat Vektorisasi Matrik

$$T = \begin{pmatrix} 0.448 & 0.248 & \cdots & 0 \\ 0.503 & 0.750 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0.535 & 0.566 & \cdots & \cdots \end{pmatrix}$$

3.3 Membuat Matrik Similarity

Menggunakan Matrik T maka dapat di bentuk matrik similarity yang berukurun 8057 x 8057 :

$$affinity\_mat = \begin{pmatrix} 0 & 0.303 & \cdots & 0.338 \\ 0.303 & 0 & \cdots & 0.321 \\ \vdots & \vdots & \ddots & \vdots \\ 0.338 & 0.321 & \cdots & 0 \end{pmatrix}$$

3.4 Membentuk Matrik diagonal dari Matrik Similirity

Untuk sebuah simpul x_i, diberikan d_i menunjukan derajat dari simpul, maka dari $d_i = \sum_{j=i}^{n} a_{ij}$ diperoleh

$$diagonal\_\deg = \begin{pmatrix} 2491.284 & 0 & \cdots & 0 \\ 0 & 3233.736 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2874.708 \end{pmatrix}$$

3.5 Laplacian Matrik

Setelah memperoleh hasil dari matrik diagonal maka langkah selanjutnya membentuk matrik *laplacian*. Matrik tersebut di peroleh dari pengurangan matrik diagonal dengan matrik affinity.

$$L = diagonal\_dgr - affinity\_mat$$
$$= \begin{pmatrix} 2.491e+03 & -3.034e-01 & \cdots & -3.384e-01 \\ -3.034e-01 & 3.234e+03 & \cdots & -3.210e-01 \\ \vdots & \vdots & \ddots & \vdots \\ -3.384e-01 & 0.321 & \cdots & 2.875e+03 \end{pmatrix}$$

3.6 Normalized Laplacian matrik

Normalisasi Matrik *Laplacian* digunakan untuk mencari nilai *eigen* yang selanjutnya diperoleh vektor *eigen*. Diberikan bobot matrik ketetanggan A dari graph G, didefinisikan sebagai :

$$L^{norm} = diagonal\_inv * affinity\_mat * diagonal\_inv$$
$$= \begin{pmatrix} -1 & 1.069e-04 & \cdots & 1.265e-04 \\ 1.069e-04 & -1 & \cdots & 1.053e-04 \\ \vdots & \vdots & \ddots & \vdots \\ 1.265e-04 & 1.053e-04 & \cdots & -1 \end{pmatrix}$$

3.7 Eigenvalue dan *vector eigen*

Nilai *eigen* dapat diperoleh dengan rumus (A- λ I) x=0 dengan A= matriks *lapnorm* dan $k = 3$ maka :

$$det(A - \lambda I) = 0$$

Nilai Eigen yang terbentuk

$$\lambda_1 = -9.222e-01, \lambda_2 = -8.574e-01, \lambda_3 = 8.882e-16$$

Selanjutnya dari nilai eigen terbentuklah vektor eigen sebagi berikut :

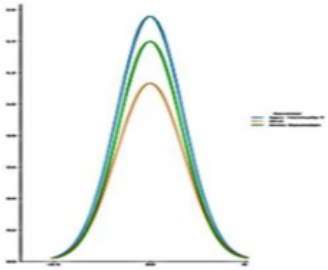
$$(A - \lambda I)x = 0$$

Maka vector eigen yang terbentuk sebagai berikut :

$$\begin{pmatrix} 0.011 & 0.002 & -0.010 \\ \vdots & \vdots & \vdots \\ 0.008 & 0.010 & 0.011 \end{pmatrix}$$

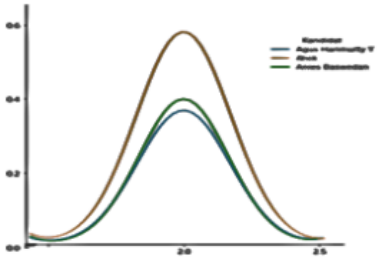
Berdasarkan vektor eigen kemudian dapat dikelompokkan menjadi 3 cluster yaitu 0, 1, 2 dimana 0 untuk negatif, 1 untuk netral dan 2 untuk positif. Dari metode *Spectral Clustering* diperoleh hasil 4571 *tweet* yang bersentimen negatif, 1899 *tweet* netral dan 1588 *tweet* positif. Pada Agus H mendapatkan sentimen Negatif sebanyak 1432 *tweet*, 381 *tweet* netral,

332 *tweet* positif. Untuk Ahok mendapatkan sentimen negatif sebanyak 1674 *tweet*, 863 *tweet* netral dan 846 *tweet* positif. Sedangkan Anies Baswedan mendapatkan 1465 *tweet* negatif, 655 *tweet* netral, dan 410 *tweet* positif.



Gambar 5. Grafik Sentimen Analisis Negatif

Dari Grafik Perbandingan tingkat sentimen negatif. Pada Sentimen negatif kandidat Agus Harimurty, tingkat komentar negatifnya paling tinggi 66.82%, diposisi kedua Anies baswedan sebanyak 57,9% dan terlihat Ahok memiliki sentimen negatif terendah sebanyak 49.48%.



Gambar 6. Grafik Sentimen Analisis positif

Di sini dapat dilihat bahwa grafik sentimen Positif tertinggi ada pada Ahok sebanyak 25% dan terendah adalah Agus 15,48%.

4. Kesimpulan dan Saran

Penelitian ini menghasilkan sebuah analisis sentimen terhadap PILKADA JAKARTA 2017 dengan menggunakan *spectral clustering*. Penelitian ini menggunanakan data *tweet* yang didapat melalui proses *scrapping* pada *software Jupyter Nootbook*. Dengan jumlah data sebanyak 8058 data yang diolah menggunakan *Jupyter Notebook*. *Tweet* tersebut di*cluster* menggunakan metode *Clustering* menjadi 3 *cluster* yaitu positif, negatif dan netral.

Didapatkan sebanyak 4571 *tweet* bernada Negatif, 1899 *tweet* bernada netral dan 1588 *tweet* bernada positif. *Tweet* bernada Negatif tertinggi pada kandidat Agus Harimurty Y, dan terendah pada Ahok, sebaliknya untuk *Tweet* bernada positif tertinggi adalah Ahok dan terendah adalah Agus Harimurty Y..

Pustaka

Aggarwal, C., Charu.2015.Data Mining:The Textbook.New York:Springer Cham Heidelberg.

Feldman,R.,S.,and Sanger,j.2007.The Text Mining handbook Advance Approaches In Analyzing Unstructured Data.New York:Cambridge University Press

Francis, L., and Flynn, M.2010. Text Mining Handbook . Casualty Actuarial Society

H. Kaur, V. Mangat dan N. 2017.A Survey of Sentiment Analysis Techniques,. 2017 International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud) (I SMAC), pp. 921 - 925.

Kurniawan, B., Effendi, S., and Sitompul, O,S. 2012. Klasifikasi Konten Berita Dengan Metode Text Mining. Dunia Teknologi Informasi Vol. 1, No. 1: Hal. 14-19.

Luxburg, U. V. 2007.A Tutorial on Spectral Clustering. Statistics and Computing.. 17(4): 395-416.

Marijan, Kacung. 2010. Sistem Politik Indonesia: Konsolidasi Demokrasi Pasca-Orde Baru. Jakarta: Penerbit Kencana Prenada Media Group

Retno Tri Vulandari. 2016. Pengelompokan Tingkat Keamanan Wilayah Jawa Tengah Berdasarkan Indeks Kejahatan Dan Jumlah Pos Keamanan Dengan Metode Klastering K-Means,vol 7 jilid 14.

Trivedi, S, A. Pardos, Z. N. Sar, G. 2008. Spectral Clustering in Educational Data Mining .

Van Dongen, S. 2000. Graph Clustering by Flow Simulation..PhD Thesis. University of Utrecht, The Netherlands

BIODATA PENULIS

A	**IDENTITAS PRIBADI**	
1	Nama Lengkap (beserta gelar)	Qonitat Rohmah Hidayati
2	Tempat Tanggal Lahir	28 Oktober 1998
10	Email	qonitat1600015013@webmail.uad.ac.id
4	No HP	081337044242
B	**IDENTITAS PROFESI**	
1	NIP	
2	NIDN/NIDK/NUPTK	
3	Asal Instansi	Universitas Ahmad Dahlan
4	Alamat Instansi	Jl.Ringroad Selatan, Kragilan, Tamanan, Kec Banguntapan,Bantul Daerah Istimewa Yogyakarta
5	Kab/Kota	Bantul/ Yogyakarta
6	Provinsi	Daerah Istimewa Yogyakarta
7	No Telp Instansi	(0274)511830
8	Lama mengajar	-
9	Pengalaman Seminar/Konferensi/Pertemuan Ilmiah	

Kegiatan	Sebagai

10	Publikasi Ilmiah	

Judul	Tahun

C	**IDENTITAS MAKALAH**	
1	Judul	Implementasi Algoritma Spectral Clustering untuk Sentimen Analisis
2	Penulis	Qonitat Rohmah Hidayati S.si dan H. Sugiyarto, M.Si., Ph.D,

HASIL CEK_IMPLEMENTASI ALGORITMA SPECTRAL CLUSTERING UNTUK ANALISIS SENTIMEN

ORIGINALITY REPORT

20%	19%	9%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eksplora.stikom-bali.ac.id Internet Source	3%
2	journal.trunojoyo.ac.id Internet Source	3%
3	Submitted to KYUNG HEE UNIVERSITY Student Paper	2%
4	docplayer.info Internet Source	2%
5	media.neliti.com Internet Source	1%
6	documents.tips Internet Source	1%
7	ejournal.gunadarma.ac.id Internet Source	1%
8	Submitted to Sim University Student Paper	1%
9	baa.uad.ac.id Internet Source	1%

10	core.ac.uk Internet Source	1%
11	vulandari.blogspot.com Internet Source	1%
12	D. J. Sherman. "Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes", Nucleic Acids Research, 01/01/2009 Publication	1%
13	journal.umpo.ac.id Internet Source	1%
14	www.univagora.ro Internet Source	1%
15	ojs.akbidylpp.ac.id Internet Source	1%
16	che.uad.ac.id Internet Source	<1%
17	www.jurnalkommas.com Internet Source	<1%
18	docplayer.net Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		