δELTΔ
Jurnal Ilmiah Pendidikan Matematika
p.ISSN: 2303 -3983 e.ISSN:2548-3994
Vol. 9 No. 1 Januari 2021 Hal 27 - 36.
DOI: http://dx.doi.org/10.31941/delta.v9i1.1229

IMPLEMENTASI *ALGORITMA SPECTRAL CLUSTERING* UNTUK ANALISIS SENTIMEN

Qonitat Rohmah [1)], Sugiyarto[2)]
[1,2)]*Universitas Ahmad Dahlan, Jl. Kapas 9, Semaki, Umbulharjo, Yogyakarta*
[1)]*qonitat1600015013@webmail.uad.ac.id* , [2)] *sugiyarto@math.uad.ac.id*

Received :
14/10/2020

Accepted :
26/11/2020

Published :
21/01/2021

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. In the second round with 2205 data, 604 positive tweets were obtained, 1123 neutral data, 479 negative data for negative tweets. In the second round Anies Baswedan received higher positive and lower negative responses than Candidate Ahok, so that the chances of winning against Anies Baswedan were higher than Ahok.

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

Abstrak

Penambangan data adalah studi yang mengumpulkan, membersihkan, memproses, menganalisis, dan memanfaatkan data. Salah satu teknik yang dikenal dalam data mining adalah teknik Spectral Clustering. Pengelompokan spektral merupakan teknik yang mengikuti pendekatan Konektivitas, dimana metode ini mengklasifikasikan titik-titik yang terhubung atau berbatasan langsung. Tujuan dari penelitian ini adalah untuk mengetahui tingkat sentimen masyarakat terhadap Pilkada Jakarta 2017 dengan menggunakan metode Spectral Clustering. Data pengujian diperoleh dari proses scraping di Twitter dari tanggal 1 Oktober 2016 sampai dengan 20 April 2017. Dalam penelitian ini digunakan data masukan berupa data tweet dan data keluaran berupa sentimen yang telah dikelompokkan menjadi 3 yaitu positif, negatif dan netral. Diperoleh 4571 data negatif, 1899 data netral, dan 1588 data positif. dengan tingkat kemenangan tertinggi di babak pertama di Ahok. Pada babak kedua dengan 2205 data diperoleh 604 tweet positif, 1123 data netral, 479 tweet negatif. Di babak kedua Anies Baswedan mendapat respon positif dan negatif yang lebih tinggi dari calon Ahok, sehingga peluang menang melawan Anies Baswedan lebih besar dari Ahok.

Kata Kunci: *Analisis Sentimen, Pilkada Jakarta 2017, Pengelompokan Spektral, Pengelompokan, Penambangan Opini*

1. Pendahuluan

Konsep data mining semakin banyak dikenal dalam berbagai macam bidang ilmu, misalnya: ilmu pendidikan, ilmu pemerintahan dan ilmu kesehatan, pada umumnya, Data mining muncul dari banyaknya jumlah data yang tersimpan dalam suatu basis data. Sehingga dari banyaknya data yang dimiliki dalam suatu basis data dapat digali untuk

δELTΔ Jurnal Ilmiah Pendidikan Matematika
Vol. 9 No. 1 Januari 2021 hal. 27 - 36

memperoleh suatu pengetahuan yang bermanfaat. Data mining adalah studi yang mengumpulkan, membersihkan, mengolah, menganalisis, dan memperoleh manfaat dari data (Charu C, Angarwal, 2015). Data mining adalah proses untuk menemukan informasi yang berguna secara otomatis di repositori data yang besar. Teknik penambangan data dikerahkan untuk menjelajahi database besar untuk menemukan pola baru dan berguna yang mungkin tidak diketahui (Van Dongen, 2000).

Salah satu teknik yang dikenal dalam data mining yaitu teknik *clustering*. *Clustering* merupakan salah satu metode Data Mining yang melakukan pemisahan/ pemecahan/ segmentasi data kedalam sejumlah kelompok *(cluster)* menurut karakteristik tertentu yang diinginkan, dalam jurnal retno Tri Wulandari metode *cluster* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (*similarity*) antara satu data dengan data yang lain. *Cluster* merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*).

Algoritma *Clustering* yang ada bermacam-macam, sebagai contoh *K-Means Clustering, Fuzzy C-Means Clustering,* dan sebagainya. *Algoritma clustering* yang sangat umum digunakan adalah *K-Means Clustering*. Metode *K-means* mudah dalam pengimplementasiannya serta memiliki waktu komputasi yang cukup cepat. Tetapi metode ini mempunyai kelemahan dalam menganalisis persebaran data serta bergantung pada inisialisasi *centroid. K-means* hanya melihat jarak data ke masing-masing *centroid* pada setiap *cluster*. Salah satu metode *clustering* lain yang diusulkan dalam memperbaiki akurasi regresi adalah *Spectral Clustering* (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 *graph* dari data yang ada. Di mana *verteks* dari *graph* tersebut merupakan setiap pada data. *Edgenya* berupa hubungan antar data yang biasanya bernilai jarak dari dua *record* yang berhubungan (Trivedi, S., A. Pardos, Z., & N. Sar, G. 2008).

Penyelenggaraan Pemilu termasuk Pilkada merupakan wujud pelaksanaan sistem demokrasi tidak langsung (*indirect democracy*). Pada sistem demokrasi tidak langsung (*indirect democracy*) atau demokrasi perwakilan (*representative democracy*), dilaksanakannya Pilkada bertujuan agar Kepala Daerah benar-benar bertindak atas nama rakyat sehingga pemilihannya harus dilakukan sendiri oleh rakyat melalui Pemilu (Marijan, 2010: 37). Oleh karena itu, sesungguhnya penyelenggaraan Pilkada adalah sarana pemberian mandat dan legitimasi dari rakyat kepada Kepala Daerah dengan harapan Kepala Daerah yang terpilih dapat memperjuangkan kepentingan rakyat. .

Terdapat banyak komentar positif dan negatif masyarakat saat sebelum pemilu dilaksanakan maupun saat pemilu sedang berlangsung mengenai pemilu yang diadakan. Twitter menyediakan sumber - sumber opini yang banyak jumlahnya, opini di Twitter belum dapat diidentifikasi secara langsung merupakan opini positif atau opini negatif. Informasi yang diterima langsung dari media Twitter jika dipahami apa adanya tanpa melakukan pengecekan terlebih dahulu sumber asli dan yang menyebarkan terpercaya atau tidak cenderung akan menjadi berita palsu atau *hoax* bahkan menjurus "kampanye hitam" kepada calon lawan di pemilihan presiden. Di Twitter ditemukan opini-opini, aspirasi atau komentar dari masyarakat yang dapat digunakan untuk mengekspresikan peristiwa yang sedang terjadi dalam hal ini adalah yang berhubungan dengan pemilihan kepala daerah dengan *hashtag* (#) #pilkada jakarta 2017. Agar opini-opini tersebut dapat dimanfaatkan dan berguna, dibutuhkan berbagai proses sehingga didapatkan suatu informasi yang penting melalui analisis sentimen.

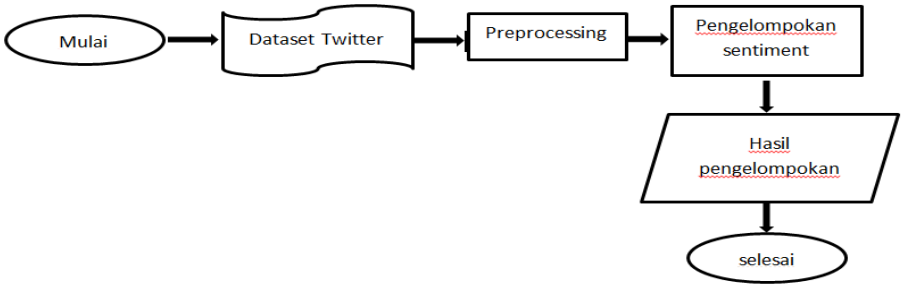
Analisis sentimen disebut juga dengan *opinion mining* (penambangan opini) yaitu proses untuk mengekstrak suatu opini atau pendapat dari dokumen untuk topik tertentu (H. Kaur, V. Mangat dan N, 2017). Analisis sentimen dilakukan untuk mengetahui kecenderungan opini seseorang terhadap sebuah peristiwa atau masalah, apakah cenderung beropini positif atau negatif. Teknik yang digunakan yaitu *text mining*. *Text mining* mengekstraksi kata kunci atau mengekstraksi pendapat dan ulasan analisis text sehingga dapat mendukung untuk memahami pendapat masyarakat dalam data *text*.

Data yang akan ditambang untuk *clustering* adalah pesan dan komentar-komentar yang sudah diposting pada media sosial Twitter. Pesan-pesan yang pernah ditulis akan diekstrasi yang akan menjadi suatu data yang akan digunakan untuk pengelompokkan jenis pesan. Variabel-variabel yang akan digunakan pada penelitian ini adalah *Tweet,* 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 Twitter untuk mengelompokan data menggunakan *Spectral Clustering*.

2. Metode Penelitian

Penelitian ini dilakukan dengan menggunakan beberapa tahapan seperti yang dijelaskan pada gambar 1, seperti pengumpulan data, pra-pemrosesan data, pengembangan sistem menggunakan metode *Spectral Clustering*, serta evaluasi sistem.

δELTΔ Jurnal Ilmiah Pendidikan Matematika
Vol. 9 No. 1 Januari 2021 hal. 27 - 36



Gambar 1. *Flowchart* penelitian

2.1 *Text Data*

Text mining adalah teknologi baru yang digunakan untuk data perusahaan yang selalu bertambah sehingga data teks yang tidak terstruktur tersebut dapat dianalisis (Francis dan Flynn, 2010). Data mining adalah disiplin ilmu yang tujuan utamanya adalah untuk menambang pengetahuan dari data atau informasi yang dimiliki. *Text mining* adalah salah satu solusi yang dapat membantu permasalahan di atas (Susanto dan Suryadi, 2010). *Text mining* mirip dengan data mining, kecuali pada teknik data mining yang didesain untuk pengerjaan data yang terstruktur pada sebuah database, tapi *text mining* dapat bekerja pada data yang tidak terstruktur atau semi terstruktur seperti email, sebuah dokumen text lengkap, html dan lain-lain. Sehingga text mining merupakan sebuah penemuan baru dari informasi yang belum diketahui dengan mengekstrak informasi dari sumber tertulis (Gupta dan Lehal, 2009).

Menurut Kurniawan, et al. (2012), langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut :

a. *Text Pre-processing*

Tindakan yang dilakukan pada tahap ini adalah :

1. *To lower case*, yaitu mengubah semua karakter huruf menjadi huruf kecil.
2. *Tokenizing*, yaitu proses penguraian deskripsi yang semula berupa kalimat – kalimat menjadi kata-kata.

b. *Feature Selection*

Pada tahap ini tindakan yang dilakukan adalah:

1. *stopword (stopword removal)* adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen. *Stopword* untuk bahasa Indonesia diperoleh dari: http://www.ranks.nl/ stopwords/indonesian (Doyle, 2010).

2. *stemming* adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya (*stem*).

2.2 Proses *Clustering* Menggunakan *Spectral Clustering*

Metode *cluster* adalah metode untuk menemukan dan mengelompokkan data yang memiliki kesamaan karakteristik antara satu data dengan data lainnya. *Cluster* adalah metode penambangan data tanpa pengawasan (H. Kaur, V. Mangat dan N. 2017). Analisis cluster mengklasifikasikan objek data hanya berdasarkan informasi yang terdapat dalam data yang menggambarkan objek dan nama. Tujuannya adalah agar objek-objek dalam suatu kelompok mirip satu sama lain dan berbeda dari objek-objek dalam kelompok lain. Semakin besar (atau homogenitas) dalam suatu kelompok dan semakin besar perbedaan antar kelompok, semakin baik atau lebih jelas pengelompokannya. Koleksi biasanya disebut sebagai clustering.

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:

δELTΔ Jurnal Ilmiah Pendidikan Matematika
Vol. 9 No. 1 Januari 2021 hal. 27 - 36

$$L = D - A$$

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

a. *Case Folding*

Proses ini merupakan proses yang mengubah huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*). Proses ini dilakukan untuk kemudian mempermudah dalam melakukan proses selanjutnya.

Input= PILGUB DKI 2017:

Output = pilgub dki 2017

Gambar 2. Contoh *Case Folding*

b. *Normalized*

Dalam data *twee*t, terdapat beberapa fitur yang tidak memiliki pengaruh pada proses selanjutnya, maka komponen-komponen tersebut dihilangkan, beberapa komponen yang perlu dihilangkan seperti, *username*, dan URL.

Input : pilgub dki 2017: pelanggaran pilkada jakarta | http://ln.is/www.bisnis.com/dmf1u … http://ln.is/bitly.com/uroyw . pembelajaran..jgn terulang di putaran ke2.

Output : pilgub dki 2017 pelanggaran pilkada jakarta … pembelajaranjgn terulang di putaran ke2

Gambar 3. Contoh *Normalized*

c. *Stopword removal*

Proses selanjutnya adalah *Stopword removal*, dalam proses ini setiap kata 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 =

Gambar 4. Contoh *Stopword Removal*

3.2. Membuat 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 :

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

δELTΔ Jurnal Ilmiah Pendidikan Matematika
Vol. 9 No. 1 Januari 2021 hal. 27 - 36

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

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

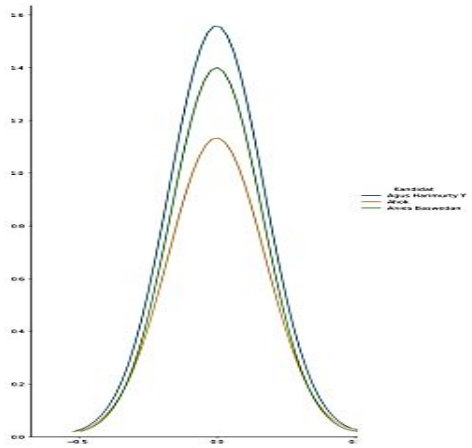
Selanjutnya dari nilai eigen terbentuklah Vektor eigen sebagi berikut :

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

Maka vector eigen yang terbentuk sebagai berikut :

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

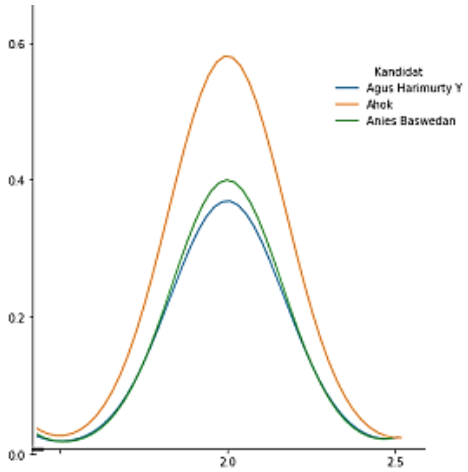
Vektor eigen kemudian dapat dikelompokkan menjadi 3 cluster yaitu 0, 1, 2. Dengan menggunakan algoritma clustering dapat dikelompokkan menjadi 3 yaitu 0 untuk negatif, 1 untuk netral dan 2 untuk positif. Dari metode *Spectral Clustering* diperoleh hasil 4571 *tweet* yang bersentimen negatif, 1899 *tweet* 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 49.48% dan terlihat Ahok memiliki sentimen negatif terendah sebanyak 57.9%.

δELTΔ Jurnal Ilmiah Pendidikan Matematika
Vol. 9 No. 1 Januari 2021 hal. 27 - 36



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%.

3. Kesimpulan

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 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 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.

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

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

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

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.