Pengembangan Rekayasa dan Teknologi, Vol 3, No. 1, Mei 2019, pp 43-47
p-ISSN: 1410-9840 & e-ISSN: 2580-8850
http://journals.usm.ac.id/index.php/jprt/index

PENGGUNAAN METODE K-MEANS PADA ANALISA DAN KLASIFIKASI CAPRES 2019 DI TWITTER

Ahmad Rifa'i [1], Galet Guntoro Setiaji [2], Vensy Vydia [3]
[1,2,3] Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang
[1] rifai@usm.ac.id. [2] gallet@usm.ac.id, [3] vensy@usm.ac.id

ABSTRAK

Media sosial merupakan wadah untuk mencurahkan kata-kata melalui tulisan pada dunia maya (*online)*, sehingga pada nantinya akan dapat dilihat oleh banyak orang baik yang dikenal maupun tidak dikenal. Twitter merupakan jejaring sosial atau media sosial yang memiliki teks terbatas, yaitu memiliki Panjang teks sebanyak 280 karakter. Metode crawling merupakan metode yag digunakan untuk pengambilan data dari media sosial yang dapat digunakan sebagai *developer application.*

Crawling data pada media sosial twitter ini menggunakan aplikasi spyder python 3.7 yang berada didalam aplikasi anaconda. Dengan aplikasi tersebut akan di dapatkan data-data berkenaan dengan capres 2019. data-data tersebut akan diolah menggunakan metode *K-Means*. Setelah dilakukan pengolahan data menggunakan K-Means di dapatkan untuk cluster 1 (Jokowi) muncul sebanyak 3 kali sedangkan untuk cluster 2 (Prabowo) muncul sebanyak 1 kali.

ABSTRACT

Social media is a place to devote words through writing to the online world, so that in the future it can be seen by many people, both known and unknown. Twitter is a social network or social media that has limited text, which has a text length of 280 characters. The crawling method is a method used to extract data from social media that can be used as an application developer.

Data crawling on social media Twitter uses the Spyder Python 3.7 application that is in the Anaconda application. With the application, the data will be obtained regarding the 2019 presidential candidate. The data will be processed using the K-Means method. After processing data using K-Means, it was obtained for cluster 1 (Jokowi) to appear 3 times while cluster 2 (Prabowo) appeared 1 time..

Keyword: Social Media, Twitter, Crawling Data, K-Means

I. Pendahuluan

Media sosial dijaman sekarang merupakan wadah untuk tempat mencurahkan kata-kata lewat tulisan, dimana tulisan itu nantinya bisa dilihat oleh semua orang yang dikenal atau tidak. Dari sini media sosial merupakan alat yang bisa mempengaruhi seseorang bahkan sekelompok orang. Media sosial yang kita kenal diantaranya facebook, instagram dan twitter yang sebagian besar media sosial ini digunakan mulai pertemanan hingga perang argument.

Dimana salah satunya sebagian peneliti juga menggunakan media sosial sebagai pengumpul data, disini paling sering digunakan yaitu media sosial twitter. Kenapa menggunakan media sosial twitter, karena media sosial ini memiliki apps yang bisa digunakan sebagai developer application. Dimana data-data yang ingin di cari dapat diambil dengan metode crawling.

Disini kita mengambil atau mengcrawling dengan focus data tentang calon presiden Indonesia tahun 2019. Dimana nanti data crawling digunakan sebagai data set, yang diolah dengan sebuah metode data mining yaitu k-means. Diharapkan dari sebuah data itu nantinya akan memberikan sebuah informasi.

II. Data Mining

Data mining didefinisikan sebagai proses mengekstrak atau menambang pengetahuan yang dibutuhkan dari sejumlah data besar (Han dan Kamber, 2006 : 5). Pada prosesnya data mining akan mengekstrak informasi yang berharga dengan cara menganalisis adanya pola-pola ataupun hubungan keterkaitan tertentu dari data-data yang berukuran besar. Data mining berkaitan dengan bidang ilmu-

ilmu lain, seperti *Database System, Data Warehousing, Statistic, Machine Learning, Information Retrieval*, dan Komputasi Tingkat Tinggi. Selain itu data mining didukung oleh ilmu lain seperti *Neural Network, Pengenalan Pola, Spatial Data Analysis, Image Database, Signal Processing*.

Pengetahuan yang dihasilkan dari proses data mining harus baru, mudah dimengerti, dan bermanfaat. Dalam data mining, data disimpan secara elektronik dan diproses secara otomatis oleh komputer menggunakan teknik dan perhitungan tertentu (Pramadhani dan Setiadi, 2014).

Menurut Maclennan, Tang, & Crivat (2009, p6). Berikut adalah fungsi *data mining* secara umum :

1. *Classification*

   Classification adalah proses untuk mencari model atau fungsi yang menggambarkan dan membedakan kelas-kelas atau konsep data. Fungsi dari *Classification* adalah untuk mengklasifikasikan suatu target *class* ke dalam kategori yang dipilih.

2. *Clustering*

   Fungsi dari *clustering* adalah untuk mencari pengelompokkan atribut ke dalam segmentasi-segmentasi berdasarkan similaritas.

3. *Association*

   Fungsi dari *association* adalah untuk mencari keterkaitan antara atribut atau *item set*, berdasarkan jumlah *item* yang muncul dan *rule association* yang ada

4. *Regression*

   Fungsi dari *regression* hampir sama dengan klasifikasi. Fungsi dari *regression* adalah bertujuan untuk mencari prediksi dari suatu pola yang ada

5. *Forecasting*

   Fungsi *dari forecasting* adalah untuk peramalan waktu yang akan datang berdasarkan *trend* yang telah terjadi di waktu sebelumnya

6. *Sequence Analysis*

   Fungsi *dari sequence analysis* adalah untuk mencari pola urutan dari rangkaian kejadian

7. *Deviation Analysis*

   Fungsi dari *deviation analysis* adalah untuk mencari kejadian langka yang sangat berbeda dari keadaan normal (kejadian *abnormal*).

2.1 Algoritma K-Means

K-Means merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok (Krisna, 2016). Metode ini mempartisi data kedalam kelompok yang lain. Adapun tujuan pengelompokan data ini adalah meminimalkan variasi didalam suatu kelompok dam memaksimalkan variasi antar kelompok. Dengan mengunakan teknik klustering dalam K-means, maka tahapan algoritma K-means adalah sebagai berikut :

a. Menentukan koordinat titik tengah setiap cluster
b. Menentukan jarak setiap onjek terhadap kordinat titik tengah
c. Mengelompokan objek-objek tersebut berdasarkan pada jarak minimumnya
d. Tentukan pusat cluster baru
e. Apakah ada selisih antar cluster lama dengan baru?jika masih ada kembali ke langkah a hingga d, jika tidak selesai.

Di dalam menentukan titik centroid kita menggunakan perhitungan jarak Euclidian distance, dengan rumus sebagai berikut :

$$d_{ij} = \sqrt{\left[(x_i - x_j)^2 + (y_i - y_j)^2\right]}$$

x_i = koordinat x untuk fasilitas i

y_i = koordinat y untuk fasilitas i

d_{ij} = jarak antar fasilitas I dan j

2.2 Proses Crawling Data Twitter

Dalam melakukan crawling data pada sosial media twitter ada beberapa proses yang terjadi diantaranya:

a. Konfigurasi API twitter

   Proses yang dilakukan sebelum melakukan crawling data yaitu melakukan twitter authentication dengan mengisikan

consumer_key, consumer_secret, access_token, access_secret.

```python
def load_api():
    ''' Function that loads the twitter API after authorizing the user. '''

    consumer_key = 'MVhn1subTp8M1Q3R37cq4Mlhe'
    consumer_secret = '6A81vtI3ZiYIlTLQ6dHkJ9z1HgS3lHO4jDmkMJNc1emYlOuYsN'
    access_token = '927346645217325056-hUD4ZembJVWpek2zQAxTzgrpoCye0B4'
    access_secret = '0rlqd1jPEzUld9PxHSeTjTKP7PcgZ7YXGXrzvHe3DTLYp'
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_secret)
    # Load the twitter API via tweepy
    return tweepy.API(auth)


def tweet_search(api, query, max_tweets, max_id, since_id, geocode):
    ''' Function that takes in a search string 'query', the maximum
        number of tweets 'max_tweets', and the minimum (i.e., starting)
        tweet id. It returns a list of tweepy.models.Status objects. '''

    searched_tweets = []
    while len(searched_tweets) < max_tweets:
        remaining_tweets = max_tweets - len(searched_tweets)
        try:
            new_tweets = api.search(q=query, count=remaining_tweets,
                                    since_id=str(since_id),
                                    max_id=str(max_id-1))
#                                    geocode=geocode)
            print('found',len(new_tweets),'tweets')
            if not new_tweets:
                print('no tweets found')
                break
            searched_tweets.extend(new_tweets)
            max_id = new_tweets[-1].id
        except tweepy.TweepError:
            print('exception raised, waiting 15 minutes')
            print('(until:', dt.datetime.now()+dt.timedelta(minutes=15), ')')
            time.sleep(15*60)
            break # stop the loop
    return searched_tweets, max_id
```

Gambar 1. Konfigurasi API Twitter

b. Konfigurasi Crawling Data

Masukkan ketentuan-ketentuan data yang akan diambil dari sosial media twitter yaitu meliputi hastag, maksimal tweet dan rentang waktu pengambilan data.

```python
def main():
    ''' This is a script that continuously searches for tweets
        that were created over a given number of days. The search
        dates and search phrase can be changed below. '''



    ''' search variables: '''
    '''search_phrases = ['jokowi','prabowo','2periode','gantipresiden']'''
    search_phrases = ['jokowi']
    time_limit = 1.5                        # runtime limit in hours
    max_tweets = 100                        # number of tweets per search (will be
                                            # iterated over) - maximum is 100
    min_days_old, max_days_old = 6, 7       # search limits e.g., from 7 to 8
                                            # gives current weekday from last week,
                                            # min_days_old=0 will search from right now
    USA = '-7.0245542,110.347024,2300km '    # this geocode includes nearly all American
                                            # states (and a large portion of Canada)
```

Gambar 2. Konfigurasi Crawling Data

2.3 Proses Olah Data dengan K-Mean

Data yang sudah di normalisasi sekitar 1000 data yang berisi tentang hastag jokowi dan hastag Prabowo. Dengan terkumpulnya data maka langkah selanjutnya adalah mengolah dengan algoritma K-Means, langkah-langkahnya sebagai berikut :

1. Menentukan Awalan Cluster Secara Random

Disini kita akan melakukan penelitian sebanyak 4 kali, dan menentukan K awal juga sebanyak 4 data K.

Ex 1		Ex 2		Ex 3		Ex 4	
120	Jokowi	224	Jokowi	114	Jokowi	248	Jokowi
0.0000	0.0000	0.0000	0.0000	0.4300	0.0000	0.0000	0.0000
0.1100	0.1100	0.1400	0.1500	0.0000	0.1500	0.1700	0.1600
0.0250	0.0100	1.8270	11.5200	0.0100	0.0230	0.0300	66.2920
0.0100	0.0000	0.8230	4.6090	0.2300	1.1700	2.6900	21.2700
0.0000	0.0000	0.0000	0.0000	0.0000	0.9560	9.6360	0.0630
0.0000	0.0000	0.0000	0.0000	0.0000	2.6940	13.9360	0.0530
0.0000	0.0180	0.0000	0.8230	0.0000	0.9560	9.6360	0.0630
0.0000	0.3400	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
215	Prabowo	398	Prabowo	71	Prabowo	111	Prabowo
0.8500	0.1100	0.0000	0.1500	0.0000	0.1500	0.0000	0.1600
6.6780	48.4400	0.0340	1.6500	0.0230	1.1700	0.0300	2.6900
0.0000	0.0000	2.5960	7.8620	0.9560	2.6940	9.6360	13.9360
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Gambar 3. Menentukan Awalan Cluster Secara Random

2. Perhitungan Jarak Pusat Cluster

Pengukuran jarak antara data dengan pusat cluster menggunakan *Euclidian distance.* Dari perhitungan tersebut didapatkan nilai dari Cluster 1 dan nilai dari Cluster 2. Nilai terendah akan diambil sebgai Cluster yang dipilih.

C1	C2	Min	C
4.7914	47.8780	4.7914	C1
69.5223	21.2119	21.2119	C2
1.0077	48.0451	1.0077	C1
9.5421	39.4339	9.5421	C1
11.4388	46.1419	11.4388	C1
0.9772	48.0741	0.9772	C1
1.6959	47.2003	1.6959	C1
2.0223	46.8740	2.0223	C1

C1	C2	Min	C
9.0814	39.8424	9.0814	C1
11.9927	46.0510	11.9927	C1
0.9563	47.9505	0.9563	C1
0.8773	48.0283	0.8773	C1
9.3527	39.5524	9.3527	C1
1.1202	47.7815	1.1202	C1
14.4568	34.5645	14.4568	C1
4.9058	44.0202	4.9058	C1

Tabel 1. Perhitungan Jarak Pusat Cluster

3. Pengelompokan Data dan Menentukan Pusat Cluster Terbaru

   Setelah dilakukan perhitungan, langkah selanjutnya adalah melakukan pengelompokan data dan menetukan pusat cluster terbaru.

Jokowi	Prabowo
0.161986	0
0.119178	0.093333
2.054158	1283.388
6.365466	0.236667
0.711456	0.012
1.532297	0.044
0.712572	0.018
0.020271	0.22

Tabel 2. Pengelompokan Data dan Menentukan Pusat Cluster Terbaru

4. Ulangi langkah ke 2 dan 3 hinga pusat cluster baru tidak berubah

   Hasil objek Interasi ke 1

Fungsi Objektif Awal	1000.00
Fungsi Objektif J	734.49
Fungsi Objektif	**265.51**

Tabel 3. Hasil Objek Iterasi ke 1

   Hasil objek Interasi ke 13 didapatkan fungsi objektif

Fungsi Objektif Awal	680.81
Fungsi Objektif J	680.81
Fungsi Objektif	**0.00**

Tabel 4. Hasil objek Interasi ke 13 didapatkan fungsi objektif

5. Dari 4 tahapan diatas dan 4 kali penelitian dihasilkan data sebagai sebagai berikut

Penelitian	Interasi	Cluster
1	5x	C1
2	6x	C2
3	7x	C1
4	4x	C1

Tabel 5. Hasil 4 Kali Tahapan

6. Cara mengitung DBI
   - Menghitung *Sum of square within Cluster* ***(SSW)***

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i)$$

SSW 1	8.956667
SSW 2	384.0227

   - Menghitung *Sum of square between cluster* ***(SSB)***

$$SSSB_{i,j} = d(c_i, c_j)$$

SSB		***Data ke i***	
		1	2
Data ke i	1	0.0000	1281.3494
	2	1281.3494	0.0000

   - Mendefinisikan ukuran rasio seberapa baik nilai atara cluster ($R_{i,j}$)

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

   - Menghitung *Davis-Bouldin Index* ***(DBI)***

$$DBI = \frac{1}{K}\sum_{i=1}^{K}{}^{max}_{i\neq j}(R_{i,j})$$

R		***Data ke i***			
		1	2	***Rmax***	**DBI**
Data ke i	1	0	0.78	0.78	**0.52**
	2	0.78	0	0.78	
				1.561	

Sehingga

a. Hasil perhitungan K-Means dipengaruhi saat penentuan nilai awal k
b. Dari 4 data awal dan 4 kali penelitian K-Means dapatkan cluster pertama yang lebih besar nilainya.
c. Dari 4 tahapan tersebut di dapatkan kemunculan Cluster 1 sebanyak 3 kali dan kemunculan Cluster 2 sebanyak 1 kali
d. Validitas Mengunakan DBI dari ke empat data, didapatkan validitas sebesar **0.52**

III. Kesimpulan

Crawling data pada media sosial twitter ini menggunakan aplikasi spyder python 3.7 yang berada didalam aplikasi anaconda. Dengan aplikasi tersebut akan di dapatkan data-data berkenaan dengan capres 2019. data-data tersebut akan diolah menggunakan metode *K-Means*. Setelah dilakukan pengolahan data menggunakan K-Means di dapatkan untuk cluster 1 (Jokowi) muncul sebanyak 3 kali sedangkan untuk cluster 2 (Prabowo) muncul sebanyak 1 kali. Dari ke empat data, didapatkan validitas menggunakan DBI sebesar **0.52**

Daftar Pustaka

Astuti, Fajar Hermawati (2013). *Data Mining*. Andi Publisher

Han, Jiawei dan Kamber, Micheline. (2007), Data Mining : Concept and Techniques Second Edition, Morgan Kaufmann Publishers.

J. MacLennan, Z. Tang and B. Crivat. (2009). Scalable Varied Density Clustering Algorithm for Large Datasets. **Journal of Software Engineering and Applications**, Vol.3 No.6

Kusrini dan EmhaTaufiq Lutfi (2009). *Algoritma Data Mining.* Andi Publisher

Prasetyo, Eko (2013). *Data Mining (Konsep dan Aplikasi Menggunakan Matlab)*. Andi Publisher

Pramadhani, A., Embun & Setiadi, T. (2014). Penerapan Data Mining untuk Klasifikasi Prediksi Penyakit Ispa dengan Algoritma Decision Tree (ID3), Jurnal Sarjana Teknik Informatika Volume 2 Nomor 1 e-ISSN: 2338-5197

Prasetyo, Eko (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Andi Publisher

Wardhani, A., Krisna. (2016). Implementasi Algoritma K-Means Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Kajen Pekalongan, Jurnal Transformatika, Volume 14, Nomor 1.