Vol 3 No. 2, Agustus 2021, pp. 221-228 **Jurnal Pepadun**

Clustering K-Means Jenis Kata Pada Laporan Kegiatan Kuliah Kerja Nyata (KKN) Universitas Lampung Menggunakan *Word2vec*

[1]Kristina Ademariana, [2]Aristoteles, [3]Favorisen Rosyking Lumbanraja, dan [4]Rico Andrian

[1,2,3,4]Program Studi Ilmu Komputer, Universitas Lampung
Jalan Soemantri Brojonegoro No. 1 Gedung Meneng, Bandar Lampung, Indonesia
e-mail : [1]kristina.ademariana1566@students.unila.ac.id, [2]aristoteles@gmail.com,
[3]favorisenrosykinglumbanraja@fmipa.unila.ac.id, [4]rico.andrian@fmipa.unila.ac.id

Abstract *— Kuliah Kerja Nyata (KKN) is a form of student service activities for the community, requesting and developing science and technology carried out off-campus within a period, linking work, and special requirements managed by the Badan Pelaksana Kuliah Kerja Nyata (BP-KKN). While carrying out KKN activities, each group of students is required to upload a report of the activities carried out in the village. In uploading the report file, there are several categories in each activity, including socialization, training, and character development. To classify the results of uploading activities one of which can be done using clustering techniques. In this research, a clustering of discussion on KKN student activities will be conducted at the University of Lampung. The text mining method is used to process KKN student activities to be more structured. Information on the KKN student activities was obtained as a feature with the Word2Vec weighting technique. The algorithm used is the K-Mean algorithm which has a high accuracy of the size of the object, so this algorithm is relatively more measurable and efficient for processing large numbers of objects. From the results of research conducted, it has been found that apply the text mining process algorithm for clustering with the K-means method on the Unila KKN Student activity data produces a value of k = 2, a lot of filtered data in the preprocess is 6284 data, using this method has not yet gotten a good association analysis because the results of the second cluster do not show the general types of words, typos and reporting activities by students who are not specifically can affect the results of clustering that is not good.*

Keywords**:** *Clustering; K-Means; Text Mining; Word2Vec.*

1. PENDAHULUAN

KKN merupakan proses pembelajaran bagi mahasiswa sekaligus wahana pemberdayaan masyarakat yang direncanakan dan dilaksanakan secara sistematis berdasarkan masalah dan potensi yang ada di masyarakat, dirumuskan serta dilakukan bersama masyarakat. Hal ini diharapkan dapat memacu kemampuan masyarakat dalam pengembangan diri dan wilayah, sehingga kesejahteraannya meningkat. Pelaksanaan KKN di Universitas Lampung dikelola oleh Badan Pelaksana Kuliah Kerja Nyata (BP-KKN). Dalam mengelola seluruh pelaksanaan KKN di Universitas lampung, BP-KKN memegang amanah mulai dari pendaftaran mahasiswa, pelaksanaan KKN, ujian, dan tahap pemberian nilai kepada mahasiswa atas kegiatan selama pelaksanaan KKN [1].

Proses pendaftaran mahasiswa dimulai dari penyerahan berkas surat pernyataan kesediaan untuk mengikuti dan menaati peraturan KKN. Kegiatan yang dilakukan oleh mahasiswa selanjutnya mendaftar *online*. Pada sistem informasi KKN untuk mengisi borang biodata lengkap beserta foto guna memudahkan BP-KKN memonitoring mahasiswa. Setelah melakukan pendaftaran, BP-KKN bertugas untuk membentuk kelompok mahasiswa dengan pembagian jumlah peserta kelompok jenis kelamin yang adil dan diusahakan tidak terdapat jurusan yang sama dalam satu kelompok. Setelah dilakukannya kegiatan pengelompokkan, dilakukannya pelaksanaan KKN selama 40 hari yang dibimbing oleh Dosen Pembimbing Lapangan (DPL) dan Dosen

Koordinator Pembimbing Lapangan (KDPL) yang mana mahasiswa wajib mengikuti seluruh tata tertib pelaksanaan KKN.

Selama melaksanakan kegiatan KKN, setiap kelompok mahasiswa diwajibkan mengunggah laporan kegiatan yang dilakukan di desa. Dalam mengunggah *file* laporan terdapat beberapa kategori di setiap kegiatan diantaranya sosialisasi, pelatihan, dan pengembangan karakter. Untuk mengelompokkan hasil dari unggahan kegiatan salah satu yang dapat dilakukan menggunakan teknik clustering. Menurut Rahman [2] *clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum. *Clustering* merupakan proses partisi satu *set* objek data ke dalam himpunan bagian yang disebut dengan *cluster*. *Clustering* merupakan metode analisis data yang mempunyai karakteristik sama akan dikelompokkan pada satu kelompok dan data yang memiliki karakteristik berbeda akan dikelompokkan pada kelompok yang lainnya [3]. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*.

Dalam penelitian ini akan dilakukan klasterisasi pembahasan kegiatan mahasiswa KKN di Universitas Lampung. Metode *text mining* digunakan untuk mengolah kegiatan mahasiswa KKN menjadi lebih terstruktur. Dalam melakukan transformasi dokumen dilakukan pembobotan kata (*word embeddings*) untuk mendapatkan vektor dari setiap kata. Menurut Kencana dan Maharani [4] *Word embeddings* merupakan sebutan dari seperangkat bahasa pemodelan dan teknik pembelajaran fitur atau *feature learning* dimana setiap kata dari kosakata (*vocabulary*) memiliki vektor yang mewakili makna dari kata tersebut dan kata-kata tersebut dipetakan ke dalam bentuk vektor bilangan riil.

Informasi di dalam kegiatan mahasiswa KKN diperoleh sebagai fitur dengan teknik pembobotan *Word2Vec*. Menurut Mikolov, dkk [5] *Word2Vec* merupakan representasi kata dalam bentuk vektor yang dibuat oleh Google. Metode *Word2Vec* bertujuan untuk menemukan hubungan tersembunyi pada kata. Setiap kata mewakili distribusi bobot pada elemennya. Metode ini memiliki 2 model arsitektur, yaitu *Continuous Bag-of-Words* (*CBOW*) dan *Skip-Gram*. Model CBOW bertujuan untuk memprediksi target kata dari konteks kata, sedangkan model *Skip-Gram* bertujuan untuk memprediksi probabilitas kata yang dapat menjadi konteks kata dari target kata yang ada [6].

Dalam menentukan banyak k kelompok yang paling optimal adalah menggunakan Metode *Elbow*. Menurut Madhulatha [7] Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik. Metode ini bertujuan untuk mengukur *Variance* antara *Centroid* dengan *dataset* Observasi yang berdekatan, atau biasa disebut sebagai WCSS (*Within-Cluster Sums Of Squares*) [8].

Metode *Clustering* yang digunakan dalam proses *cluster* adalah algoritma *K-Means*. Tujuan dari *K-Means*, yaitu untuk meminimalkan dari fungsi objektif yang diatur dalam proses pengelompokan, pada umumnya akan berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok [9]. Algoritma *K-Means* memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar.

2. METODOLOGI PENELITIAN

2.1. Alat dan Bahan

2.1.1 Perangkat Lunak

Perangkat keras yang digunakan untuk pembuatan sistem berupa sebuah komputer dengan spesifikasi sebagai berikut:

1. Sistem Operasi Windows 10 Home Single Language 64 Bit

2. *ClickCharts Diagram Flowchart Software*

   ClickCharts Diagram Flowchart Software adalah perangkat lunak yang digunakan untuk membuat representasi visual dari suatu proses yang kompleks, membuat *value stream* dan diagram alir, serta optimisasi proses. Keunggulan dari pembuatan *flowchart* yaitu dapat menampilkan proses yang sangat terperinci dan rumit menjadi lebih mudah dipahami [10].

3. Python

   Python merupakan bahasa pemrograman tingkat tinggi yang bersifat interpretatif dan *multiplatform* (dapat bekerja di berbagai *platform* seperti MS Windows, Linux, Macintosh, dan lainnya). Python dikembangkan oleh Guido Van Rossum pada tahun 1990 di Amsterdam yang merupakan kelanjutan dari bahasa pemrograman ABC. Python dapat digunakan secara bebas oleh siapapun [11].

4. Jupyter *Notebook*

   Jupyter membutuhkan Python untuk diinstal karena didasarkan pada bahasa Python. Salah satu alat yang akan mengotomatisasi instalasi Jupyter dari GUI yaitu *software* Anaconda. Bagian utama Jupyter *Notebook* yaitu metadata untuk mengatur dan menampilkan *Notebook*, nomor versi *software* yang digunakan untuk membuat *Notebook*, dan daftar sel [12].

5. *Natural Language Toolkit* (*NLTK*)

   NLTK merupakan platform terkemuka yang bersifat open-source dan gratis untuk membangun program Python, NLTK yang digunakan dalam penelitian ini menggunakan versi 3.4.5. NLTK disebut sebagai “perpustakaan yang luar biasa untuk bermain dengan bahasa alami” karena menyediakan antarmuka yang mudah digunakan dan serangkaian pustaka pemrosesan teks untuk klasifikasi, tokenisasi, *stemming*, *tagging*, *parsing*, penalaran semantik, dan pustaka pemrosesan bahasa alami lainnya. NLTK tersedia untuk Windows, Mac OS X, dan Linux [13].

6. Scikit-Learn

   Pada penelitian ini digunakan Scikit-Learn versi 0.21.3. Scikit-Learn adalah pustaka Python *opensource* yang mengimplementasikan serangkaian pembelajaran mesin, pra-proses, *cross validation*, dan algoritma visualisasi menggunakan antarmuka terpadu. Awalnya, Scikit-Learn dikembangkan oleh David Cournapeau pada tahun 2007 dan dikeluarkan pertama kali (v0.1 beta) pada akhir Januari 2010 [14]. Keunggulan yang terdapat dalam Scikit-Learn diantaranya memiliki berbagai fitur untuk klasifikasi, regresi dan algoritma *clustering* untuk data mining dan analisis data [15].

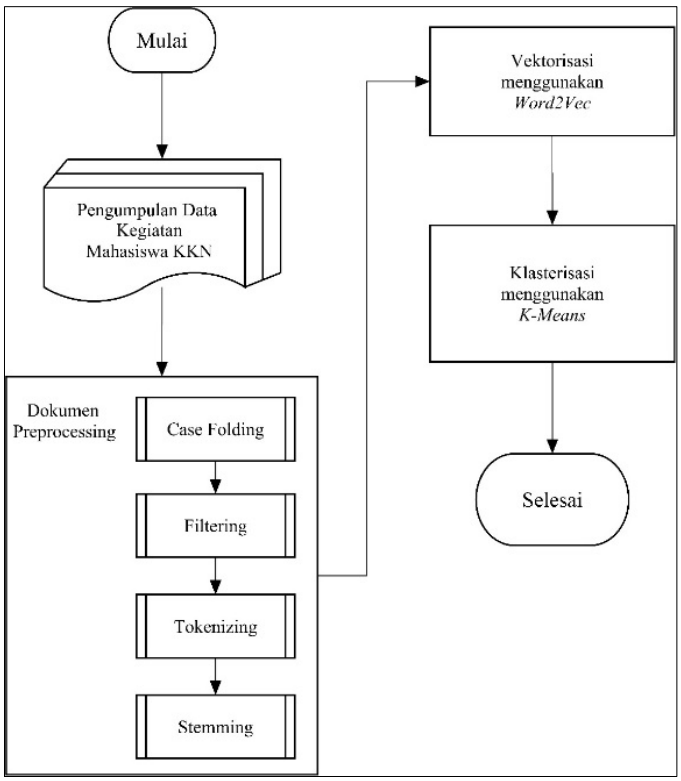
2.1.2 Perangkat Keras

Perangkat keras yang digunakan untuk pembuatan sistem berupa sebuah komputer dengan spesifikasi sebagai berikut:

1. Processor: Intel® CoreTM i7-8550U (2.0 GHz, 3MB L3 Cache)
2. System type: 64-bit Operating System, x64-based processor
3. RAM: 32 GB DDR4-2400, LPDDR3-2133
4. Harddisk: 500 GB HDD SATA, 5400 RPM

2.2. Metode

Implementasi text mining pada clustering kegiatan KKN menggunakan *tools Natural Language Toolkit* (*NLTK*) sedangkan proses hingga tahap evaluasi menggunakan *library* Scikit-Learn yang tersedia pada Python. Tahapan penelitian ini dilakukan dengan beberapa langkah yang dapat dilihat pada Gambar 1.



Gambar 1. Alur penelitian implementasi *Text Mining* dan *Word2Vec* untuk klasterisasi kegiatan mahasiswa KKN Unila

Berikut ini merupakan tahapan yang dilalui dalam penelitian ini:

2.2.1 Pengumpulan data kegiatan mahasiswa KKN Universitas Lampung

Pada penelitian ini, pengumpulan data hasil Kegiatan Mahasiswa KKN Unila yang digunakan yaitu berupa laporan Kegiatan Mahasiswa KKN Universitas Lampung dengan format .CSV (*Comma Separated Value*). Data laporan kegiatan Mahasiswa KKN Universitas Lampung yang digunakan berjumlah 14 kabupaten, 95 kecamatan, 864 desa, dan 11626 kegiatan mahasiswa.

Tabel 1. *User stories* iterasi 1

npm	nama_desa	nama_kecamatan	nama_kabupaten	kegiatan
1442011023	Rantau Tijang	Pardasuka	Pringsewu	sosialisasi mengenai pentingnya. . .
1414121074	Ambarawa	Ambarawa	Pringsewu	memberitahu kepada para petani akan . . .
1414121074	Ambarawa	Ambarawa	Pringsewu	sosialisasi dan membuat cabai menjadi . . .
1414121074	Ambarawa	Ambarawa	Pringsewu	sosialisasi dan . . .

2.2.2 Praproses data kegiatan mahasiswa KKN Universitas Lampung

Praproses pada laporan kegiatan diuraikan sebagai berikut. Pertama, tahap *case folding* untuk mengubah semua huruf dalam laporan kegiatan mahasiswa KKN Unila menjadi huruf kecil sehingga menghasilkan bentuk laporan kegiatan yang standar dan seragam. Kedua, mengimplementasikan fungsi *filtering* yang terdiri dari tahap *remove number* untuk menghapus angka pada laporan kemudian diganti menjadi *whitespace*, tahap *remove whitespace* untuk menghapus spasi yang berlebih, tahap *remove punctuation* untuk menghapus tanda baca, dan tahap *remove stopwords* untuk menghilangkan kata yang tidak memiliki makna sehingga menghasilkan *array* kata yang bermakna saja. Ketiga yaitu mengimplementasikan tokenisasi untuk memotong kalimat yang sudah melalui kelima proses sebelumnya menjadi kata/token. Selanjutnya yaitu *stemming* untuk mencari kata dasar dari setiap kata hasil tokenisasi dengan membuang imbuhan di awal (*prefix*) dan di akhir (*sufix*) kata tersebut. Setelah melakukan proses *stemming* langkah selanjutnya yaitu membersihkan data yang hanya berisi 3 kata. Setelah membersihkan data dari kalimat yang hanya berisi 3 kata, langkah selanjutnya melakukan tahap pemodelan yang menggunakan data berjumlah 1000 dan tahap terakhir memperbaiki kesalahan kata dalam pengetikan (*typo*).

2.2.3 *Feature extraction* hasil praproses kegiatan KKN Unila

Dari tahap praproses diperoleh korpus kegiatan dalam bentuk *string*. Kemudian dilakukan transformasi dokumen dari tipe *string* menjadi bentuk vektor dengan menggunakan teknik pembobotan *Word2Vec*. Metode *Word2Vec* bertujuan untuk menemukan hubungan tersembunyi pada kata. Setiap kata mewakili distribusi bobot pada elemennya. Metode ini memiliki 2 model arsitektur, yaitu *Continuous Bag-of-Words* (*CBOW*) dan *Skip-Gram*. Model *CBOW* bertujuan untuk memprediksi target kata dari konteks kata, sedangkan model *Skip-Gram* bertujuan untuk memprediksi probabilitas kata yang dapat menjadi konteks kata dari target kata yang ada.

2.2.4 Penentuan jumlah *cluster* menggunakan metode *elbow*

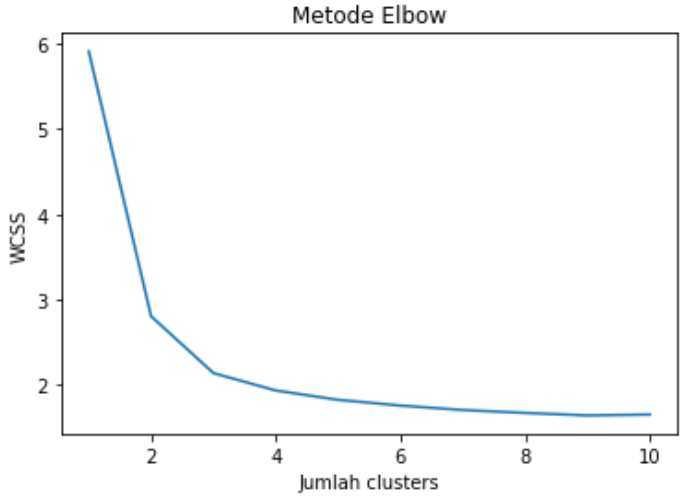
Dari hasil pembobotan *Word2Vec* langkah selanjutnya yaitu menentukan jumlah *cluster* menggunakan metode *Elbow*. Metode *Elbow* digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* yang optimal dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik.

2.2.5 Proses *clustering* menggunakan pendekatan metode *K-Means*

K-means mempartisi data ke dalam suatu kelompok dimana data yang berkarakteristik sama akan dimasukkan ke dalam satu kelompok sama, sedangkan data yang memiliki karakteristik yang berbeda akan dikelompokkan ke dalam kelompok lainnya. Tujuan dari pengelompokan ini untuk meminimalkan dari fungsi objektif yang diatur dalam proses pengelompokan, pada umumnya akan berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok.

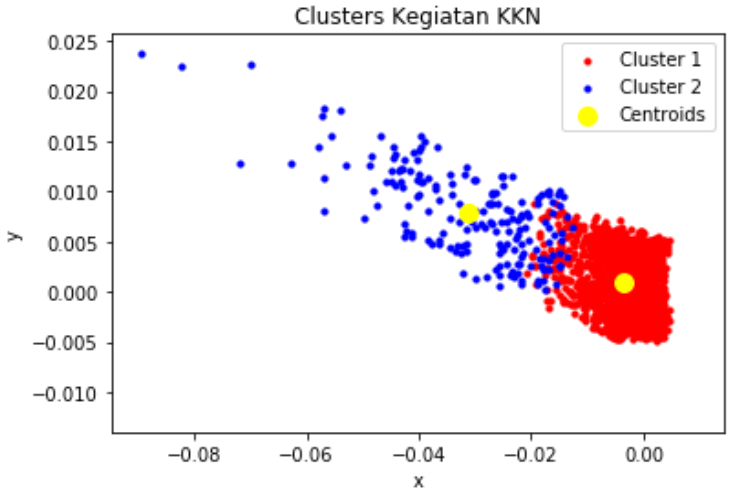
3. HASIL DAN PEMBAHASAN

Hasil dari tahap praproses diperoleh 6284 data, setelah dibersihkan dari data yang memiliki kurang dari 3 kata selanjutnya adalah melakukan pemodelan menggunakan 1000 data. dan memperbaiki data yang *typo* (kesalahan ketik). Kemudian dari 1000 data kegiatan Mahasiswa KKN Unila dilakukan transformasi data yaitu pembobotan pada setiap kata yang diperoleh yaitu 1870 kata. Selanjutnya menentukan jumlah *cluster* menggunakan Metode *Elbow*. Untuk menentukan jumlah *cluster* terbaik digunakan Metode *Elbow*. Metode ini bertujuan untuk mengukur *Variance* antara *Centroid* dengan *dataset* Observasi yang berdekatan, atau biasa disebut sebagai WCSS (*Within-Cluster Sums Of Squares*). Pada metode *elbow* nilai *cluster* terbaik yang akan diambil dari nilai *Sum of Square Error* (*SSE*) yang mengalami penurunan yang signifikan dan berbentuk siku.



Gambar 2. Hasil fungsi metode *Elbow*

Pada Gambar 2 menunjukkan bahwa bentuk *elbow* (siku) terlihat hasil yang diperoleh adalah k=2. Setelah menentukan jumlah *cluster* kemudian melakukan *clustering* menggunakan metode *K-Means*.

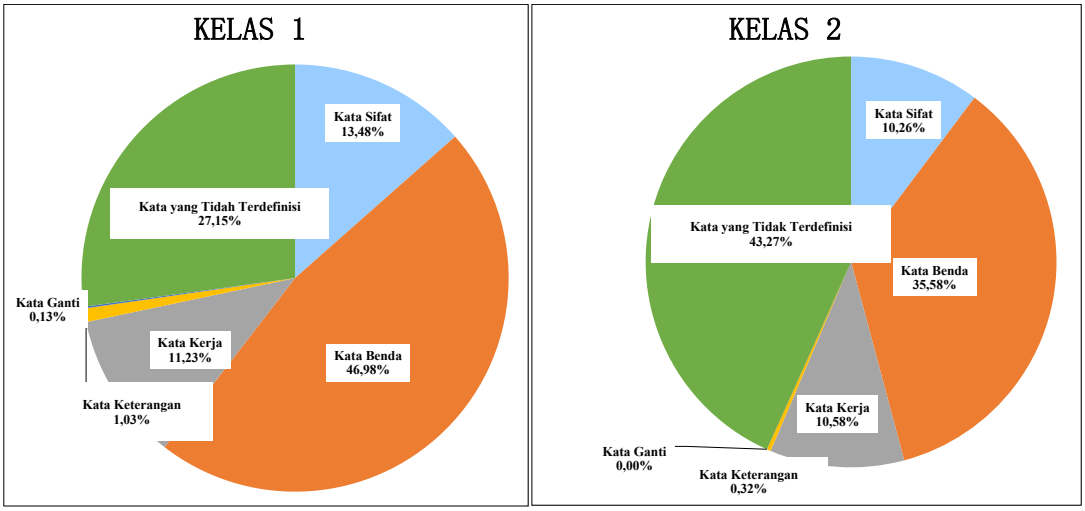


Gambar 3. Hasil *K-Means Clustering*

Dari hasil *Clustering K-Means* pada Gambar 3, jarak setiap *data input* terhadap masing-masing *centroid* menggunakan rumus jarak Euclidean (*Euclidean Distance*) hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid*. Berikut adalah persamaan *Euclidean Distance*:

$$d(x,y) = ||x-y|| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 : i = 1,2,3,\ldots.n} \quad (1)$$

Langkah selanjutnya adalah memberi label pada plot masing-masing kelas berdasarkan kelas kata dalam Kamus Besar Bahasa Indonesia (KBBI). Pelabelan kata berdasarkan kelas kata diantaranya kata sifat, kata benda, kata kerja, kata keterangan, kata ganti, dan kata yang tidak terdefinisi dalam kelas kata pada Kamus Besar Bahasa Indonesia (KBBI).



Gambar 4. Hasil pelabelan *cluster* 1 dan *cluster* 2

Berdasarkan Gambar 3 dapat dilihat jumlah kelas kata terbanyak pada *cluster* 1 dari tiga besar adalah yang pertama kelas kata benda dengan jumlah 46,98%, kedua kata yang tidak terdefinisi dengan jumlah 27,15%, yang ketiga kata sifat dengan jumlah 13,48% sedangkan jumlah kelas kata terbanyak pada *cluster* 2 dari tiga besar adalah yang pertama kata yang tidak terdefinisi dengan jumlah 43,27%, kedua kata benda dengan jumlah 35,58%, yang ketiga kata kerja dengan jumlah 10,58%.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, diperoleh beberapa kesimpulan bahwa penerapan algoritma proses *text mining* untuk melakukan *clustering* dengan metode *K-means* pada data kegiatan Mahasiswa KKN Unila menghasilkan nilai k=2, banyak data yang terfilter dalam praproses sejumlah 6284 data, menggunakan metode ini belum mendapatkan analisis asosiasi yang baik karena hasil *cluster* ke 2 tidak menunjukkan jenis kata secara umum, kesalahan ketik (*typo*) dan pelaporan kegiatan oleh mahasiswa yang tidak spesifik dapat mempengaruhi hasil *clustering* yang tidak baik.

DAFTAR PUSTAKA

[1] BP-KKN, "*Buku Panduan KKN Unila*," Badan Pelaksana Kuliah Kerja Nyata - Universitas Lampung, 2018. [Online]. Available: http://kkn.unila.ac.id/wp-content/uploads/2018/07/Buku-Panduan-KKN-Unila-Periode-II-2018.pdf. [Accessed 22 October 2018].

[2] A. R. Tegar, Wiranto & R. Anggrainingsih, "Coal Trade Data Clustering Using K-Means (Case Study PT. Global Bangkit Utama)," *ITSMART: Jurnal Ilmiah Teknologi dan Informasi,* vol. 2, no. 1, pp. 24-31, 2017.

[3] Gustientiedina, M. H. Adiya & Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *Jurnal Nasional Teknologi dan Sistem Informasi,* vol. 5, no. 1, pp. 17-24, 2019.

[4] K. B. A. W. Kencana & W. Maharani, "Klasifikasi Opini Pada Fitur Produk Berbasis Graph Opinion Classification for Product Feature Based on Graph," *e-Proceeding of Engineering,* vol. 4, no. 2, pp. 3148-3155, 2017.

[5] T. Mikolov, K. Chen, G. Corrado & J. Dean, "Efficient Estimation of Word Representations in Vector Space," 7 September 2013. [Online]. Available: https://arxiv.org/pdf/1301.3781.pdf. [Accessed 27 October 2013].

[6] A. F. Niasita, P. P. Adikara & S. Adinugroho, "Analisis Sentimen Pembangunan Infrastruktur di Indonesia dengan Automated Lexicon Word2Vec dan Naive-Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 3, pp. 2673-2679, 2019.

[7] T. S. Madhulatha, "An Overview on Clustering Methods," *IOSR Journal of Engineering,* vol. 2, no. 4, pp. 719-725, 2012.

[8] Minitab Express Support, "Interpret all statistics and graphs for Multiple Regression," Minitab, 2019. [Online]. Available: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/multiple-regression/interpret-the-results/all-statistics-and-graphs/.

[9] E. Prasetyo, *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*, Yogyakarta: Andi, 2012.

[10] NCH, "ClickCharts Diagram & Flowchart Software," NCH Software, [Online]. Available: https://www.nchsoftware.com/chart/index.html. [Accessed 21 September 2019].

[11] J. Enterprise, *Trik Cepat Menguasai Pemrograman Python*, Jakarta: PT Elex Media Komputindo, 2016.

[12] D. Toomey, *Learning Jupyter*, Birmingham: Packt, 2016.

[13] A. Mitrani, "NLTK Applications for NLP and Python," Medium: Towards Data Science, 11 October 2019. [Online]. Available: https://towardsdatascience.com/nltk-applications-for-nlp-and-python-dc8c5381668a. [Accessed 29 April 2019].

[14] M. Mishra, "Hands-On Introduction To Scikit-learn (sklearn)," Medium: Towards Data Science, [Online]. Available: https://towardsdatascience.com/hands-on-introduction-to-scikit-learn-sklearn-f3df652ff8f2. [Accessed 15 September 2019].

[15] N. Kumar, "Learning Model Building in Scikit-learn : A Python Machine Learning Library," Geeks for Geeks, 6 August 2019. [Online]. Available: https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library. [Accessed 2019].