CLUSTERING BERITA MENGGUNAKAN ALGORITMA TF-IDF DAN K-MEANS DENGAN MEMANFAATKAN SUMBER DATA *CRAWLING* PADA SITUS DETIK.COM

I Made Arta Purniawan[a1], Gusti Made Arya Sasmita [a2], I Putu Agus Eka Pratama [b3]
[a]Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Bukit Jimbaran, Bali, Indonesia, telp(0361)701806
e-mail: [1]kadeajus@gmail.com, [2]aryasasmita@it.unud.ac.id, [3]eka.pratama@unud.ac.id

Abstrak

Clustering berita bertujuan untuk mengidentifikasi setiap kelompok berita yang terbentuk dari implementasi metode K-Means yang didasarkan dari proses pembobotan kata menggunakan Algoritma TF-IDF (Term Frequency Inverse Document Frequency). Proses clustering menggunakan berita hasil crawling dari situs detik.com selama kurun waktu satu tahun (2018) yang berjumlah 124.509 berita dan disimpan dalam bentuk file CSV (Comma Seperated Value). Sebelum melakukan proses clustering, dataset sebelumnya harus melalui tahap text-processing berupa: case folding, tokenizing, stopword removal, dan stemming. Metode TF-IDF dan K-Means digunakan untuk proses pengelompokan / clustering. Metode TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia, kemudian dilanjutkan dengan Metode K-Means untuk proses pengelompokan berdasarkan ciri yang sama / kemiripan antar dokumen. Dalam prosesnya, terdapat dua kali implementasi metode K-Means, masing-masing menggunakan 16 centroid dan 12 centroid. Ini dikarenakan pada proses pertama, terdapat kelompok / cluster yang tidak bisa di identifikasi karena mengandung kata yang umum, sehingga diperlukan implementasi kedua. Berdasarkan hasil pengujian terhadap 124.509 berita, terdapat 27 kelompok berita yang berhasil di identifikasi dengan kemampuan aplikasi yang cukup memadai dalam memproses data dalam ukuran besar.

Kata kunci: *TF-IDF, K-Means, Clustering, Data Mining, Dataset Crawling*

Abstract

News clustering aims to identify each news group that is formed from the implementation of the K-Means method which is based on the word weighting process using the TF-IDF (Term Frequency Inverse Document Frequency) Algorithm. The clustering process uses news crawled from the detik.com site for a period of one year (2018), totaling 124,509 news stories and stored in the form of a CSV (Comma Seperated Value) file. Before carrying out the clustering process, the previous dataset must go through a text-processing stage in the form of: case folding, tokenizing, stopword removal, and stemming. The TF-IDF and K-Means methods are used for the clustering process. The TF-IDF method assigns weights to each keyword in each category to find the similarity of keywords to the available categories, then continues with the K-Means Method for the grouping process based on similar characteristics / similarities between documents. In the process, there are two implementations of the K-Means method, each using 16 centroids and 12 centroids. This is because in the first process, there are groups / clusters that cannot be identified because they contain common words, so a second implementation is needed. Based on the results of testing on 124,509 news stories, there are 27 news groups that have been successfully identified with adequate application capabilities in processing large data.

Keywords : *TF-IDF, K-Means, Clustering, Data Mining, Dataset Crawling*

1. Pendahuluan

Berita merupakan hal yang wajib untuk di ikuti setiap harinya oleh masyarakat. Dengan kemajuan Teknologi Informasi dewasa ini yang semakin pesat, berita sudah dapat diakses kapan saja dan dimana saja dengan sangat cepat baik dari kalangan anak – anak hingga orang dewasa. Berita dapat diperoleh mulai dari media sosial populer seperti Facebook, Instagram, Twitter bahkan sudah merambah ke *platform* pesan singkat seperti Whatsapp, Messenger, Telegram, LINE Chat.

Berita yang beredar tak sedikit hasil dari *sharing* beberapa portal berita online seperti detik.com, kompas.com, tempo.co, dan lain sebagainya. Dari sekian banyak contoh portal berita yang ada, detik.com adalah salah satu portal berita yang cukup populer di kalangan masyarakat Indonesia.

Menurut situs it-jurnal.com, detik.com merupakan portal berita paling populer yang disebut sebagai pelopor perusahaan media online di Indonesia, situs ini menyediakan berita terbaru dan komprehensif dari Indonesia dan seluruh dunia. Didirikan pada tahun 1998 dan bergabung dengan Transmedia di bawah CT Corp sejak Agustus 2011

Tentunya dalam kurun waktu lebih dari 9 tahun tersebut portal berita detik.com memiliki jutaan berita yang beredar dan dapat dibaca oleh masyarakat umum. Dari hal tersebut sebagai pengamat, meneliti *cluster* atau kelompok berita pada kurun tahun tertentu sangat diperlukan guna mengetahui berita atau kejadian apa saja yang dibicarakan dari sekian ratusan ribu berita yang di muat. Hal ini bisa menjadi dasar penelitian selanjutnya untuk mengetahui tren berita selama kurun waktu tertentu. Untuk mengetahui *cluster* berita setiap tahunnya, diperlukan sistem yang handal untuk menangani ratusan ribu berita yang kemudian akan penulis sebut *dataset* untuk diolah dan di visualisasikan berdasarkan tujuan dari penelitian. Berbeda dengan klasifikasi, Pendekatan berbeda dilakukan pada clustering. Tidak ada nama kelas atau batasan jumlah yang ditentukan, artinya *cluster* yang akan dibentuk tergantung dari masukan pengguna dan mencari jumlah *cluster* yang optimal dari beberapa percobaan yang dilakukan [1].

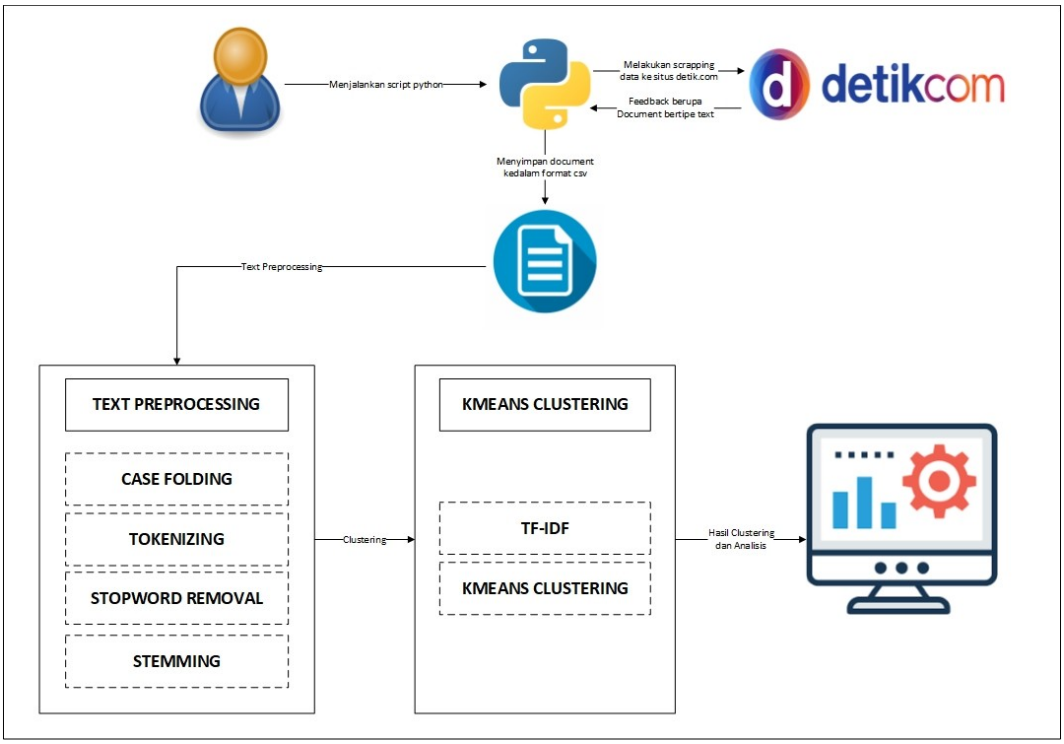
Dari hal tersebut muncul keinginan penulis untuk melakukan penelitian tentang bagaimana mengetahui *Cluster* Berita selama kurun waktu satu tahun berdasarkan hasil *crawling* pada portal berita detik.com. Berita yang diperoleh dari *crawling* pada portal berita detik.com akan di kelompokan menggunakan salah satu algoritma dari *Unsupervised Machine Learning*, yaitu K-Means. Sebelum diterapkannya metode K-Means, terlebih dahulu *dataset* harus melewati tahap *pre-processing* dan menghitung bobot dengan menggunakan metode TF-IDF (*Term Frequency–Inverse Document Frequency*). Hasil penelitian berupa persentase berita terhadap banyaknya berita dalam satu kelompok atau *cluster*.

2. Metodelogi Penelitian

Metode dalam melakukan penelitian ini meliputi, gambaran umum sistem, alur penelitian, sumber data dan instrumen pembuatan sistem.

2.1 Gambaran Umum Sistem

Sistem yang dibuat dapat melakukan crawling pada situs detik.com perdasarkan periode waktu tertentu dan disimpan kedalam file bertipe CSV, dan selanjutnya diproses kembali untuk agar hasil yang diperoleh sesuai dengan tujuan penelitian yaitu mengetahui kategori/topik berita selama tahun 2018. Gambaran umum sistem dapat dilihat pada Gambar 1



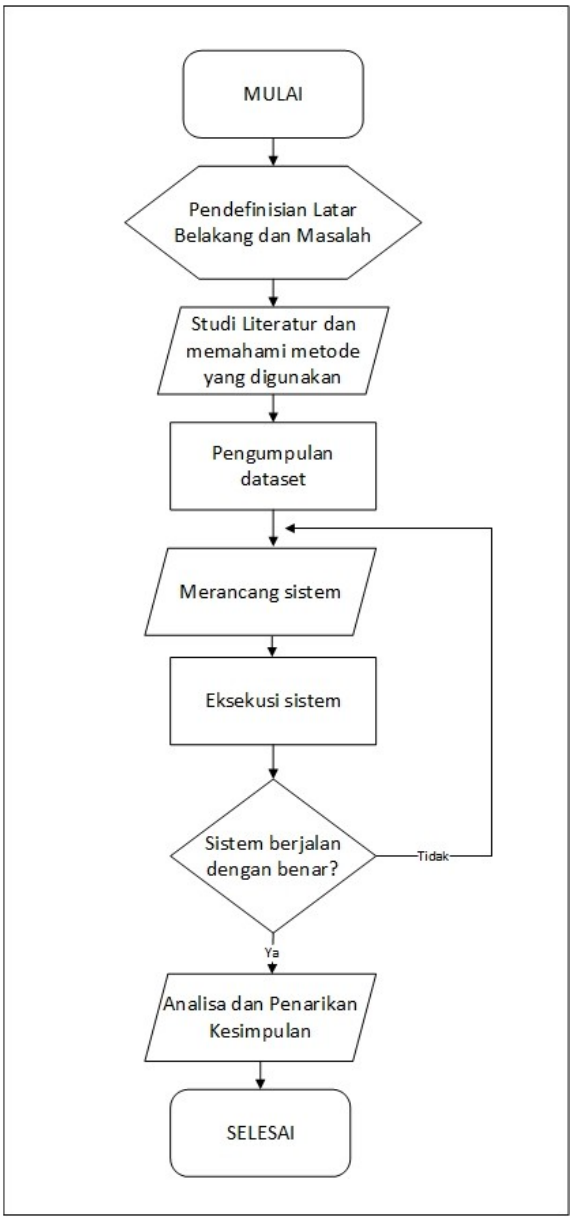
Gambar 1. Gambaran Umum Sistem

Sistem clustering berita pada penelitian ini bertujuan untuk membuat sebuah sistem yang mampu mengelompokan data berdasarkan bobot atau ciri yang sama hasil dari penerapan metode TF-IDF dan dikelompokan menggunakan algoritma K-Means dengan nilai Centroid yang sudah penulis tentukan yaitu sebanyak 16 Cluster. Pada tahap awal sistem melakukan crawling berita pada situs detik.com dengan mengirim parameter tanggal dan mengembalikan dokumen berupa text yang selanjutnya disimpan kedalam file bertipe CSV. Pada tahap selanjutnya dataset akan dibersihkan terlebih dahulu menggunakan teknik text preprocessing yang bertujuan untuk menghilangkan noise pada dataset sehingga meningkatkan nilai akurasi clustering nantinya. Sebelum ke tahap clustering dataset yang sudah dibersihkan akan dihitung bobotnya menggunakan metode TF-IDF yang berfungsi untuk mengubah data text menjadi data vektor yang merupakan data yang diperlukan agar algoritma K-Means dapat berjalan dan mengelompokan dataset berdasarkan ciri atau bobot yang sama. Hasil pengelompokan selanjutkan di tampilkan dan di identifikasi untuk mendapatkan kategori/topik berita selama tahun 2018.

2.2 Alur Penelitian

Alur Penelitian yang dilakukan dalam penelitian ini dapat dijabarkan melalui proses berikut.

1. Mendefinikan latar belakang penelitian
2. Studi literatur terkait dengan implementasi algoritma TF-IDF dan K-Means untuk melakukan *clusterig* terhadap *dataset* berupa text.
3. Mempelajari dan memahami metode – metode yang digunakan dalam proses *crawling* dan *clustering dataset* berupa text.
4. Melakukan pengumpulan *dataset* dengan metode *crawling* pada situs detik.com dan disimpan dalam format CSV file (*Comma Separated Value*)
5. Menganalisa dan merancang *script* program python untuk melakukan *clustering* mengunakan algoritma TF-IDF dan K-Means
6. Implementasi menggunakan software Jupyter Notebook berdasarkan perancangan yang telah dilakukan.
7. Melakukan analisa terhadap hasil *dataset* yang sudah diolah menggunakan algoritma TF-IDF dan K-Means.



Gambar 2. Alur Penelitian

2.3 Sumber Data

Sumber data yang digunakan dalam penelitian ini adalah sumber data primer. Data Primer adalah data yang langsung diperoleh oleh penulis melalui sumber pertama. Data diperoleh melalui *crawling* pada situs detik.com menggunakan *script* python. Data berupa dokumen text yang telah di publikasikan melalui situs detik.com dari rentang tanggal 1 Januari 2018 sampai dengan 31 Desember 2018 yang selanjutnya disimpan kedalam file dengan format CSV (*Comma Separated Value*), dapat dilihat pada Tabel 1

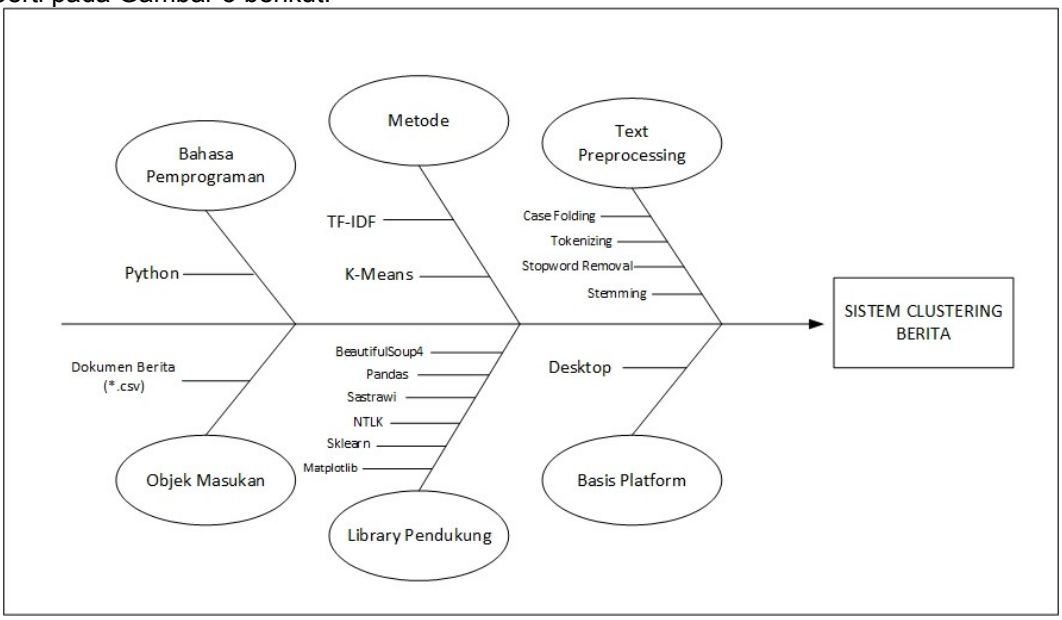
Tabel 1. Data Penelitian

No	Bulan	Jumlah
1	Januari	10.599 berita
2	Pebruari	9.437 berita
3	Maret	11.084 berita
4	April	10.451 berita
5	Mei	10.912 berita
6	Juni	10.142 berita
7	Juli	11.586 berita
8	Agustus	10.975 berita
9	September	9.886 berita
10	Oktober	10.643 berita
11	Nopember	9.490 berita
12	Desember	9.304 berita
Total		**124.509 berita**

Berita yang berjumlah 124.509 akan disimpan kedalam satu buah file bertipe CSV dan akan di lanjutkan ke proses selanjutnya yaitu Text Preprocessing, sebelum dilakukannya clustering pada *dataset*.

2.4 Analisis Sistem

Fishbone diagram atau Cause dan Effect diagram merupakan salah satu tools yang digunakan untuk mengambarkan hubungan antara sebab dan akibat agar dapat menemukan akar permasalahan. Diagram fishbone dari sistem clustering berita meliputi bahasa pemrograma, objek masukan, metode, library pendukung, text preprocessing dan basis flatform, seperti pada Gambar 3 berikut.



Gambar 3. *Fishbone* Diagram

Gambar 3 merupakan perancangan diagram fishbone dari sistem. Diagram fishbone menggambarkan batasan masalah dan komponen pendukung dalam penelitian yang dapat dijabarkan sebagai berikut.

1. Bahasa pemrograman yang digunakan untuk crawling berita pada situs detik.com, text preprocessing, clustering dataset dan visualisasi hasil adalah bahasa pemrograman python.
2. Objek masukan berupa file bertipe CSV (Comma Separated Value) hasil dari crawling pada situs detik.com
3. Metode yang digunakan untuk ekstrasi fitur pada text atau pembobotan adalah TF-IDF (Term Frequency – Inverse Document Frequency) sedangkan metode atau algoritma yang digunakan untuk clustering dataset adalah K-Means.
4. Library pendukung adalah komponen pendukung bahasa pemrograman python untuk menyelesaikan suatu masalah dengan cepat. Library pendukung yang digunakan penulis dalam melakakukan penelitian diantaranya: BeautifulSoup4 digunakan untuk crawling pada situs detik.com, Pandas digunakan untuk mengolah dataset, Sastrawi digunakan untuk stemming kata yang akan dibahas dalam text preprocessing, NTLK merupakan library python untuk proses pengolahan kata, dalam kasus ini digunakan untuk tokenizing kata, Sklearn merupakan library python untuk machine learning dan dalam kasus ini digunakan untuk proses clustering dokumen yang telah dilakukan proses pembobotan sebelumnya, Matplotlib digunakan untuk visualisasi data hasil.
5. Text preprocesing merupakan tahapan pertama sebelum dataset dapat diimplementasikan algoritma K-Means. Dataset terlebih dahulu dibersihkan untuk menaikan tingkat akurasi clustering. Tahapan ini meliputi: Case Folding yaitu merubah semua kata ke huruf kecil, Tokenizing yaitu merubah kata menjadi bentuk token,

Stopword Removal meghilangakan kata penghubung dan kata – kata yang tidak perlu, Stemming mengubah kata menjadi kata dasanya menggunakan library Sastrawi.

6. Basis platform yang digunakan adalah basis platform desktop.

2.5 Instrumen Pembuatan Sistem

Berikut merupakan beberapa alat pendukung yang digunakan dalam pembuatan sistem.

1. Spesifikasi Perangkat Keras
   a. Intel Core i5 4200U CPU @ 1.60GHz 2.30GHz
   b. Memory DDR3 8 GB 1600 MHz
   c. NVIDIA Geforce 840M 2GB
   d. Hardisk 1 TB
   e. Mouse
   f. Keyboard
2. Spesifikasi Perangkat Lunak
   a. Sistem Operasi Windows 10 Pro Build 18363
   b. Python versi 3.8.4 64 bit
   c. Visual Studio Code versi 1.47.1
   d. Browser Google Chrome versi 84.0.4147.89 64 bit

3. Kajian Pustaka

Kajian pustaka berupa pembahasan dari penelitian – penelitian sebelumnya dan teori – teori penunjang yang secara umum digunakan dalam penelitian ini.

3.1 Web Crawler

Web crawler atau yang dikenal juga dengan istilah web spider atau web robot adalah program yang bekerja dengan metode tertentu dan secara otomatis mengumpulkan semua informasi yang ada dalam suatu website. Web crawler, yang sering disebut crawler saja, akan mengunjungi setiap alamat website yang diberikan kepadanya, kemudian mengorek, mengambil, dan menyimpan semua informasi yang terdapat didalam website tersebut [2].

3.2 Text Mining

Text mining atau penambangan teks adalah proses ekstraksi atau pemisah pola yang berupa pengetahuan dan informasi-informasi yang penting dari sumber dokumen teks seperti dokumen PDF, dokumen Word dan CSV File dan lain sebagainya [3].

3.3 Text Preprocessing

Tujuan dari pemrosesan awal atau preprocessing adalah untuk mempersiapkan text menjadi data yang siap diproses. Proses yang dilakukan pada tahap ini meliputi case folding atau mengubah seluruh huruf yang ada pada dokumen menjadi huruf kecil serta menghilangkan karakter selain huruf, filtering atau tahap mengambil kata penting dengan cara menghilangkan stopwords, tokenization atau tahapan dimana kumpulan karakter dalam teks yang telah melalui proses case folding akan dipecah kedalam satuan kata (token), dan stemming yang merupakan untuk mentransformasikan kata-kata yang terdapat dalam dokumen ke kata-kata akar atau dasarnya dengan menggunakan berbagai aturan tertentu [4].

3.4 TF-IDF (Pembobotan)

TF-IDF (Term Frequency Inverse Document Frequency) merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen (Evan, 2014). TFIDF adalah sebuah algoritma yang umumnya digunakan untuk pengolahan data besar (Kamath, 2014).

3.5 Clustering

Clustering merupakan proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas atau cluster dan meminimumkan

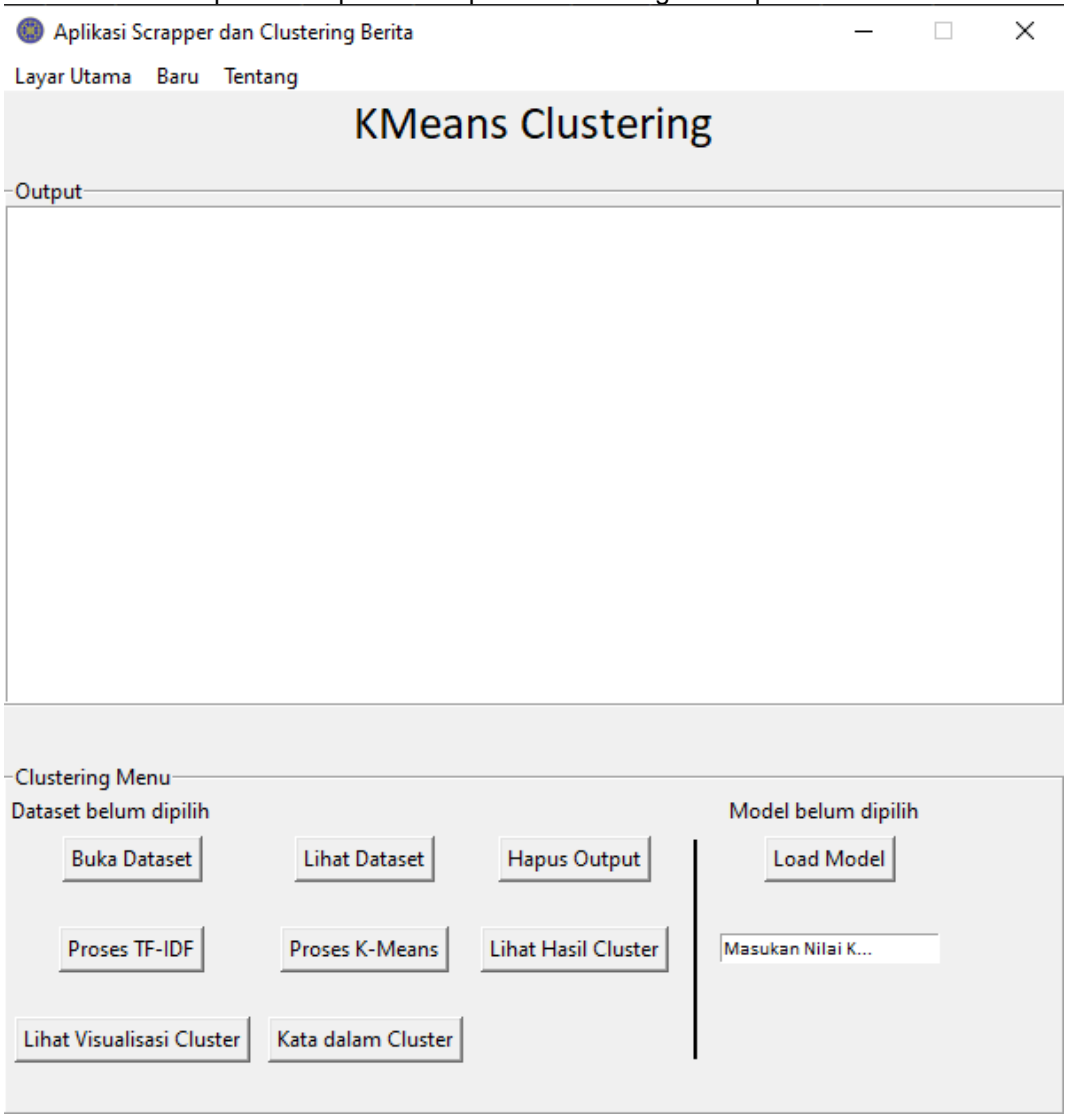
kesamaan antar cluster menurut Tan, Steinbach dan Kumar (2006). Sementara clustering akan membagi data ke dalam grup-grup yang mempunyai objek yang karakteristiknya sama [5].

4. Hasil dan Pembahasan

Membahas tentan hasil penelitian yang dilakukan yaitu Clustering berita menggunakan algoritma TF-IDF dan K-Means dengan memafaatkan sumber data crawling pada situs detik.com.

4.1 Tampilan Aplikasi *Clustering* pada situs detik.com

Berikut merupakan tampilan dari aplikasi Clustering Berita pada situs Detik.com



Gambar 4. Tampilan aplikasi *clustering* pada situs detik.com

Gambar 4. merupakan tampilan dari aplikasi Clustering berita pada situs Detik.com. Terdapat dua buah segment, dimana pada segment pertama berisikan output atau nilai keluaran dari aplikasi, seperti Dataframe, status pembobotan, clustering dan kata yang sering muncul dalam tiap cluster, guna proses identifikasi cluster yang terbentuk. Dan pada segment kedua terdapat Entry Widget yang digunakan untuk memasukan nilai K dan beberapa tombol diantaranya:

1. Buka Dataset, digunakan untuk memilih dataset / data hasil crawling pada situs detik.com.
2. Lihat Dataset, digunakan untuk melihat dataframe berita.
3. Hapus Output, digunakan untuk menghapus tampilan output.
4. Proses TF-IDF, digunakan untuk proses pembobotan kata.

5. Proses K-Means, digunakan untuk clustering dataset
6. Lihat Hasil Cluster, digunakan untuk melihat hasil cluster yang terbentuk yang disimpan dalam dataframe baru.
7. Lihat Visualisasi Cluster, digunakan untuk melihat tampilan grafik sebaran cluster yang terbentuk berdasarkan hasil Proses K-Means.
8. Kata dalam cluster, digunakan untuk melihat kata-kata yang terkandung dalam setiap cluster yang berguna untuk mengidentifikasi setiap cluster.
9. Load Model, digunakan untuk memuat model yang sudah disimpan sebelumnya agar cluster yang terbentuk sesuai dengan model yang disimpan.

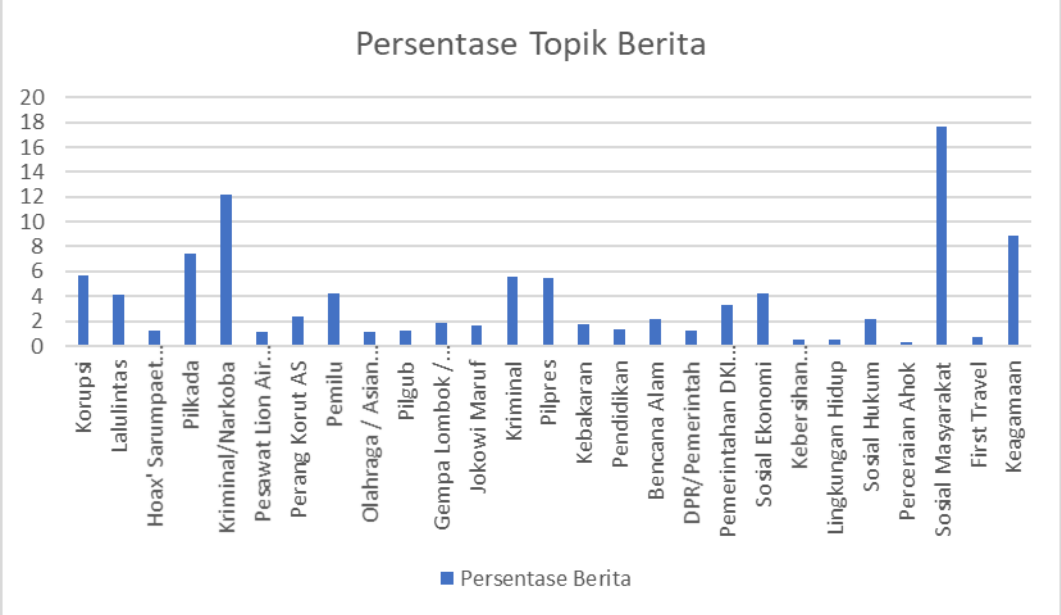
4.2 Analisa Hasil Sistem

Hasil dari implementasi TF-IDF dan Kmeans pada dataset menghasilkan label untuk setiap berita didalam dataset. Untuk mengetahui topik / kategori, maka dilakukan analisa secara manual dengan melihat kata yang sering muncul dari setiap cluster. Tabel 2 merupakan hasil Analisa dari *cluster* yang terbentuk.

Tabel 2. Persentase*Cluster/*Kelompok Berita

No	Bulan	Jumlah
1	Cluster 0 / Korupsi	5,67 %
2	Cluster 1 / Lalulintas	4,16 %
3	Cluster 2 / Hoax Sarumpaet dan Amien Rais	1,21 %
4	Cluster 3 / Pilkada	7,41 %
5	Cluster 4 / Kriminal / Narkoba	12,14 %
6	Cluster 5 / Pesawat lion air jatuh	1,16 %
7	Cluster 6 / Perang Korut AS	2,38 %
8	Cluster 7 / Pemilu	4,24 %
9	Cluster 8 / Olahraga / Asian Games	1,12 %
10	Cluster 9 / Pilgub	1,20 %
11	Cluster 11 / Gempa Lombok / Tsunami Palu	1,82 %
12	Cluster 12 / Jokowi Maaruf	1,66 %
13	Cluster 13 / Kriminal	5,56 %
14	Cluster 14 / Pilpres	5,42 %
15	Cluster 15 / Kebakaran	1,72 %
16	Re-Cluster 0 / Pendidikan	1,30 %
17	Re-Cluster 1 / Bencana Alam	2,12 %
18	Re-Cluster 2 / DPR – Pemerintah	1,24 %
19	Re-Cluster 3 / Pemerintah DKI Jakarta	3,33 %
20	Re-Cluster 4 / Sosial Ekonomi	4,22 %
21	Re-Cluster 5 / Kebersihan Lingkungan	0,53 %
22	Re-Cluster 6 / Lingkungan Hidup	0,50 %
23	Re-Cluster 7 / Sosial Hukum	2,19 %
24	Re-Cluster 8 / Perceraian Ahox	0,31 %
25	Re-Cluster 9 / Sosial Masyarakat	17,67 %
26	Re-Cluster 10 / First Travel	0,67 %
27	Re-Cluster 11 / Keagamaan	8,91 %

Hasil akhir analisis terbentuk 27 Kategori Berita dari 119.422 berita didalam dataset. Berikut penulis sajikan dalam bentuk diagram pada gambar 5.



Gambar 5. Diagram Persentase Topik Berita

Gambar 5 menunjukan persentase Cluster berita yang ada pada situs detik.com selama tahun 2018. Persentase diambil dari jumlah berita tiap cluster dibagi total berita yang diproses dikalikan 100.

Dari gambar 5 dapat penulis perhatikan terdapat Cluster berita yang bersifat Umum dan Khusus. Cluster yang bersifat Khusus mengindikasikan topik / tren tersebut pernah viral atau sering diberikatan. Ini dikarenakan pada tahap ekstraksi fitur / pembobotan dataset, sebaran datanya paling sempit dikarenakan banyaknya berita yang membahas topik tersebut sehingga dalam tahap clustering membentuk sebuah cluster utuh.

5. Kesimpulan

Crawling dataset menggunakan bahasa pemrograman Python dan library BeautifulSoup4 mampu dan handal dalam melakukan pengambilan data text dalam jumlah besar pada portal berita online detik.com. Terdapat dua proses yang dilakukan, diantaranya Pengambilan Data URL dengan total waktu yang dibutuhkan selama 1 Jam 55 Menit 20 Detik dan Pengambilan Data Berita dengan total waktu yang dibutuhkan selama 2 Hari 18 Jam 1 Menit 39 Detik. Total berita yang diperoleh sebanyak 124.509 berita.

Dari 124.509 berita yang diproses, Algoritma TF-IDF dan K-Means dapat mengindetifikasi sebanyak 27 cluster berita yang dibagi kedalam 2 tahapan clustering. Tahap pertama sebanyak 16 cluster dan tahap kedua sebanyak 12 cluster. Hal ini dikarenakan pada tahapan awal terdapat satu cluster yang tidak dapat diidentifikasi karena terdapat kata-kata yang masih umum sehingga harus dilakukan clustering ulang.

Secara keselurahan, Bahasa pemprograman Python dapat menangani data dalam jumlah besar (Big Data) dengan baik. Hal ini dapat dilihat dari Diagram Penggunaan Processor, Diagram Penggunaan Memory dan Diagram Waktu Komputasi yang masih wajar dengan spesifikasi komputer yang dapat dikatakan sedang.

Daftar Pustaka

[1] Husni, Yudha Dwi Putra Negara, M. Syarief (2015). Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means. *Jurnal SimanteC,* vol. 4, no. 3, 159 - 160

[2] Wayan M. Wijaya (2019). TEKNOLOGI BIG DATA Sistem Canggih dibalik Google Yahoo! Facebook IBM.

[3] Edy Susanto, Viny Christanti Mawardi, Manatap Dolok Lauro (2021). Aplikasi Clustering Berita Dengan Metode K Means Dan Peringkas Berita Dengan Metode Maximum Marginal Relevance. *Jurnal Ilmu Komputer dan Sistem Informasi,* 62-63

[4] Ni Komang Widyasanti, I Ketut Gede Darma Putra, Ni Kadek Dwi Rusjayanthi (2018). Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia. *Merpati,* vol. 6, no. 2, 121-122

[5] Muhammad Sholeh hudin, M Ali Fauzi, Sigit Adinugroho (2018). Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2, no. 11, 5519-5520